International Benchmarking

Mark Schneider
American Institutes for Research

June 2, 2009

Send all correspondence to:
Mark Schneider
American Institutes for Research
1000 Thomas Jefferson St. NW
Washington, DC 20007
Voice: 202-403-5510
E-mail: mschneider@air.org

Despite a growing fascination with international comparisons of student performance and the feedback they provide on how young Americans are doing compared with their age-mates in other countries, current international assessments cannot generate a great deal of reliable policy advice. Indeed, many of the policy conclusions drawn from these assessments seem to be motivated as much by existing ideas as by strong evidence. The latest wrinkle in the nation's interest in these international assessments is the drumbeat for state-level participation in the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Assessment (PISA), with a somewhat weaker chorus asking for state-level participation in the Trends in International Mathematics and Science Study (TIMSS).

Tom Friedman may have made a fortune with his idea that "the world is flat." Waving this same banner, many states may spend a fortune for international assessments they believe, to use the words of the National Governors Association, will "identify policy solutions to U.S. education system shortcomings."[1] Indeed, given the worldwide reach of these assessments, it is hard not to view these data as a rich source for policy advice. But the limits on these assessments are often overlooked.

"International benchmarking," the term used to describe these efforts, has at least two components.[2] The first is the comparison of the relative performance of students in the United

[1] From the news release describing the Governors Education Symposium in June 2008, http://www.nga.org/portal/site/nga/menuitem.6c9a8a9ebc6ae07eee28aca9501010a0/?vgnextoid=68f3b5cd2977a110VgnVCM1000001a01010aRCRD.

[2] As an example of both state interest and the use of the term, in December of 2008, the National Governors Association, the Council of Chief State School Officers, and Achieve released a report, *Benchmarking for Success: Ensuring U.S. Students Receive a World-class Education*, calling for state scores on international assessments.

States versus their peers in other countries. Here, the resulting "league tables" are the center of attention and the results of these tests are often described and often decried as a "horse race"—a fixation on who's winning and who's losing. The second function is identifying activities, usually culled from the practices in high-performing countries, that provide guidance on how to improve student performance.

A QUICK GUIDE TO INTERNATIONAL STUDENT ASSESSMENTS

There are three main international student assessments, known widely by their acronyms.

- **PIRLS (Progress in International Reading Literacy Study)** is an assessment of 4th-grade reading, http://nces.ed.gov/surveys/pirls/
- **TIMSS (Trends in International Mathematics and Science Study)** is an assessment of 4th- and 8th-grade[3] science and math, http://nces.ed.gov/timss/
- **PISA (Programme for International Student Assessment) is an assessment of** reading, math, and science "literacy" among 15-year-olds, http://nces.ed.gov/Surveys/PISA/

In the United States, the results of all three student assessments are often compared to **NAEP** (the National Assessment of Educational Progress, http://nces.ed.gov/nationsreportcard/about/).

Within the United States, the scope of NAEP dwarfs the international assessments. For example, in 2007, the NAEP 8th-grade math assessment involved over 150,000 students in around 7,000 schools. In contrast, the 8th-grade 2007 TIMSS assessed around 7,400 students in fewer than 250 schools. PISA in 2006 involved only 5,600 15-year-olds in around 170 schools.

[3] There is a TIMSS Advanced for 12th-grade science and math, but only a few countries participate.

The difference in size of these assessments is important—in reading and math in the 4th and 8th grade, NAEP can report state-by-state performance every 2 years and, through the Trial Urban District Assessment program, NAEP can also present data on a growing number of large school districts.[4] In contrast, the small size of the TIMSS and PISA samples makes any state-level breakout practically impossible.

HOW DO THESE ASSESSMENTS DIFFER?

PISA is a self-proclaimed "yield study" assessing the total "literacy" of 15-year-olds and is therefore not tied to specific curricula. It also has an emphasis on globalization and 21st century skills and claims to be assessing the skills that young adults will need in the emerging global economy.[5]

According to the OECD:

> PISA seeks to measure how well young adults, at age 15 and therefore approaching the end of compulsory schooling, are prepared to meet the challenges of today's knowledge societies—what PISA refers to as "literacy." The assessment is forward looking, focusing on young people's ability to use their knowledge and skills to meet real-life challenges, rather than merely on the extent to which they have mastered a specific school curriculum. http://www.oecd.org/dataoecd/51/27/37474503.pdf

[4] State-level results in science and writing have also been made available, although not on the regular schedule of math and reading, which are mandated by the No Child Left Behind Act.

