NEPC

NATIONAL EDUCATION POLICY CENTER

International Test Score Comparisons and Educational Policy

A Review of the Critiques

Martin Carnoy

Stanford University

October 2015

National Education Policy Center

School of Education, University of Colorado Boulder
Boulder, CO 80309-0249
Telephone: (802) 383-0058
Email: NEPC@colorado.edu
http://nepc.colorado.edu

This is one of a series of briefs made possible in part by funding from The Great Lakes Center for Education Research and Practice.

GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

http://www.greatlakescenter.org
GreatLakesCenter@greatlakescenter.org

Briefs published by the National Education Policy Center (NEPC) are blind peer-reviewed by members of the Editorial Review Board. Visit http://nepc.colorado.edu to find all of these briefs. For information on the editorial board and its members, visit: http://nepc.colorado.edu/editorial-board.

Suggested Citation:

Carnoy, M. (2015). *International Test Score Comparisons and Educational Policy: A Review of the Critiques*. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/international-test-scores.

International Test Score Comparisons and Educational Policy: A Review of the Critiques

Martin Carnoy, Stanford University

Executive Summary

In this brief, we review the main critiques that have been made of international tests, as well as the rationales and education policy analyses accompanying these critiques—particularly the policy analyses generated by the Program for International Student Assessment (PISA) of the Organization for Economic Cooperation and Development (OECD).

We first focus on four main critiques of analyses that use average PISA scores as a comparative measure of student learning:

- Critique #1: Whereas the explicit purpose of ranking countries by average test score is to allow for inferences about the quality of national educational systems, the ranking is misleading because the samples of students in different countries have different levels of family academic resources (FAR).
- Critique #2: Students in a number of countries, including the United States, have made large FAR-adjusted gains on the Trends in International Mathematics and Science Study (TIMSS) test 1999-2011, administered by the International Association for the Evaluation of Educational Achievement (IEA). However, they have shown much smaller, or no, gains on the FAR-adjusted PISA test. This raises issues about whether one test or the other is a more valid measure of student knowledge.
- Critique #3: The error terms of the test scores are considerably larger than the testing agencies care to admit. As a result, the international country rankings are much more in "flux" than they appear.
- Critique #4: The OECD has repeatedly held up Shanghai students and the Shanghai educational system as a model for the rest of the world and as representative of China, yet the sample is not representative even of the Shanghai 15-year-old population and certainly not of China. In addition, Shanghai schools systematically exclude migrant youth. These issues should have kept Shanghai scores out of any OECD comparison group and raise serious questions about the OECD's brand as an international testing agency.

This brief also discusses a set of critiques around the underlying social meaning and educational policy value of international test comparisons. These comparisons indicate

how students in various countries score on a particular test, but do they carry a larger meaning? There are four main critiques in this regard.

First, claims that the average national scores on mathematics tests are good predictors of future economic growth are, at best, subject to serious questions and, at worst, gross misuse of correlational analysis. The U.S. case appears to be a major counterexample to these claims. Japan is another.

Second, the use of data from international tests and their accompanying surveys have limited use for drawing educational policy lessons. This is because cross-sectional surveys such as the TIMSS and PISA are not amenable to estimating the causal effects of school inputs on student achievement gains. Further, unlike TIMSS, PISA neither directly measures teacher characteristics and practices, nor can it associate particular teachers with particular students. Yet, again in the case of the OECD, there seem to be no end of asserted policy lessons—none with appropriate causal inference analysis, many based on questionable data, and others largely anecdotal—proposed by the same agency that developed and applied the test.

Third, critiques have pointed to the conflict of interest that arises because the OECD (and its member governments) acts simultaneously as testing agency, data analyst, and interpreter of results for policy purposes.

Fourth, a recent critique questions the relevance of nation-level test score comparisons of countries with national educational systems to other countries with more diverse and complex systems—such as the United States, with its 51 (including the District of Columbia) highly autonomous geographic educational administrations. This newest critique goes beyond the questions raised about the validity of international test comparisons and even beyond the careless way results are used to draw conclusions about "good" educational policies. PISA and TIMSS scores for U.S. states show large variation in student performance among states. PISA results for U.S. states are available only in 2012, but FAR-adjusted TIMSS scores are available for a number of U.S. states over more than a decade. These show large performance gains for some states and smaller gains for others. The critique suggests that from the standpoint of U.S. educational analysts and politicians, it would seem much more relevant and interesting to employ state-level test results over time to understand the policies high-gaining states implemented in the past 20 years than to examine other countries' educational policies—if, indeed, it is their educational policies—that are behind any large test score gains made in the decade of the 2000s.

Recommendations

- PISA and TIMSS should report all international test results by FAR (family academic resource) subgroups of students with different levels of resources such as books in the home or mother's education. Relatedly, PISA and TIMSS should report all changes in test scores over time by FAR subgroups. Compare country results by student FAR subgroup together with country aggregate averages.

- The OECD and the IEA should make the individual-level student micro-data for the PISA and TIMSS tests available at the same time as the results are formally announced. This would allow international researchers to produce independent analyses of the data within a week of the time when the OECD's and IEA's versions of the results appear.

- Beyond allowing access to individual-level student micro-data immediately, the OECD should separate international tests' design, application, and results from data analysis and policy recommendations derived from the tests and surveys. The OECD should also include independent academic expert appointments to PISA's decision-making board, which governs the application and use of test data, as is the practice at the IEA for the TIMSS test.

- In the United States, the National Center of Educational Statistics should publish PISA reading, mathematics and science scores and TIMSS mathematics and science scores by FAR group, particularly FAR-adjusted PISA and TIMSS scores over time. There will be a golden opportunity to do this for the PISA and TIMSS together in December 2016, when the results of the 2015 PISA and TIMSS will be announced in the same month.

- In the United States, policymakers should shift focus away from why students in other countries may do "better" than U.S. students as a whole and instead focus on why student achievement gains have been greater in some U.S. states and lower in others.

International Test Score Comparisons and Educational Policy: A Review of the Critiques

Introduction

Since its inception in the year 2000, the Program for International Student Assessment (PISA),[1] an international test of reading, mathematics and science, has indicated that American 15-year-olds perform more poorly, on average, than 15-year-olds in many other countries. This finding is consistent with results from another international assessment of 8th-graders, the Trends in International Mathematics and Science Survey (TIMSS).[2]

Both the PISA and the TIMSS rank countries, large and small, by the performance of stratified random samples of students on a particular test. However, their methodology varies in important ways. In the case of PISA, the sample is based on student age and school, which means that in some countries 15 year-old students are drawn from multiple grades. In TIMSS, the sample is based on grade as well as classroom and school. PISA does not interview teachers, and sampled students cannot be linked with their teachers; in TIMSS, teachers are interviewed and students and teachers can be linked. PISA and TIMSS both test mathematics and science. PISA additionally tests reading.

Extensive educational research in the United States has demonstrated that students' family and community characteristics powerfully influence their school performance.

Extensive educational research in the United States has demonstrated that students' family and community characteristics powerfully influence their school performance. From such tests, many journalists and policymakers have concluded that American student achievement lags woefully behind that in many comparable industrialized nations, that this shortcoming threatens the nation's economic future, and that these test results therefore suggest an urgent need for radical school reform. Journalists and policymakers in many other countries similarly show intense interest in results, in some cases because their students did quite poorly (Latin American countries), and in others, because their students did very well, which creates the kind of national pride associated with national sports victories (Finland, Estonia, Korea, Singapore).

