On the merits of, and myths about, international assessments

Oren Pizmony-Levy, James Harvey, William H. Schmidt, Richard Noonan, Laura Engel, Michael J. Feuer, Henry Braun, Carla Santorno, Iris C. Rotberg, Paul Ash, Madhabi Chatterji and Judith Torney-Purta
(Author affiliations can be found at the end of the article)

Abstract

Purpose – This paper presents a moderated discussion on popular misconceptions, benefits and limitations of International Large-Scale Assessment (ILSA) programs, clarifying how ILSA results could be more appropriately interpreted and used in public policy contexts in the USA and elsewhere in the world.

Design/methodology/approach – To bring key issues, points-of-view and recommendations on the theme to light, the method used is a "moderated policy discussion". Nine commentaries were invited to represent voices of leading ILSA scholars/researchers and measurement experts, juxtaposed against views of prominent leaders of education systems in the USA that participate in ILSA programs. The discussion is excerpted from a recent blog published by *Education Week*. It is moderated with introductory remarks from the guest editor and concluding recommendations from an ILSA researcher who did not participate in the original blog. References and author biographies are presented at the end of the article.

Findings – Together, the commentaries address historical, methodological, socio-political and policy issues surrounding ILSA programs *vis-a-vis* the major goals of education and larger societal concerns. Authors offer recommendations for improving the international studies themselves and for making reports more transparent for educators and the public to facilitate greater understanding of their purposes, meanings and policy implications.

Originality/value – When assessment policies are implemented from the top down, as is often the case with ILSA program participation, educators and leaders in school systems tend to be left out of the conversation. This article is intended to foster a productive two-way dialogue among key ILSA actors that can serve as a stepping-stone to more concerted policy actions within and across national education systems.

Keywords Accountability, TIMSS, Assessment policy, International assessments, PISA, Educational quality, ILSA, National competitiveness

Paper type Technical paper

Introduction

Madhabi Chatterji, *Guest Editor, Teachers College, Columbia University*[2]

Front-page headlines and editorial sections of newspapers around the world today grab our attention frequently by announcing the latest results from various International Large-Scale Assessment (ILSA) programs. The media are quick to

Quality Assurance in Education
Vol. 22 No. 4, 2014
pp. 319-338
Emerald Group Publishing Limited
0968-4883
DOI 10.1108/QAE-07-2014-0035

highlight the standings of particular nations on ILSA scores, commenting on their implications for global competitiveness. Provocative and satirical headlines like the following, "Chinese Third-Graders Falling behind US High School Students in Math, Science" (The Onion, 2013; www.theonion.com/articles/report-chinese-thirdgraders-falling-behind-us-high,31464/), appeared after the publication of the 2012 Programme for International Student Assessment (PISA) reports in December 2013.

World-wide attention notwithstanding, ILSA reports tend to be miscast by the media, and are often over-interpreted and over-generalized in public and policy discussions (Backhoff, 2013; Feuer, 2013; Laurie, 2013; Plisko, 2013; Wagemaker, 2013 in Chatterji, 2013). There is ambivalence among educators, researchers and the public about ILSA programs and their findings, especially when recent results like those of the PISA in Shanghai lead to instant international economic comparisons. Some of the critiques are well-founded. Others are not. There are *myths* surrounding what ILSA reports can and cannot tell us that need to be discussed and even questioned. There are also *merits* to ILSA programs, some of which need to be underscored, while others are clarified or qualified.

ILSA programs are technically complex and multi-layered research endeavors. In the typical case, the public sees ILSA rankings of countries based on average student achievement, measured via sample-survey testing of students, focusing on subjects like math, reading and science. The assessments on which these rankings are based are not the simple tests that most members of the public may envision, where scores consist of percentages of items answered correctly. The questions included in the ILSA tests, their scoring mechanisms and final scale properties, the samples of students tested and ways in which reports present differences between countries are complex and difficult to explain in lay language. Furthermore, many participants – including researchers, public educators and policymakers – do not understand either the historical forces that brought ILSAs into existence, nor the political and educational factors that are influencing their rapid expansion and sustenance internationally (currently, participating nations are at 65 for PISA).

There are other questions, as well. How are ILSA programs impacting students, teachers and leaders in education systems, and what are their repercussions in other sectors of society? How might we transcend the media "hype" and steer ILSA programs into more productive international conversations on educational policy?