[5] In 1997, OECD launched the DeSeCo Project (Definition and Selection of Competencies: Theoretical and Conceptual Foundations, http://www.deseco.admin.ch/) to identify a foundation for assessing skills for the future; PISA has been influenced strongly by this effort. What 21st-century skills are is an open debate. See, for example, Jay Mathew's cogent analysis "The Latest Doomed Pedagogical Fad: 21st-Century Skills" and "The Rush for '21st-Century Skills' New Buzz Phrase Draws Mixed Interpretations From Educators." Both appeared in the *Washington Post* on Monday, January 5, 2009; B02. Also see Elena Sivla's Education Sector report, http://www.educationsector.org/usr_doc/MeasuringSkills.pdf.

Another OECD publication notes that "…PISA sets out to assess certain key competencies that young adults need to possess in order to be successful in tomorrow's global village and global economy."[6] In that same report, the writers argued that "Some countries see themselves as part of the global village and economy and tend to value skills assessed in PISA…Countries that are consciously modernising their educational systems are changing the goals of their education systems and are usually making use of those of PISA. These countries see the development of global citizens that are able to live and work in any country in the world, endorse the skills assessed in PISA and examine ways to incorporate them into their curricula."

Never modest, the report releasing the 2003 PISA was titled *Learning for Tomorrow's World* and the 2006 PISA results *Science Competencies for Tomorrow's World*.

TIMSS is more grade- and curriculum-centered and far more modest in its claims. As described in the most recent release of the 2007 data: "TIMSS is designed to align broadly with mathematics and science curricula in the participating countries. The results, therefore, suggest the degree to which students have learned mathematics and science concepts and skills likely to have been taught in school."

PISA has not hesitated in making the most of these differences and the high visibility of the OECD has propelled PISA forward to become the most visible of the international assessments.

WHO PARTICIPATES IN THESE ASSESSMENTS AND DOES IT MATTER?

As an OECD product, the countries that participate in PISA differ from those that participate in TIMSS. In the 2006 PISA, a total of 57 countries (all 30 OECD members and 27

[6] *External Evaluation of the Policy Impact of PISA* EDU/PISA/GB(2008)35/REV1 13 November 2008, p. 10.

nonmember countries and economies) participated. All of the United States' major trading partners and competitors participate in PISA.

While as many countries and educational jurisdictions participated in the 2007 TIMSS as participated in PISA 2006, the list of participants differs. Only about half of the OECD countries participated in the 4th-grade TIMSS and even fewer participated in the 8th-grade assessment.

OECD does allow "partner countries" to participate in PISA, and indeed the number of partner countries is almost equal to the 30 OECD countries; however, OECD calculates a PISA test score average based only on the 30 member countries. In contrast, the TIMSS international average is based on all participating countries, which includes many less-developed countries such as Jordan, Romania, Morocco, and South Africa.[7]

We can compare TIMSS with the last PISA given in 2006 to see how widely different the comparative results can be. Remember that PISA is an assessment of 15-year-olds, so the closest comparison is with the 8th-grade TIMSS.

In TIMSS 2007, and focusing on math, 8th-grade students in the United States scored higher than the international TIMSS average of 500. However, we were 24 points *below* the OECD average math score in PISA 2006. Further, if we look at the highest-performing students in the United States compared to international averages, in TIMSS, the United States looks pretty good, with 3 times the percentage of 8th-grade students in the top 10% compared to the

[7] To provide comparisons between the 2007 results and prior results, the scores of students who participated in 2007 are scaled to be comparable with scores in prior administrations of TIMSS. The international average was established based on the 1995 TIMSS, which even then had a large number of low-performing less-developed countries participating. If more low-performing countries continue to join TIMSS, the average for that year's test will fall below 500 (indeed, in 2007, the average across all countries, unweighted by population, was 473 for 4th-grade math and 452 for 8th-grade math).

international median. However, in PISA, only 1.3% of U.S. students were in the highest proficiency level in 2006 PISA math—this was half the OECD average and in the same range as Greece, Mexico, Portugal, and Turkey.

These differences suggest why it is important to pay attention to the countries included in different international assessments.

BENCHMARKING TO GAUGE COMPARATIVE PERFORMANCE

We can use the information in TIMSS and PISA to gain more insight into the performance of American students. In Figure 1 below, I combine information on 8th-grade math performance from the 2007 TIMSS report, the 2007 NAEP, and the 2006 PISA math assessment of 15-year-old students.