Upon release of the 2011 TIMSS results, for example, U.S. Secretary of Education Arne Duncan called them "unacceptable," saying that they "underscore the urgency of accelerating achievement in secondary school and the need to close large and persistent achievement gaps."[3] A year later, when the 2012 PISA results were released on "PISA Day" 2013, Duncan pointed to positive reductions in high school dropouts and increasing

numbers of Hispanic students attending college. And yet, he also concluded that "The big picture... is straightforward and stark: It is a picture of educational stagnation." He went on to say that, "The problem is not that our 15-year olds are performing worse today than before. The problem instead is that they are not making progress.[4] Yet students in many nations...are advancing, instead of standing still." As he had argued three years earlier, when the 2009 PISA results were announced, the critical nature of such lack of progress is that " In a knowledge-based, global economy, where education is more important than ever before, both to individual success and collective prosperity, our students are basically losing ground. We're running in place, as other high-performing countries start to lap us."[5]

A 2013 comprehensive report titled *What Do International Tests Really Show about American Students' Performance?* (Carnoy and Rothstein)[6] criticized the way U.S. students' performance on international tests was being interpreted, writing that it was

> "...oversimplified, exaggerated, and misleading. It ignores the complexity of the content of test results and may well be leading policymakers to pursue inappropriate and even harmful reforms that change aspects of the U.S. education system that may be working well and neglect aspects that may be working poorly" (p. 7).

The main argument in the report was that by not adjusting international test scores for major national differences in students' family academic resources, such as books in the home or mother's education, journalists and policymakers mistakenly attributed the poor performance of U.S. students entirely to the quality of U.S. education. The second part of the report's argument was that focusing on "national progress in test scores" failed to differentiate gains over time for disadvantaged and advantaged students and how these compared to gains over time for similarly advantaged or disadvantaged students in other countries.

Education analysts in the United States pay close attention to the level and trends of test scores disaggregated by socioeconomic groupings. Indeed, a central element of U.S. domestic education policy is the requirement that average scores be reported separately for racial and ethnic groups and for children who are from families whose incomes are low enough to qualify for the subsidized lunch program. A school with high proportions of disadvantaged children may be able to produce great "value-added" for its pupils, although its average test score levels may be low. It would be foolish to fail to apply this same understanding to comparisons of international test scores.

Extensive educational research in the United States has demonstrated that students' family and community characteristics powerfully influence their school performance. Children whose parents read to them at home, whose health is good and can attend school regularly, who do not live in fear of crime and violence, who enjoy stable housing and continuous school attendance, whose parents' regular employment creates security, who are exposed to museums, libraries, music and art lessons, who travel outside their immediate neighborhoods, and who are surrounded by adults who model high educational achievement and attainment will, on average, achieve at higher levels than children

without these educationally relevant advantages. Much less is known about the extent to which similar factors affect achievement in other countries, but we should assume, in the absence of evidence to the contrary, that they do.

Countries' educational effectiveness and their family academic resource composition can change over time. Consequently, comparisons of test score trends over time by family academic resource group provide more useful information to policymakers than comparisons of total average test scores at one point in time, or even of changes in total average test scores over time.

The third part of the argument in the report was that different international and national tests produced different pictures of mathematics achievement gains by U.S. students in the same period, 1999-2012. Significantly, students in several other countries, such as England/UK, Finland, and Russia perform very differently on the TIMSS and PISA mathematics tests. These differences suggest caution in using any single test as a basis for education reform. Because the full range of knowledge and skills that describe "mathematics" cannot possibly be covered in a single brief test, the report suggested that policymakers should carefully examine whether an assessment called a "mathematics" test necessarily covers knowledge and skills similar to those covered by other assessments also called "mathematics" tests, and whether performance on these different assessments can reasonably be compared. The report pointed out that, for example, American adolescents perform relatively well on algebra questions, and relatively poorly on geometry questions, compared to adolescents in other countries. Thus, student math achievement in the United States would compare more favorably to student achievement in other countries when a test has more algebra items and fewer geometry items. Whether there is an appropriate balance between these topics on any particular international assessment is rarely considered by policymakers who draw conclusions about the relative performance of U.S. students from that assessment. Similar questions arise with regard to a "reading" test. There are undoubtedly sub-skills covered by international reading and math tests on which some countries are relatively stronger and others are relatively weaker. Authors Carnoy and Rothstein recommended that these differences should be investigated before policy conclusions are drawn from international test scores.

Review: The Critical Discussion of the Relevance (and Validity) of International Test Results for Policy-Making

Carnoy and Rothstein's research and others' have produced a much broader discussion in the past three years about the relevance of rankings on international tests for U.S. or any other country's educational policy reform strategies.

In assessing these critiques, the reader should keep in mind that the international tests have yielded interesting results for many countries, helping them—when adjusted for students' family background—to benchmark their students' performance relative to other countries. More important, the results can help benchmark changes in scores within their own countries. The descriptive analyses developed by the OECD reports are often very

informative. However, they are descriptions, and descriptions most relevant to analyzing test score patterns within countries. The point of most of the critiques we review is that the test results are used in ways beyond their capacity as measures of student learning and far beyond their capacity for drawing policy conclusions. In large diverse countries, such as the United States, another important critique is that they are not very useful for benchmarking relative to other countries.

Higher than reported possibility of error in PISA rankings

One set of critiques, by a number of European statisticians and mathematicians, argues that the PISA rankings (and, by implication, the TIMSS rankings) are subject to much larger errors than reported in the PISA results.[7] Part of the critique is that the structure of the questions making up the various parts of the PISA test may not correctly capture differences in students' subject knowledge since questions have to be designed ("smoothed out") to estimate math or reading or science knowledge across a wide range of cultures. Another part is that PISA imputes "plausible values" for each student's score because not all students answer the same questions on the test. That is a common practice, but the Rasch model used to estimate plausible values relies on items having similar degrees of difficulty in different countries. Although PISA seeks to eliminate questions that are biased toward particular countries, Danish researcher Svend Kreiner "was not able to find two items in Pisa's tests that function in exactly the same way in different countries," which invalidates the Rasch model.[8] Further, in "off years," when one of the three subjects is not the main focus of the PISA test, only a minority of students actually takes the test. PISA's main focus in 2000 and 2009 was reading, in 2003 and 2012, it was mathematics, and in 2006, it was science. In the "off years" of 2003, 2006, and 2012, a country's reading scores were estimated on the basis of only 40% of the student sample actually taking the test, and not all those answering the same questions. This was also the case for the mathematics test in 2000, 2006, and 2009, and for the science test in 2000, 2003, 2009, and 2012. The OECD's PISA team admits to the validity of some of these critiques. but they simultaneously deny that their testing and imputation methods raise serious questions about the credibility of PISA rankings.

PISA politics and the Shanghai case

Another discussion was about the representativeness of PISA samples and, simultaneously, about the political role of PISA in making claims about educational quality in various countries based on PISA test results. In 2009, China participated for the first time in the PISA (Hong Kong has participated since 2000 and has taken the TIMSS from 1995-2011). Although students in a number of provinces took the test, the OECD only reported—or was allowed to report—the Shanghai (highest) scores. These turned out to be by far the top of the PISA rankings. Students in Shanghai outscored students in Singapore and Hong Kong (other "Chinese" cities) by about 0.5 standard deviations in mathematics and 0.3 standard deviations in reading. Thanks to these results and the publicity given them by the OECD's media team, Shanghai students' performance quickly became a benchmark for how well

students worldwide should be able to achieve academically. Shanghai scores also became conflated with China's *national* performance.