To answer such questions, this moderated policy discussion in QAE is titled: *On the Merits of, and Myths about, International Assessments*. From their respective contexts and expert vantage points in the USA, the participants bring to the table extensive knowledge and direct experience with the "ILSA phenomenon". Perspectives include those of ILSA researchers, policy scholars and measurement experts: Oren Pizmony-Levy, William Schmidt, Laura Engel and Michael J. Feuer, Iris C. Rotberg and Henry Braun, in order of presentation; and school district superintendents and education leaders: James Harvey, Richard Noonan, Carla Santorno and Paul Ash, also in order of appearance. The discussion is excerpted from a recent blog published in *Education Week*, co-facilitated by James Harvey of the National Superintendents Roundtable and myself (http://blogs.edweek.org/edweek/assessing_the_assessments). To conclude, Judith Torney-Purta provides an outside ILSA researcher's thoughts and recommendations.

Back to the future on international assessments

Oren Pizmony-Levy, *Teachers College, Columbia University*[1]

Fascinated by the immense growth and visibility of ILSA programs, I spent the past few years exploring the socio-historical roots of ILSAs through archival materials and interviews with key-informants (Pizmony-Levy, 2013). Here, I discuss two major changes that took place in the world of ILSAs over the past 50 years, which allow us to better understand the ILSA phenomenon and perhaps, the intended and unintended consequences of ILSA programs.

First change: a shift in ownership from researchers to governments. The International Association for the Evaluation of Educational Achievement (IEA), established in 1958, was the first organization to conduct ILSAs. The IEA emerged from a working group of scholars under the auspices of the United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Education in Hamburg, Germany. Key figures in the group included Professor Benjamin Bloom (University of Chicago), Professor Torsten Husén (Stockholm University) and Professor Arthur Foshay (Teachers College, Columbia University). For these scholars, comparing educational systems using large-scale quantitative data was driven by intellectual objectives, for example:

> If custom and law define what is educationally allowable within a nation, the educational systems beyond one's national boundaries suggest what is educationally possible (Foshay, 1962, p. 7).

The IEA General Assembly (GA) is the organization's governing body. For many years, a majority of the countries were represented at the GA by individuals affiliated with an academic or research institute, while the rest of the countries were represented by individuals affiliated with governmental agencies. Since the mid-1990s, that pattern has been reversed. In 1986, the proportion of representatives from governmental agencies was 43.3 per cent; this figure jumped to 59.6 per cent in 1998 and to 73.4 per cent in 2012. Representatives from academic and research institutes correspondingly declined from 56.7 to 40.4 per cent and then to 26.6 per cent. In an interview, a high-ranking official at the IEA commented:

> When the first math study started [1964], these were researchers who became interested in exploiting (or understanding) the variance between countries. And then, more countries joined, because they thought that would be a good thing. Now, it became much of a more governmental thing, not so much a research thing. Governments, misguidedly, I think, jumped in on the bandwagon to see who is better or worse than Americans (excerpted from Pizmony-Levy, 2013).

This transformation is also evident in other organizations that conduct ILSAs. Throughout its somewhat shorter history, the PISA has been run by the Organisation for Economic Co-operation and Development (OECD), which is an inter-governmental organization. Moreover, the Governing Board of PISA has been populated by individuals affiliated with governmental agencies.

Second change: from research to policy […] and educational quality and accountability. In the early decades, the leading rationale for operating ILSA programs was formulated in terms of basic educational research. Indeed, in reading the IEA mission statement from 1968, we can find a strong emphasis on scholarship:

> [The] IEA is an international, non-profit-making, scientific association […] whose principal aims are: (a) to undertake educational research on an international scale; (b) to promote

> research aimed at examining educational problems common to many countries, in order to provide facts which can help in the ultimate improvement of educational systems […] (IEA, 1968, p. 1).

At that time, the IEA had a diverse portfolio of studies on topics such as literature education, English and French as foreign languages, classroom environment, civic education and computers in education. The IEA publications of that time avoided the "horse race" discourse among countries and results were presented in alphabetical order of countries.