TIMSS, PISA, and NAEP are all on different scales, so we need a way of expressing differences in a common metric—and effect sizes does just that. In Figure 1, we see that the difference between the United States and Chinese Taipei, the highest-performing jurisdiction in the 8th-grade TIMSS math test, is quite large—an effect size of 1.0. To further gauge the size of this difference, the next bar in the figure shows the standardized differences between Massachusetts, the highest-performing state in NAEP, and Mississippi, the nation's lowest-performing state.

- In terms of an analogy: *the United States is to Taipei as Mississippi is to Massachusetts*.

The distance between the United States and Korea, the highest-performing nation in PISA 2006, is also on the same order as the difference between Mississippi and Massachusetts. The gaps between black and white students in the United States and between students in high- and low-poverty schools are larger than the international gaps and the gap between the

performance of white and Hispanic students is the same size as the gap between the United States and Korea in PISA.

These comparisons give an immediate sense of the challenges facing the nation and show how international data can add context to our understanding of the condition of our education system. Moreover, as Erik Hanushek and his colleagues have demonstrated, the cognitive skills reflected in these tests are related to economic growth; doing well on PISA or TIMSS may be a leading indicator of how well a country's economy will perform in the future relative to others.

GENERATING POLICY ADVICE FROM INTERNATIONAL ASSESSMENTS

Figure 1 reflects the first aspect of benchmarking: the "how are we doing?" question. But what policy advice can be garnered from these studies? The quick answer is not as much as many would have you believe.

First, both PISA and TIMSS are cross sectional and do not allow longitudinal analysis at the student level, which is increasingly the way researchers prefer to measure growth in student achievement and to identify the factors associated with such changes. In addition, there are definitional problems in data across the many countries that lead to problems in comparability and interpretation. While both the International Association for the Evaluation of Educational Achievement (IEA) and the OECD work hard to make sure that measures are comparable, large gaps remain. One of the most interesting measurement issues flows from the importance given to attitudes in PISA. There are large gaps in how people in different countries understand similar questions. This issue of "measurement equivalence" and how to address it is a fundamental challenge for international assessments.[8]

[8] In the 2008 Brown Center report on American education, *How Well Are American Students Learning?*, Tom Loveless shows how poorly PISA is doing meeting this challenge.

Despite these limits, both TIMSS and PISA are used for policy recommendations. However, in the world of policy advice, IEA is a bit player compared to OECD.

OECD is a high-level intergovernmental organization with wide influence across many domains of finance, trade, environment, agriculture, technology, and taxation as well as education and its pronouncements have a gravitas that cannot be matched by the IEA. And just as OECD does not have a modest voice in describing the scope of PISA, its voice in pushing the policy lessons that can be "learned" from PISA has been equally amplified.

One problem is organizational: PISA combines the collection and release of statistical data with policy advice in a single unit. In contrast, in the United States, and indeed in most governments, these two functions are separated. In this country, strict Office of Management and Budget (OMB) guidelines separate the release of federal statistical reports by time and space from any policy statements (for example, during the last Administration, the Secretary of Education was never present at the release of any National Center for Education Statistics [NCES] reports, including the high-visibility NAEP report cards).

This pressure from policy makers for advice based on PISA interacts with this unhealthy mix of policy and technical people. The technical experts make sure that the appropriate caveats are noted, but the warnings are all too often ignored by the needs of the policy arm of PISA. As a result, PISA reports often list the known problems with the data, but then the policy advice flows as though those problems didn't exist. The reader is too easily caught up in the advice and forgets those boring, hard-to-read caveats. As a result, some have argued that PISA has become a vehicle for policy advocacy in which advice is built on flimsy data and flawed analysis.[9]

[9] See the "critical bibliography" assembled by Joachim Wuttke, http://www.messen-und-deuten.de/pisa/biblio.htm#43. Much of the critical work has come from Germany, where PISA has had a

POLICY ADVICE BUILT ON WEAK DATA

One of the most important issues in education research is how school resources affect student performance. Clearly, the stakes are high in identifying the best ways to allocate money, teachers, school leadership, and the like. The demands on PISA to offer policy advice on these issues is strong—countries and education ministers demand information on these issues and don't want to know how cross-sectional data using flawed measures can lead to bad advice or advice that seems "right" but that really has little or no basis in fact.

For example, consider Chapter 5 of the 2006 PISA science, which is devoted to analyzing the effects of school resources on student performance. The chapter includes all the appropriate caveats—Box 5.1 on page 215 notes that PISA is cross-sectional, that some things are not measured well, and that other important factors are unmeasured. Moreover, the school characteristics that are measured are from the student's current school—and in many countries a 15-year-old might have been in that school for only a year or two. The report notes that "the combination of these restrictions limits the ability of PISA to provide direct statistical estimates of the effects of school resources on educational outcomes" (p. 215).