The Brookings Institution's Tom Loveless[9] and James Harvey, Executive Director of the National Superintendents Roundtable,[10] were sharply critical of the representativeness of the Shanghai sample based on simple calculations of Shanghai's population of 15 year-olds. They claimed that the PISA sample had excluded most of Shanghai's large rural migrant population of 15 year-olds because migrant youth did not have the right to attend Shanghai's schools. This is known as the *hukou* policy. Loveless also questioned why the OECD had agreed to publish the results from only one Chinese city when the China sample had included twelve other Chinese provinces, all with lower scores, some much lower, and whether the Shangahi scores were representative of 15 year-olds from China's rural population, which accounts for 66% of the Chinese population. Loveless went further, essentially arguing that OECD's endorsement of the Shanghai results was an implicit endorsement of China's *hukou* policy of forcing rural migrants' teen age children to return to their provinces of origin to attend school. That policy has since been "reformed," but Loveless and Harvey's argument that Shanghai scores were greatly biased upward rather than genuinely representative of all Shanghai's 15 year-old students' "true" performance proved valid. PISA's director Andreas Schleicher ultimately admitted before a British House of Commons Education Select Committee that the Shanghai sample only represented 73% of that province's 15 year-olds—this after a year of denying any problems with the Shanghai results.[11]

The controversies over the validity of international tests as measures of students' knowledge and the representativeness of (PISA) samples reveal an important aspect of these tests. Not surprisingly, the agencies producing (and selling) them have a vested interest in defending them against all critiques, even when those critiques prove to be correct. The OECD in particular has consistently pulled out all stops in defending even the most indefensible uses of "national" test score rankings, such as publishing and touting the results for an unrepresentative sample from one Chinese city when the test was also applied in many other Chinese provinces but not published. If the validity of these tests comes into question, what is the reason for nations to pay dearly to participate in them and to know their results? As critics have learned, transparency beyond rather opaque technical appendices to reports is not in a testing agencies' interest.

Are PISA scores indicators of economic growth potential?

A third theme of discussion on international test score comparisons in the past two years is why or whether they are at all relevant for anything besides knowing how students in various places have performed on a particular test. The underlying argument, pushed by the Hoover Institution's Eric Hanushek and recently joined by Paul Peterson and Ludgar Woessmann, is that average national mathematics test scores are the single best predictor of national economic growth in the period 1960-2010.[12] Thus, when Arne Duncan and others—for example, Marc Tucker[13] —argue that U.S. students test far below our "competitors" in Korea, Shanghai, and Singapore,[14] they are explicitly drawing on the

Hanushek et al. premise that our mathematics score is an accurate predictor of future economic progress.

A puzzling aspect of the broad acceptance of the mathematics-test score-as-predictor-of-U.S.-economic-performance argument is that Hanushek himself defines the United States as an exception to this "rule."[15] U.S. students have participated in international tests since they were first tried in the mid-1960s with the First International Mathematics Study (FIMS), and so there is a long history of results in international mathematics performance comparisons. In the FIMS, a large sample (6,700) of U.S. 13 year-olds in 395 schools—both numbers larger than today's TIMSS or PISA samples—scored next to last (Sweden was lower), a full standard deviation behind Japanese students and lower than students in 9 of the 12 countries that took the 70 item test. In the 1987 Second International Mathematics Study (SIMS), U.S. 13 year-old students did reasonably well on arithmetic, algebra, and descriptive statistics items, but near the bottom in answering correctly on geometry and measurement items. However, overall, students in countries such as Japan, Netherlands, Belgium, France, and Hungary did better than students in the United States.[16]

Ravitch's position reflects a growing critique of Hanushek-Peterson-Woessmann premise that the United States is doomed to a future of slower progress because of its low PISA mathematics scores

Diane Ravitch correctly claims, "U.S. students have never been top performers on the international tests. We are doing about the same now on PISA as we have done for the past half-century."[17] Based on an article by Keith Baker,[18] Ravitch also argues that for the 12 nations that took the FIMS in the mid-1960s, there was "no relationship between a nation's economic productivity and its test scores. Nor did the test scores bear any relationship to quality of life or democratic institutions. And when it came to creativity, the U.S. 'clobbered the world,' with more patents per million people than any other nation."[19] Ravitch's position reflects a growing critique of Hanushek-Peterson-Woessmann premise that the United States is doomed to a future of slower progress because of its low PISA mathematics scores: "Never do they [Hanushek et al.] explain how it was possible for the U.S. to score so poorly on international tests again and again over the past half century and yet still emerge as the world's leading economy, with the world's most vibrant culture, and a highly productive workforce."[20] Norman Eng makes this argument in a different form. "Learning in school," he writes, "is largely characterized by narrow, detached, and contrived experiences, whereas work—especially the highly skilled jobs that drive the economy—incorporates more active, cross disciplinary, and out-of-the box thinking."[21]

Eng's view is supported by a study recently published in *Science*. It concludes that while Chinese freshmen college students score higher than U.S. college freshman on tests of physics content knowledge, there is no difference in terms of scientific reasoning. "The results suggest that the large differences in K–12 STEM education between the United States and China do not cause much variation in students' scientific-reasoning abilities. The results from this study are consistent with existing research, which suggests that

current education and assessment in the STEM disciplines often emphasize factual recall over deep understanding of science reasoning."[22]

Another point raised in this discussion of the link between test scores and economic development is that the average U.S. student does not stop learning at the age of 15, but continues on to higher education. For the almost 40% of U.S.15 year-olds who complete four-years of college, the additional gain in academic skills is higher than for college students in East Asia (and Europe) because of the higher quality of U.S. universities and the greater emphasis placed in the U.S. on post-high school learning. There is evidence, for example, that between the first and third year of their college education, U.S. engineering students reduce the difference between their average score on an Educational Testing Service (ETS) critical thinking test and the average score achieved by Chinese engineering students by half.[23]

There is a growing critique of the mathematics test score-economic growth link coming from a somewhat different direction, typified by Andrew Hacker's recent article in the *New York Review of Books*,[24] pieces by Ross Eisenbrey and Norman Matloff from the Economics Policy Institute,[25] and research on skill gaps, shortages, and mismatches by Peter Cappelli.[26] This critique centers on the alleged shortage of science, technology, engineering, mathematics (STEM) and other education-related skills in the U.S. labor market, certainly the underlying premise of the political hysteria surrounding U.S. students' low mathematics scores. In his review of Michael Teitelbaum's, *Falling Behind?* (2014), Hacker writes:

> ...*Falling Behind?* makes a convincing case that even now the U.S. has all the high-tech brains and bodies it needs, or at least that the economy can absorb. Teitelbaum points out that "US higher education routinely awards more degrees in science and engineering than can be employed in science and engineering occupations." Recent reports reinforce his claim. A 2014 study by the National Science Board found that of 19.5 million holders of degrees in science, technology, engineering, and mathematics, only 5.4 million were working in those fields, and a good question is what they do instead. The Center for Economic Policy and Research, tracing graduates from 2010 through 2014, discovered that 28% of engineers and 38% of computer scientists were either unemployed or holding jobs that did not need their training.[27] (Hacker, 2015, p. 33).

The more likely reason for pushing the notion that the U.S. is short on STEM talent is that U.S. high tech companies want greater leeway in bringing in STEM workers from abroad (primarily India) or keeping foreign students with U.S. earned PhDs here on H-1 visas. Such H-1 visa workers tend to be locked into jobs at lower salaries. The notion that they are smarter than available U.S. workers or bring talents not available in the U.S. turns out to be a myth. They earn lower salaries, are less likely to work in R&D, and, when graduated from U.S. institutions, register fewer patents and generally have PhDs from less prestigious universities than their U.S.-born counterparts working in high-tech.[28]

Cappelli's more general analysis of education-related skills comes to the same conclusion: "There is very little evidence consistent with the complaints about skills and a wide range of evidence suggesting that they are not true. Indeed, a reasonable conclusion is that over-education remains the persistent and even growing situation of the US labor force with respect to skills."[29]

The mathematics test score-economic growth proponents can argue that higher average mathematics scores are a good predictor of general productivity, not just of how many students are able to get college degrees in STEM fields. But the earnings-math score relation, while statistically significant and positive, is surprisingly small. In the most cited study, by Murnane, Willett, and Levy,[30] a standard deviation higher math score in the early 1990s was associated with a 9% higher wage. More recent research, by Castex and Dechter,[31] uses data from the 1980s and 2000s and shows that during these two decades the return to cognitive ability declined by 30% to 50% for men and women and that returns to years of education increased. They argue that the decline in returns to ability can be attributed to differences in the growth rate of technology between the 1980s and 2000s.