Nonetheless, the chapter proceeds with dozens of charts and tables relating different school resources to student outcomes. Finally the chapter ends with "policy implications"—but the foundations for these implications are weak.

---

profound effect. Indeed, according to a recent OECD report, *External Evaluation of the Policy Impact of PISA* EDU/PISA/GB(2008)35/REV1 13 November 2008, the effects of the 2000 PISA in Germany were compared to the *Sputnik* shock and even the French Revolution. There has also been debate in Finland, in part because their students do so well in PISA but relatively poorly in TIMSS. See, for example, "The PISA survey tells only a partial truth of Finnish children's mathematical skills," http://solmu.math.helsinki.fi/2005/erik/PisaEng.html.

For example, the report states that "what is noticeable about the strongest effects measured in this chapter is that they are not the ones most closely associated with finite material resources, such as the distribution of good teachers. Rather, such effects are related to how schools and the school system are run—for example, the amount of time that students spend in class and the extent to which schools are accountable for results" (p. 277).

This is powerful stuff: hold schools accountable and other issues (such as the allocation of good teachers) don't matter. Of course, longitudinal research has consistently highlighted the importance of good teachers and PISA doesn't have a good measure of teacher quality—but that's beside the point.

What is the support for the powerful endorsement of accountability?

First of all, "accountability" turns out to be the public posting of school-level results—because no other measures of accountability were statistically significantly tied to PISA scores once student socioeconomic status (SES) was introduced (OECD 2006, 243). But the value-laden word "accountability" is used rather than "posting"—so the reader can easily have a mistaken impression that a much wider concept has been found to be important.

The report further notes that Australia, Canada, Finland, Japan, and Korea are the five OECD countries that show both above-average student performance in science and below-average impact of socioeconomic background on performance (a highly valued PISA outcome). Among these five countries, the percentage of students who attend schools that post achievement data ranges from 4% in Finland to 64% in Canada—the average across all five is 31% and the OECD average is 38%. Best practices might suggest *not* posting results, but the PISA report argues that accountability matters and that posting results is a way of improving performance.

To summarize PISA's research method supporting "accountability" as a means of improving performance: PISA begins with important caveats (cross-sectional data, bad measurement of important factors, and so on). It then produces complicated analyses that yield mixed results (at best). It focuses on one relationship in its suite of accountability measures—but uses the broader term anyway—and notes that high-performing countries vary in their practices regarding posting. Despite these problems PISA concludes that student performance is related to "the extent to which schools are accountable for their results."

MINING FOR NUGGETS AS A FOUNDATION FOR POLICY ADVICE

Much of the policy advice coming from these studies is based on identifying the "best practices" found in the highest-performing countries. For example the *Benchmarking for Success* report noted above specifically calls for drawing upon international best practices ("Action 4").

As is typical of the style of benchmarking studies, that report notes some best practices drawn from high-ranking PISA countries. A 2007 report by McKinsey and Company, *How the Best Performing Countries Come Out on Top*, is based entirely on identifying practices in high-performing PISA countries.[10]

All research methods have their limits, but all too often the presentation of "best practices" falls into a classic trap of "selecting on the dependent variable." The practices of high-performing countries are presented without any evidence about the extent to which these practices are also implemented in *low*-performing ones. Moreover, there is often just passing reference to the contextual effects of many reforms—one size often doesn't fit all. As a result, the presentation of best practices is often akin to a series of "just so" stories, without sufficient evidence to support claims of effectiveness.

[10] http://www.mckinsey.com/locations/ukireland/publications/pdf/Education_report.pdf

This mining for nuggets can lead to strange stories. The McKinsey benchmarking study, for example, discusses the importance of the high status of teachers in student achievement. As an illustration, the report praises South Korea's entry system of elementary school teachers (which is based on scoring highly on the national College Entrance Exam). In contrast to this rigorous recruitment of elementary school teachers, the report notes that Korea's secondary teacher training system is wide open, which in turn has created an oversupply of secondary school teachers—the result? The "…status and attractiveness of secondary school teaching has declined in South Korea, making it unattractive to high performers" (p. 19). Yet Korea is among the highest-performing countries in PISA's math assessment both overall and in the percentage of 15-year-olds in the highest achievement level. What is the "best practice" advice one can take away from this high-performing country regarding recruitment of teachers?

In short, standards of evidence and the research practices that are found in much PISA-based analysis would not pass muster in the equivalent U.S. statistical agencies and among most researchers in the United States.