To put this in perspective, if average U.S. PISA math scores in 2012 (481) were to equal Finland's (519), and if earnings were a good proxy for productivity, productivity/earnings in the U.S. would increase by about 3% if we assume the 1980s returns to higher ability, but only about 2% or less if we assume the lower estimates for the 2000s. If Korean math scores were the target, productivity/earnings in the U.S. would go up by 6% or 3-4%, depending on the assumed returns. These are not mind-boggling gains. Moreover, the U.S. 8th grade math scores on the Long Term Trends (LTT) survey of the National Assessment of Education Progress (NAEP), which we discuss later in this report, increased about 0.6 of a standard deviation in the past 34 years (1978-2012), perhaps playing a role in greatly increased worker productivity in this period—but not resulting in any significant increase in average real wages or weekly earnings.[32] Meanwhile, profit rates have skyrocketed. If this is what the math score-economic growth proponents have in mind, then our main objective as a nation in increasing mathematics scores is to increase company profits...but not necessarily worker wages. Only a deep cynicism would make that a convincing case for educational reforms targeting increased math scores.

Poorly drawn policy lessons from international test comparisons

A fourth theme that has emerged from international test comparisons is the validity of the educational policy conclusions that pundits, politicians, educators, and especially the OECD draw from them. Carnoy and Rothstein argued in their 2013 report[33] that national average test scores are poor measures for developing educational policy because they often reflect large differences in the socio-economic background of students taking the test in various countries. Students from low-academic resource backgrounds may score higher in one country compared to others of similar background in other countries, and may score steadily higher on the PISA or TIMSS over time, as low academic resource students do in the United States. Students from higher socio-economic background may score lower than

similar students in other countries and make only small or no gains over time. That should produce a different set of policy recommendations from the case of students from higher resource backgrounds doing relatively well and making gains over time while students from low socioeconomic background do relatively poorly compared to similar students in other countries and make small gains. Average score differences may not distinguish between these two cases.

The OECD and others not only tend to ignore such patterns of relative scores and the patterns of change across time, but leap to the very broad conclusion that the reason test scores are high in some countries and low in others is due to particular educational policies. Based on the PISA 2012, for example the OECD produced two volumes that focused on developing policy recommendations. The first of these volumes, *What Makes Schools Successful? Resources, Policies and Practices*, recommends, for example, supporting disadvantaged schools, attracting more qualified teachers and principals, investing more in pre-primary education, and allowing greater teacher and school autonomy.[34] These are generally "appealing" policies, but they are, at best, correlational, and in some, the correlation is not strong. The second of the volumes, *Strong Performers and Successful Reformers—Lessons from PISA 2012 for the United States*, turns its attention to the United States, arguing that the U.S. cannot blame its relatively low performance on the socio-economic characteristics of its students and recommending the policies derived from the *What Makes Schools Successful?* volume.[35]

More generally, the main line coming out of international comparisons is to "copy" the policies of higher scoring countries. Because students in in East Asian countries, such as Korea, Japan, Singapore, and, most recently, Shanghai achieve such high levels of test scores, the OECD and the media consistently feature these countries as having exemplary educational systems. Some reasons given for educational quality in East Asia are the high level of teacher skills, particularly in mathematics, high teacher pay, and, in some countries, such as Korea, rather equal distribution of students from different social class backgrounds across schools. Others have similarly argued that test scores tend to be higher, on average, in countries with more equal income distribution.[36] Again, these "reasons" are at best correlational and are not based on causal analysis. Even more suspect is the notion that the higher test scores are the result mainly of school quality rather than the massive amount of out-of-school tutoring and test-prep taken by East Asian students.[37]

Families in some cultures are more likely to put great emphasis on academic achievement, particularly on achievement as measured by tests. They act on this emphasis by investing heavily in their children's out-of-school academic activities, including in "cram courses," which focus on test preparation.[38] In a world that puts high value on test scores, perhaps such intensive focus on children's academic achievement should be applauded. However, whether it is a good choice for middle and high schoolers to spend most of their waking hours studying how to do math and science problems, and whether it is likely that families in other societies would buy into this choice for their children, are highly controversial issues and certainly only somewhat related to the quality of schooling taken by students in a particular society.

Until 2012, when its students' test scores declined sharply, Finland was also touted as having a model educational system, mainly because of its highly trained teachers and the autonomy teachers and principals in Finland allegedly have in their classrooms and schools. Teacher education and classroom teaching in Finland indeed seem very good, but neither the OECD nor the Finns ever offered any evidence approaching causality to support these claims.

It is especially difficult to make any empirical inferences as to the positive effects of teachers on student achievement using PISA data because PISA does not include a teacher survey. Its scant information on teachers is drawn from school average data found on the principal questionnaire. The claims about the positive effects of school autonomy are also suspect—first, because they are only correlational and second, because autonomy is estimated as an index based on a few questions in the principal questionnaire, not on any externally validated measure of whether principals and teachers actually make independent educational decisions regarding school processes.

Germany and Poland are particularly interesting cases. Students in those two countries scored near the OECD average in 2003 and have made large gains since. In our 2012 report, we cited German empirical studies showing that the gain in test scores from 2000-2009 came from gains made by children from Slavic country immigrant families. The gains of ethnic German students were negligible.[39] An OECD report on lessons for the U.S. from other countries discusses German reforms but concedes that there seems to be no empirical link between those reforms and German test score gains.[40] One study of the Polish reform argues that Poland's 1999 reform postponing vocational tracking from the 9th to the 10th grade lifted by one standard deviation the PISA reading scores of students who would have gone to vocational school in the 9th grade. The study argues that this explains much of Polish reading test score gains in 2000-2006.[41] Yet, thanks to a special follow up sample in Poland, that same study is also able to show that in 10th grade, the 9th grade cohort entering the vocational education track "lost" the gains they had made in 9th grade.[42]

In sum, cross-sectional surveys such as the TIMSS and PISA are not amenable to estimating the causal effects of school inputs on student achievement gains.[43] Neither follows individual students over time as they proceed through the school system, so we cannot tell how the gains in each grade are related to the school resources the student faced in that grade. The PISA survey is an even worse candidate than the TIMSS for drawing conclusions about which educational factors contribute to higher student scores. PISA is not a classroom-based survey such as the TIMSS, so no connection can be made between the student taking the PISA test and his or her teacher. Further, PISA does not survey teachers, so no data are available on their characteristics or their teaching methods. PISA asks students about the kind of teaching and curriculum they experienced, but these data are not related to a particular year of school.

Nevertheless, lack of causal evidence has not stopped the OECD and others from drawing conclusions from PISA data on what works to increase student test scores. As noted, the OECD has published several reports recommending what countries and even schools

should do to increase their students' learning even though there is no causal evidence for these claims. The latest of these makes general recommendations to U.S. policymakers from the PISA data about how to improve U.S. education.[44] Yet, out of 55 countries that have taken the PISA mathematics test over a number of years, only 18 trend upward, and of these 18, about one-half are low-scoring. The OECD has focused heavily on the high scoring countries and big gainers, but it has failed to explain why students in countries with "good" school systems, such as Finland, Australia, New Zealand, Canada, Belgium, Hungary, Czech Republic, and Sweden, did worse on the PISA mathematics test in 2012 than in 2003.[45]

Recent Developments

Should testing agencies be doing their own policy analysis and making policy recommendations?

The Brookings Institution's Tom Loveless has, as noted, raised critiques of the Shanghai PISA survey and their inappropriate use by the OECD to promote Shanghai's educational reforms and, implicitly, to suggest that the results reflect high quality throughout China's educational system. Based on these critiques, he has also proposed several reforms of PISA itself.[46] The reforms apply to the way PISA is governed and how policy recommendations are derived from the PISA results. Loveless holds up the U.S. National Assessment of Educational Progress (NAEP) and the International Association for the Evaluation of Educational Achievement (IEA)—the parent organization for the TIMSS—as governance models for the PISA. The governing bodies of those two organizations include non-government representatives such as statisticians, educators, and policy analysts who have no vested interest in how results are presented.

Loveless also argues that PISA, like the National Center for Educational Statistics (NCES)—the U.S. government agency that oversees the NAEP—should "separate the policy analysts from the data collectors." The NCES administers the NAEP and presents the results. NCES officials do not speculate why some states do better than others, nor do they make policy recommendations to states on how they can improve their students' performance. In contrast, Loveless points out,

> The same subunit of OECD plans and administers PISA, analyzes the data, and makes policy recommendations. PISA data releases are accompanied by thematic volumes. The 2012 data, for example, were joined by volumes on equity, student engagement, and the characteristics of successful schools. The title of the 2012 volume on school characteristics reveals its ambitions: "What Makes Schools Successful?[47]

As we have already noted, this same OECD subunit has produced two reports, in 2011 and 2013, advising the United States on how to improve its educational system, again based on data produced by the subunit itself.[48] Further, as noted, these policy recommendations are all based on analyses of test results for a cohort of students at a single point in time, or on

time trends of test results for different cohorts of students, each at a single point in time. In Loveless' words, "Skilled policy analysts are cautious in making policy recommendations based on cross-sectional data because they provide weak evidence for policy guidance."[49]

Loveless raises another point that bears further discussion. "It is strange," he says, "that the U.S. participates in an international test that violates the constraints it imposes on its own national assessment." Strange, indeed. Not only does the U.S. Department of Education participate in the PISA, it fosters the violation of the constraints imposed in its own NCES by asking the OECD for reports recommending educational reforms in the U.S. The use of an "outside" agency (in which the U.S. DOE participates) to issue policy recommendations to promote Secretary Duncan's educational agenda is not that different from Loveless' description of the OECD's role in promoting the Shanghai Municipal Education Committee's education reforms. Zhang Minxuan, the Shanghai coordinator for PISA, was also the Committee's vice-director general and therefore deeply involved in the reforms.

Of course, in the Shanghai case (as in Finland, Estonia, and Poland), the OECD's role is to laud educational policies because of high PISA scores, whereas in the U.S., the OECD supports Secretary Duncan's endless criticism ("stagnation," "failure") of U.S. education. After more than six years at the helm, Secretary Duncan receives no respite from U.S. PISA performance. Yet, he seems to find new ways to use the PISA scores and OECD reports to diminish evidence of high levels of U.S. performance where they exist and of the gains actually being made. It is difficult to explain such behavior, even in terms of a motivated political bureaucrat trying to use "tough love" to exhort positive change in a huge, highly decentralized organization.

The case for U.S. states looking to each other to improve education

In addition to all the other problems of drawing accurate, transferable educational policy lessons from other countries, or of even arguing that school system quality is the main (causal) reason that students in some countries score higher on a particular test than students in other countries, there is the issue of whether large, federally-run educational systems such as the United States' or Germany's or Brazil's should be discussed in national terms. Is there such a thing as U.S. education? Are statements identifying U.S. national scores on the PISA or TIMSS as "U.S. educational performance" just a self-serving construction of the international testing agencies and the U.S. Department of Education?

There are commonalities across the many educational administrative entities in the U.S. But there are also differences, even among the 50 states plus D.C. In addition, many school districts operate rather independently of states, adding more complexity to the system. Some of these districts, such as Los Angeles, New York, and Chicago, have more schools and greater student enrollment than do some countries taking the PISA and TIMSS.

Should policymakers turn away from international comparisons and toward the U.S.' own state-based systems to gain insights into school system improvement? The case for looking

inward, across states, within the United States, is compelling. On international tests, student performance and performance gains in, say, mathematics vary greatly among U.S. states. Nine states took the 2011 TIMSS and three states took the 2012 PISA. In some of these states, such as Massachusetts and Connecticut on the PISA, and in seven of nine states that took the TIMSS, higher socioeconomic class students scored as high or higher in mathematics and much higher on the PISA reading test than similarly advantaged students from European countries and Canada. Based on NAEP data and the link between the 2011 NAEP and the 2011 TIMSS,[50] if students in Texas, Vermont, and North Carolina had taken the PISA in 2012, they, too, probably would have scored very high. In other states, such as Florida on the PISA and California and Alabama on the TIMSS, higher socioeconomic class students scored far below similarly advantaged students in higher scoring states and other countries.

A second argument for learning from our own states' experiences is that the conditions and context of education are more similar among U.S. states than between the United States and other countries. Teacher labor markets are not drastically different, and the state systems are regulated under the same federal rules. If students in some states make much larger gains than in other states, the policies that produced those larger gains were implemented in the same general "national political context" as in states where students make lower gains, although the political context does vary from state to state. There are far greater contextual differences between countries, including the varying role that out of school tutoring and "cram" courses play in many countries,[51] as well as the national emphasis placed on international test performance.[52]

Further, some countries' curriculum and evaluative test instruments are gradually adapted to teach to international tests, particularly to the PISA test, which has many items that do not resemble most countries' existing evaluation instruments or curriculum.[53] The introduction of the Common Core in the United States and the new tests developed to evaluate students in states adopting the Common Core are good examples of tilting toward the PISA test. This can raise scores on one type of test but lower scores on another—Finnish students' contrasting performance on the PISA and TIMSS is a good example.

In addition, it is puzzling that Finland should have become such a mecca for U.S. educational reformers because of their PISA performance, when middle and higher family resource students in more diverse and more populous Massachusetts scored as high or higher than the Finns in the PISA 2012 mathematics and reading tests.

A third argument for learning from our own states rather than from other countries is that large gains seem to have made in many states since the 1990s, particularly in mathematics. Students in five U.S. states took the TIMSS test in the 1990s and again in 2011. Of these, students in Massachusetts, Minnesota, and North Carolina made very large mathematics gains compared to cohorts in Finland and Korea, and as large or larger than students in England. Twenty years ago, based on international test score comparisons, researchers analyzing the 1995 TIMSS test and its results were able to argue convincingly that students in the U.S. did poorly on the 8th grade TIMSS mathematics test mainly because only one-quarter of U.S. students were exposed to algebra and even less to geometry by the 8th

grade.[54] They also argued that the U.S. math curriculum was a "mile wide and an inch deep."[55] This was a good use of international comparisons to make policy changes in mathematics. Yet, since 1995, some states have reformed their mathematics education, making for large gains in student performance. It also turns out that whereas all groups have made large gains, in some states, such as Indiana, the largest gains appear to be among disadvantaged students, and in others, such as Connecticut, the gains appear to be largest for advantaged students. In three states—Minnesota, Massachusetts, and North Carolina—large gains are spread across all family academic resource groups. Differences in reforms implemented by states in this period may reveal important lessons for policymakers.[56]

Although there is a strong logic to moving in this direction, it does not seem to be resonating with our own Secretary of Education. He dismisses Massachusetts' students' high performance because "... the percentage of high-performing students in Massachusetts—the U.S.'s highest-performing state—is dwarfed by the percentage of advanced students in top-performing systems, such as Shanghai, Singapore, and Korea. In math, 19 % of Massachusetts' students are high-performers. But in Shanghai, in China's top-performing system, 55 %of students—almost triple Massachusetts' rate—are high-performers in math."[57]

It is quite a stretch to argue that the quality of Massachusetts' educational system should be judged on the basis of how well all students in Massachusetts perform on the PISA test compared to the performance of a non-representative sample of students in Shanghai, or compared to perhaps more representative samples in Singapore and Korea. These comparison groups in Asia have been shown to invest many more hours per day and vastly larger amounts of family resources studying mathematics outside school than do students in the United States.[58]

Nevertheless, there is one group in Massachusetts that may be comparable to students in Shanghai, Singapore, and Korea in terms of such out-of-school activities: self-identified Asian-origin students. Students who identified themselves as of Asian origin in the Massachusetts PISA sample scored 569 in the 2012 PISA mathematics test, significantly higher than the average in Korea (554), and about the same as students in Singapore (573).[59] One possibility would be that Asian-origin students in Massachusetts are from families that have greater family academic resources than students in Singapore or Korea. But available data show otherwise. The distribution of students in the Korean and Singapore PISA samples across books in the home categories suggests that there are more low family academic resource Asian-origin students in Massachusetts than in Korea or Singapore. About the same proportion of Asian-origin students in Massachusetts as in Korea and Singapore are from high academic resource family as measured by books in the home. When we measure family academic resources by parents' highest level of education, about 60 %of the Asian-origin students in Massachusetts report that their parents only had middle or high school education—much higher than in Korea and Singapore. Only 23 %of Asian-origin students in Massachusetts reported parents with college education, whereas in Korea, it was 53 %and in Singapore, 30 percent.