GETTING STATE RESULTS FROM INTERNATIONAL TESTS

It is unlikely that governors will change their desire for state PISA scores in the face of the problems noted here. Setting aside the technical problems, there are practical considerations that policy makers should consider before moving forward.

How will international assessments fit into the already complex world of large-scale assessments? For example, it's conceivable that a state could do well in TIMSS and poorly in PISA (this happened in Finland) or a state could improve over time in NAEP and not in PISA.

If a state participates in PISA, can it align its curricula to both NAEP and PISA? Massachusetts is notable for its decision to align its curriculum and assessments with NAEP. As

a result, it is the highest-performing state in 8th-grade math and its proficiency standards are among the most rigorous in the nation. But if aligning state standards with PISA becomes a Massachusetts policy goal, can it align with NAEP through grade 8 and then spend the next 2 years aligning with PISA?

Can PISA really inform policy makers about how to improve the state's schools system? Remember, PISA assesses mathematical and science "literacy," a broader domain encompassing skills and knowledge learned both inside and outside of school. PISA 2006 cautions, "If a country's scale scores in reading, scientific or mathematical literacy are significantly higher than those in another country, it cannot automatically be inferred that the schools or particular parts of the education system in the first country are more effective than those in the second. However, one can legitimately conclude that the cumulative impact of learning experiences in the first country, starting in early childhood and up to the age of 15 and embracing experiences both in school and at home, have resulted in higher outcomes in the literacy domains that PISA measures." In short, can PISA really be used to identify which parts of the education pipeline are working well and which need improvement?

Getting schools and students to take no- or low-stakes tests is increasingly difficult. In the 2006 PISA assessment, the United States barely made the minimal school participation rate to be included in the analysis. If governors and chief state school officers are behind a state administration of PISA or TIMSS, getting school participation may be easier, but student engagement in low-stakes tests declines as they get older—something that is probably not fixable by gubernatorial exhortations.

These tests are not cheap. To get reliable estimates, about 1,500 students per state will need to be tested. The cost is around $500,000 for PISA and somewhat less for each grade of TIMSS. Nationwide, that's around $25 million if all states participated.

There may be cheaper alternatives. Gary Phillips has placed NAEP and TIMSS on the same scale and generated state scores that allowed comparison with nations that have taken TIMSS.[11] Phillips has also used the same technique to compare the performance of the 11 large urban districts that are in NAEP's Trial Urban District Assessment (TUDA) to the international rankings of TIMSS.[12] This statistical linking can be more easily done with TIMSS than with PISA, since TIMSS is administered at the same grade levels as NAEP and their purposes and frameworks are similar.

An alternative to statistical linking for PISA would be "small area estimation." These estimates are model based and "borrow" information from other data available for the state together with any state-level PISA data collected. The results are often known as "indirect" projections to distinguish them from standard or "direct" estimates. Further research would be needed to determine the feasibility of conducting small area estimates to generate state-level PISA scores. But this idea may be worth pursuing.[13]

THE DOUBLE-EDGED SWORD OF STATE-LEVEL PISA RESULTS

American students do not perform well in PISA compared most other OECD countries. In the aftermath of the latest dismal results from the 2006 PISA, the clamor for state-level PISA

[11] http://www.air.org/publications/documents/phillips.chance.favors.the.prepared.mind.pdf

[12] http://www.air.org/news/documents/Counting%20on%20the%20Future.pdf

[13] As an example, see http://nces.ed.gov/naal/estimates/overview.aspx in which the NCES produced state and country estimates of adults with low literacy based on the National Assessment of Adult Literacy (NAAL). Phillips is also working on a way of generating PISA scores by embedding PISA questions in state assessments.

scores was palpable. State interest in state PISA has receded given the current financial crisis and with the fading of the fanfare surrounding PISA 2006, but the economy will eventually turn around and PISA 2009 is in the offing. When those results are released in 2010 state interest will again grow.

While TIMSS has its partisans and several states have actually chosen to participate in it, momentum is behind PISA. If we do implement state PISA, what should states expect?

First, states would get a PISA score that would allow them to compare themselves to other PISA participants. In some cases, this would provide bragging rights ("Our students scored better than those in Korea"). In most states, disappointing results would provide reform-minded governors with ammunition to push for policy changes. But along with the PISA scale score would come all of the OECD's policy advice, which might make it harder for governors to choose the policy options they prefer. *Caveat emptor*.

Figure 1. International Benchmarking Can Provide Insights into U.S. Student Performance