Those who would use international test score results to bash U.S. education should take note of all the favorable data about U.S. education that can be drawn not only from Massachusetts, but from a number of other states. These are not random factoids. They reflect real educational policies that have worked, and worked in places as different as Texas, Massachusetts, Minnesota, Vermont, and North Carolina. If we extend our analysis to national test data (NAEP), we also find big gains in places where students began with very low scores, such as D.C., Hawaii, and Louisiana. There are many more lessons to be learned about improving U.S. education from their experiences than from Shanghai.

Discussion and Conclusions

Rankings of countries' educational systems based on international test scores and policy lessons drawn from high scoring countries' educational systems have taken on a life of their own, turning some national educational systems into superstars, to be admired and flooded with educational tourists, and other national systems, such as the United States', into sad sacks, criticized and mocked for being "stagnant" and "failing." Lurking behind all this adulation or condemnation is the alleged link between current international test scores and future national economic performance—a future of prosperity versus a future of stagnation.

Our review of the critiques of the claims surrounding international tests and the future prospects for countries that do well or poorly on these tests suggests that much of the international testing enterprise and its ideological influence has the substance of a house of cards.

To start, the critiques undercut a number of fundamental premises of the average national test score rankings. First, educational "success" needs to be based not on average reported test scores, but on test scores adjusted for the family academic resources of the students taking the tests and on the gains made by different groups of students in each country—or, at the least, gains adjusted to account for varied family resources available to students within each national sample. Second, the error terms in the test scores are underestimated, suggesting that if rankings were based on more accurate error terms, substantial changes in rankings could result. Third, in the case of the OCED and its PISA test, the conflation of international politics and country rankings—as epitomized by OECD misrepresentations of the randomness of the Shanghai PISA sample and of Shanghai scores as representative of China—raises serious questions about the validity of the rankings and how they are used to promote educational policy.

At another level, the critiques raise questions about the meaning and importance of international test comparisons for anything beyond the comparison of how students in various countries score on a particular test. First, claims that the average national scores on mathematics tests are good predictors of future economic growth are, at best, subject to serious questions and, at worst, gross misuse of correlational analysis. The U.S. case appears to be a major counterexample to these claims (Japan is another). Second, the use of data from international tests and their accompanying surveys have limited use for

drawing educational policy lessons. Yet, again in the case of the OECD, there seem to be no end of policy lessons—none with appropriate causal inference analysis, many based on questionable data, and others largely anecdotal—proposed by the same agency that developed and applied the test.

The questions raised about the validity of international test comparisons, about the way test results are used for policy analysis and recommendations, and about the use of high or low test results as a base for extolling or bashing educational systems as the main reason for student performance have produced a new set of critiques.

These new critiques question: (a) the conflict of interest role of the OECD (and its member governments) in acting simultaneously as testing agency, data analyst, and interpreter of results for policy purposes; and (b) the relevance of national test score comparisons for large federal educational systems such as the United States, with its 51 (including the District of Columbia) highly autonomous geographic educational administrations.

An example of the importance of these critiques is U.S. student performance in mathematics on international tests. There is ample reason to agree with international testing proponents that, on average, the U.S. educational system could be improved to teach students mathematics better. As we have noted, studies from the 1995 TIMSS test were able to make a convincing empirical argument that students in the U.S. did poorly on the 8th grade TIMSS mathematics test mainly because only one-quarter of U.S. students were exposed to algebra and even fewer to geometry by the 8th grade.[60] This is a good example of an influential policy analysis undertaken by researchers *independent of the organization* that applied the test.

Unfortunately, all the valid critiques of international testing…are not going to make those tests go away.

Yet, even though many independent researchers use PISA data to try to find student practices and school inputs that lead to improved student outcomes, almost all analyses of the PISA data and educational policy recommendations come from the large group of analysts, the publications, and the public relations machine of the OECD itself. The OECD chooses the policy lines it wants to present, produces the correlational analysis to support them, and then promotes its policy conclusions through the governments that support the entire enterprise. As Tom Loveless accurately points out in the case of the OECD's analysis of the alleged effect of pre-school education on PISA performance, these analyses are usually statistically biased and their policy conclusions misleading.[61]

Another contrast to draw from the earlier analysis of TIMSS regarding U.S. math performance is that since 1995, a number of U.S. states have reformed their mathematics education, producing large gains in student scores. Rather than comparing unadjusted average mathematics performance for U.S. students with other countries' on the single PISA test, we could base comparisons of U.S. student performance on the TIMSS mathematics test, adjusted for variations in family academic resources (FAR) in the

student sample. As we have noted, those scores increased considerably since 1995 and 1999[62] —and we could compare the test score changes for various states administering the TIMSS in 1999 and 2011, again adjusting for FAR variations within the state student samples in those two years. We note that students in Minnesota made a 45 point gain in FAR adjusted TIMSS math scores in 1995-2011, the same as students in the U.S. as a whole—a gain as large as that in England and larger than the gain in Korea. In 1999-2011, students in Massachusetts and North Carolina made even larger 54 and 52 point gains (about one-half a standard deviation), while gains by Indiana and Connecticut students substantively equaled gains by U.S. students as a whole (20 points, or 0.2 standard deviations). The Massachusetts and North Carolina student math gains were far larger than gains in Finland (-5 adjusted points), England (24 adjusted points), and Korea (16 adjusted points) during the same period.

From the standpoint of U.S. policymakers, it seems much more relevant and interesting to understand the policies states such as Massachusetts, North Carolina, and Minnesota implemented in the past 20 years to promote such high mathematics gains than to examine other countries' educational policies—if, indeed, it is their educational policies—behind large test score gains on either the PISA or the TIMSS during the decade of the 2000s. We have discussed Germany and Poland's gains on the PISA test as cases in point.

Recommendations

Unfortunately, all the valid critiques of international testing—particularly of using those tests to judge the quality of educational systems and, even worse, of claiming that test scores of 8th grade students or 15 year olds in school are good predictors of a country's (or a state's) future quality of life—are not going to make those tests go away. Neither will these critiques stop pundits and politicians from misusing test results nor stop international agencies such as the OECD from trying to shape educational policies with statistically biased empirical analyses.

Nevertheless, there are changes that could be made to reduce misuse. Based on this review, it is recommended:

- PISA and TIMSS should report all international test results by FAR (family academic resource) subgroups of students with different levels of resources such as books in the home or mother's education. Relatedly, PISA and TIMSS should report all changes in test scores over time by FAR subgroups. Compare country results by student FAR subgroup together with country aggregate averages.

- The OECD and the IEA should make the individual-level student micro-data for the PISA and TIMSS tests available at the same time as the results are formally announced. This would allow international researchers to produce independent analyses of the data within a week of the time when the OECD's and IEA's versions of the results appear.

- Beyond allowing access to individual-level student micro-data immediately, the OECD should separate international tests' design, application, and results from data analysis and policy recommendations derived from the tests and surveys. The OECD should also include independent academic expert appointments to PISA's decision-making board, which governs the application and use of test data, as is the practice at the IEA for the TIMSS test.

- In the United States, the National Center of Educational Statistics should publish PISA reading, mathematics and science scores and TIMSS mathematics and science scores by FAR group, particularly FAR-adjusted PISA and TIMSS scores over time. There will be a golden opportunity to do this for the PISA and TIMSS together in December 2016, when the results of the 2015 PISA and TIMSS will be announced in the same month.

- In the United States, policymakers should shift focus away from why students in other countries may do "better" than U.S. students as a whole and instead focus on why student achievement gains have been greater in some U.S. states and lower in others.

Notes and References

1 PISA is sponsored by the Organization for Economic Cooperation and Development (OECD). See http://www.pisa.oecd.org/ and http://nces.ed.gov/surveys/pisa/. *PISA was administered to 15-year-olds in 2000, 2003, 2006, 2009, 2012.*

2 TIMSS was administered by the International Association for the Evaluation of Educational Achievement (IEA) to 8th-graders in 1995, 1999, 2003, 2007, and 2011. See http://timss.bc.edu/ and http://nces.ed.gov/timss/. An international test of reading, the Progress in International Reading Literacy Study (PIRLS), was administered only to 4th-graders in 2001, 2006, and 2011. TIMSS was also administered to 4th-graders simultaneously with the 8th-grade administration. We do not discuss 4th-grade scores, either from PIRLS or from TIMSS, in this report.

3 Duncan, A. (2012). Statement by U.S. Secretary of Education Arne Duncan on the release of the 2011 TIMSS and PIRLS assessments. U.S. Department of Education, December 11. http://www.ed.gov/news/press-releases/statement-us-secretary-education-arne-duncan-release-2011-timss-and-pirls-assess. Retrieved June 10, 2015.

4 Duncan, A. (2013). The threat of educational stagnation and complacency: Remarks of U.S. Secretary of Education Arne Duncan at the release of the 2012 Program for International Student Assessment (PISA)." U.S. Department of Education, December 3. http://www.ed.gov/news/speeches/threat-educational-stagnation-and-complacency. Retrieved July 20, 2015.

5 Duncan, A. (2013). The threat of educational stagnation and complacency: Remarks of U.S. Secretary of Education Arne Duncan at the release of the 2012 Program for International Student Assessment (PISA)." U.S. Department of Education, December 3. http://www.ed.gov/news/speeches/threat-educational-stagnation-and-complacency. Retrieved July 20, 2015.

6 Carnoy, M. & Rothstein, R. (2013). *What do international tests really show about American student performance?* Washington, D.C.: Economic Policy Institute.

7 Stewart, W. (2013). Is PISA fundamentally flawed? *Times Education Supplement Magazine,* July 26, updated September 27, 2014. https://www.tes.co.uk/article.aspx?storycode=6344672. Retrieved June 30, 2015.

8 Stewart, W. (2013). Is PISA fundamentally flawed? *Times Education Supplement Magazine,* July 26, updated September 27, 2014.

9 Loveless, T. (2013). PISA's China problem. Brookings Institution, Brown Center Chalkboard. October 9. http://www.brookings.edu/research/papers/2013/10/09-pisa-china-problem-loveless. Retrieved June 30, 2015; Loveless, T. (2014). Lessons from the PISA-Shanghai controversy." Brookings Institution, Brown Center Chalkboard. March 18. http://www.brookings.edu/research/reports/2014/03/18-pisa-shanghai-loveless. Retrieved June 30, 2015.

10 Harvey, J. (2015). Ten things you need to know about international assessments. *Washington Post*, February 3.

11 Stewart, W. (2014). More than a quarter of Shanghai pupils missed by international Pisa rankings. *Times Education Supplement*, March 6. https://www.tes.co.uk/news/school-news/breaking-news/more-a-quarter-shanghai-pupils-missed-international-pisa-rankings. Retrieved July 20, 2015.

12 Hanushek. E., Peterson, P., and Woessmann, L. (2013). *Endangering prosperity*. Washington, D.C.: Brookings Institution Press.

13 Tucker, M. (2015). Are we just fooling ourselves? Is American education a colossal failure? *Education Week*. April 16.

14 Duncan, A. (2012). Statement by U.S. Secretary of Education Arne Duncan on the release of the 2011 TIMSS and PIRLS assessments. U.S. Department of Education, December 11. http://www.ed.gov/news/press-releases/statement-us-secretary-education-arne-duncan-release-2011-timss-and-pirls-assess. Retrieved June 10, 2015.

Duncan, A. (2013). The threat of educational stagnation and complacency: Remarks of U.S. Secretary of Education Arne Duncan at the release of the 2012 Program for International Student Assessment (PISA).” U.S. Department of Education, December 3. http://www.ed.gov/news/speeches/threat-educational-stagnation-and-complacency. Retrieved July 20, 2015.

15 Hanushek. E., Peterson, P., and Woessmann, L. (2013). *Endangering prosperity*. Washington, D.C.: Brookings Institution Press.

16 Medrich, E. & Griffith, J. (1992). *International mathematics and science assessment: What have we learned?* Washington, D.C.: U.S. Department of Education, National Center for Educational Statistics, Office of Educational Research and Improvement. NCES 92-011.

17 Ravitch, D. (2013). My view of the PISA Scores. Diane Ravitch’s Blog, p. 2.

18 Baker, K. (2007). Are international tests worth anything? *Phi Delta Kappan*, 89 (2), 101-104.

19 Ravitch, D. (2013). My view of the PISA Scores. Diane Ravitch’s Blog, p. 2.

20 Ravitch, D. (2013). My view of the PISA Scores. Diane Ravitch’s Blog, p. 3.

21 Eng, N. (2014). Should U.S. panic over latest international creative problem-solving test scores? *American School Board Journal*, May 7, p. 1. http://www.asbj.com/HomePageCategory/Online-Features/ReadingsReports/BonusArticles/Should-US-Panic-Over-Latest-International-Creative-Problem-Solving-Tests-Scores.pdf

22 Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., Liu, Q., Ding, l. et al. (2014). Learning and scientific reasoning. *Science*, 323 (5914), 586-587.

23 Loyalka, P., Kardanova, E., Liu, L., Novdonov, V., Shi, H., Enchicova,

K., Johnson, N., & Mao, L. (2015). Where are the skilled engineers coming from? Assessing and comparing skill levels and gains in engineering programs across the US, China, and Russia. Stanford University, Working Paper.

24 Hacker, A. (2015). The frenzy about high-tech talent. *New York Review of Books*, 62 (12), 33-35.

25 Eisenbrey, R. (2013). America’s Genius Glut. *New York Times*, Opinion Pages. February 7; Matloff, N.. 2013. “Are foreign students the ‘best and brightest’?” Economic Policy Institute Briefing Paper, February 28. http://www.epi.org/publication/bp356-foreign-students-best-brightest-immigration-policy. Retrieved June 30, 2015.

26 Cappelli, P. (2014). Skill gaps, skill shortages, and skill mismatches: Evidence for the US. Cambridge, MA: National Bureau of Economic Research, Working Paper, No. 20382. http://www.nber.org/papers/w20382. Retrieved July 21, 2015.

27 Hacker, A. (2015). The frenzy about high-tech talent. *New York Review of Books*, 62 (12), 33-35, p. 33.

28 Eisenbrey, R. (2013). America’s genius glut. *New York Times*, Opinion Pages. February 7; Matloff, N.. 2013. Are foreign students the “best and brightest”? Economic Policy Institute Briefing Paper, February 28. http://www.epi.org/publication/bp356-foreign-students-best-brightest-immigration-policy. Retrieved June 30, 2015.

29 Cappelli, P. (2014). Skill gaps, skill shortages, and skill mismatches: Evidence for the US. Cambridge, MA: National Bureau of Economic Research, Working Paper, No. 20382. http://www.nber.org/papers/w20382. Retrieved July 21, 2015.

30 Murnane, R., Willett, J., and Levy, F. (1995). The growing importance of cognitive skills in wages. *Review of Economics and Statistics*, 77 (2), 251-266.

31 Castex, G. & Dechter, E. (2012). The changing roles of education and ability in wage determination. University of New South Wales, Australian School of Business Research Paper No 2012 Econ 43. UNSW Australian School of Business Research Paper No. 2012 ECON 43. Retrieved July 24, 2015.

32 *Economic Report of the President*, 2014, Table B-15.

33 Carnoy, M. & Rothstein, R. (2013). *What do international tests really show about American student performance?* Washington, D.C.: Economic Policy Institute.

34 OECD, PISA (2013). *PISA 2012 results: What makes schools successful? Resources, policies and practices (volume IV)*. Paris: OECD, chapter 6.

35 Organization of Economic Cooperation and Development (OECD), Programme of International Student Assessment (PISA). (2013b)*., Lessons from PISA 2012 for the United States, strong performers and successful reformers in education*. Paris: OECD.

36 See also Chiu, M.M. (2015). Family inequality, school inequalities, and mathematics achievement in 65 countries: Microeconomic mechanisms and rent seeking and diminishing marginal returns. *Teachers College Record*, 117 (1), 1-32.

37 Bray, M. (2006) Private supplementary tutoring: comparative perspectives on patterns and implications. *Compare*, 36 (4), 515-530.

38 There is a vast literature on cram school in Korea (*hagwon*), Japan (*juku*), and other Asian countries. We only cite a few references; however, there is no doubt that a high percentage of students in these countries spend a considerable amount of time during their middle school and high school years in cram schools/courses in addition to studying for tests and completing other work for "regular" school. Families invest major resources in extra instruction. Amazingly, this is rarely mentioned when discussing whether such behavior or levels of investment are broadly transferable to other societies. See Bray, M. (2006) Private supplementary tutoring: comparative perspectives on patterns and implications. *Compare*, 36 (4), 515-530; Ripley, A. (2013). *The smartest kids in the world*. New York*:* Simon and Shuster.

39 Stanat, P., Rauch, D. & Segeritz, M. (2010). Schülerinnen und schüler mit migrationshintergrund. In Klieme, E., Artelt, C.,Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W., & Stanat, P. (eds.*), PISA 2009. Bilanz nach einem Jahrzehnt*. Münster, Germany: Waxmann, 200-230.

40 Organization of Economic Cooperation and Development (OECD), Programme of International Student Assessment (PISA). (2011). *Strong performers and successful reformers in education—Lessons from PISA for the United States*. Paris: OECD. http://dx.doi.org/10.1787/9789264096660-en

41 Jakubowski, M., Patrinos, H., Porta, E. E., & Wisniewski, J. (2010). The impact of the 1999 education reform in Poland. Washington, D.C.: World Bank Policy Research Working Paper No. 5263.

42 It is also possible that Poland, like Estonia and a number of other countries, considers improved performance on the PISA test as a means of achieving greater "legitimacy" in the international community, and has gradually reformed its curriculum and internal student evaluation instruments to match the types of questions asked on the PISA test. As students become more familiar with such questions, it would not be surprising that they would perform better. See Carnoy, M., Khavenson, T., & Ivanova, A. (2014a). Using TIMSS and PISA results to inform educational policy: A study of Russia and its neighbors." *Compare*, 45 (2), 248-271.

43 Carnoy, M., Khavenson, T., Loyalka, P., Schmidt, W., & Zakharov, A. (2014). Revisiting the relationship between international assessment outcomes and educational production: Evidence from a longitudinal PISA-TIMSS sample. Graduate School of Education, Stanford University (mimeo).

44 Organization of Economic Cooperation and Development (OECD), Programme of International Student Assessment (PISA). (2013). *Lessons from PISA 2012 for the United States, strong performers and successful reformers in education*. Paris: OECD.

45 Organization of Economic Cooperation and Development (OECD), Programme of International Student Assessment (PISA). (2013). *PISA 2012 results: What students know and can do: student performance in mathematics, reading and science (Volume I)*. Paris: OECD, Figure I.2.16.

46 Loveless, T. (2014). Lessons from the PISA-Shanghai controversy. Brookings Institution, Brown Center Chalkboard. March 18. http://www.brookings.edu/research/reports/2014/03/18-pisa-shanghai-loveless. Retrieved June 30, 2015.

47 Loveless, T. (2014). Lessons from the PISA-Shanghai controversy." Brookings Institution, Brown Center Chalkboard. March 18. http://www.brookings.edu/research/reports/2014/03/18-pisa-shanghai-loveless. Retrieved June 30, 2015.

48 Organization of Economic Cooperation and Development (OECD), Programme of International Student Assessment (PISA). (2011). *Strong performers and successful reformers in education—Lessons from PISA for the United States*. Paris: OECD. http://dx.doi.org/10.1787/9789264096660-en; Organization of Economic Cooperation and Development (OECD), Programme of International Student Assessment (PISA). (2013b)., *Lessons from PISA 2012 for the United States, strong performers and successful reformers in education*. Paris: OECD.

49 Loveless, T. (2014). Lessons from the PISA-Shanghai controversy. Brookings Institution, Brown Center Chalkboard. March 18. http://www.brookings.edu/research/reports/2014/03/18-pisa-shanghai-loveless. Retrieved June 30, 2015.

50 National Center for Educational Statistics. (2013). *The nation's report card: U.S. states in a global context: Results from the 2011 NAEP-TIMSS linking study*. Washington, D.C.: NCES.

51 Bray, M. (2006). Private supplementary tutoring: comparative perspectives on patterns and implications. *Compare*, 36 (4), 515-530; Ripley, A. (2013). *The smartest kids in the world*. New York: Simon and Shuster.

52 Carnoy, M., Khavenson, T., & Ivanova, A. (2014). Using TIMSS and PISA results to inform educational policy: A study of Russia and its neighbors." *Compare*, 45 (2), 248-271.

53 Carnoy, M., Khavenson, T., & Ivanova, A. (2014). Using TIMSS and PISA results to inform educational policy: A study of Russia and its neighbors." *Compare*, 45 (2), 248-271.

54 Schmidt, W.H., McKnight, C.C., Houang, R. Wang, H., Wiley, D. et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

55 Schmidt, W.H., McKnight, C.C., & Raizen, S.A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, The Netherlands: Kluwer.

56 Carnoy, M., Garcia, E., Khavenson, T. (2015). Are differences in student performance among U.S. states more useful for educational policy than international comparisons? Washington, D.C.: Economic Policy Institute (forthcoming).

57 Duncan, A. (2013). The threat of educational stagnation and complacency: Remarks of U.S. Secretary of Education Arne Duncan at the release of the 2012 Program for International Student Assessment (PISA)." U.S. Department of Education, December 3. http://www.ed.gov/news/speeches/threat-educational-stagnation-and-complacency. Retrieved July 20, 2015.

58 Bray, M. (2006). Private supplementary tutoring: comparative perspectives on patterns and implications. *Compare*, 36 (4), 515-530; Wantanabe, M. (2013). *Juku: The stealth force of education and the deterioration of schools in Japan*. Independently Published.

59 The test score results for Asian-origin students in Massachusetts are similarly very high in the 2011 TIMSS and the 2013 NAEP.

60 Schmidt, W.H., McKnight, C.C., Houang, R. Wang, H., Wiley, D. et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

61 Loveless, T. (2014). Lessons from the PISA-Shanghai controversy. Brookings Institution, Brown Center Chalkboard. March 18. http://www.brookings.edu/research/reports/2014/03/18-pisa-shanghai-loveless. Retrieved June 30, 2015.

62 Adjusting the TIMSS 1995 score using 2011 family academic resource (FAR) sample distribution (books in the home), the average U.S. 8th grade math score increased from 466 in 1995 to 510 in 2011. Adjusting the 1999 math score using 2011 FAR sample proportions, the average U.S. math score increased from 490 in 1999 to 510 in 2011. The 2011 U.S. TIMSS score is about the same as in England or Finland, but much lower than Korea's.