> The aim is to develop a systematic study of educational outcomes in the school systems of the cooperating countries. The question we wish to ask is not "Are the children of country X better educated than those of country Y?" To us this seemed a false question begging all the important issues we need to study (Minutes of The IEA Project, 17-22 October 1960, p. 9).

In the past two decades, however, this research-oriented rationale was replaced by a more policy-oriented rationale, one that is more linked to accountability, educational quality indicators and, sadly, to international competition. The current version of the IEA's mission statement emphasizes the provision of "international benchmarks to assist policymakers", and the collection of "high-quality data" for facilitating the implementation of accountability policies in education. The portfolio of topics covered by ILSAs is less diverse and tends to focus on the basics: mathematics, science and reading. Finally, publication of ILSAs reinforces the "horse race" discourse by presenting results in "league" tables and "report cards" of ranked countries that improved and declined in student achievement, without describing the technical limitations of such inferences for those who are non-specialists.

Where do we go from here? Some recommendations. The two inter-related changes in the world of ILSAs are important because they unmask the link between politics (in a broad sense) and practices of ILSA programs. We should aim to bring back research and researchers to the front lines of ILSA programs. Indeed, there are some preliminary signs that this is happening. For example, since 2004 the IEA has organized a biennial International Research Conference (IRC) that is intended to create a forum for scholars from different countries who are interested in further exploring ILSA program data. The 5th IEA IRC was held in Singapore in 2013 with over 140 participants (www.unescobkk.org/education/news/article/5th-iea-international-research-conference-irc-2013/).

Another example is the OECD's Thomas J. Alexander Fellowship for ILSA scholars (www.oecd.org/edu/thomasjalexanderfellowship.htm). This program is funded by the Open Society Foundation that supports innovative analysis of data from PISA and other OECD-sponsored ILSAs.

Catnip[3] for politicians: international assessments

James Harvey, *National Superintendents Roundtable, USA*[1]

As a boy in Ireland, I attended Mass every Sunday, during which priests with their backs to the congregation mumbled in incomprehensible Latin. None of us knew what it meant, but we were assured these mysteries were good for the soul.

In education today, ILSA measurement experts occupy the exalted status of Irish priests. With their figurative back to the schools, these prelates genuflect at the altar of Item Response Theory and mumble confidently among themselves in a language known

as "psychometrics". No one in the school congregation understands a word of it, but we are assured these mysteries are based on science.

Lack of transparency in ILSAs. It is a troubling situation. Assessment results are but the tip of an iceberg (Harvey, 2014). All the assumptions that produce the numbers the public sees are buried beneath the waterline. Discussion between the ILSA measurement experts who develop the assessments, ILSA researchers who analyze and prepare reports and front-line educators who must act on them is almost non-existent. It is a broken conversation.

Pizmony-Levy's review (this issue) of the history of the governance of ILSAs introduces a new element that makes the situation even more complex. Researchers once dominated the governance mechanisms of ILSAs. That is not true today. Nearly three-quarters of the members of the policy governance body of the IEA, for example, are from government or the bureaucracy. It is a safe bet that many of these people do not possess even a rudimentary understanding of the complexities of ILSA reports. It is "voodoo that we do so well", as the arcana of ILSA psychometrics was aptly described by Jakob Wandall of Denmark at a 2012 conference hosted by the Assessment and Evaluation Research Initiative at Teachers College, Columbia University (Wandall, 2013).

The ILSA horse race. Pizmony-Levy's history (this issue) reminds me of Torsten Husén's 1982 testimony before the National Commission on Excellence in Education, of which I was the deputy director (Husén, 1982). Husén, one of the towering figures in the early history of educational research on ILSAs in Europe and of IEA, made a plea to the commission: Do not interpret ILSA reports as a "horse race" between nations. ILSAs were not developed by bookies to handicap horses but by researchers to help clarify the goals each nation pursued for itself through its schools. Unfortunately, the politically appointed members of the excellence commission could not resist the "catnip" of ranking nations by mean results. In its startling 1983 report, *A Nation at Risk*, the commissioners thundered: "On 19 [international] academic tests, American students were never first or second and, in comparison with other industrialized nations, were last seven times" (National Commission of Excellence in Education, 1983).

What would Husén make of where we are today? It is not just that the horse-race mentality of mean scores on ILSAs dominates the conversation, domestically and internationally. It is also that bureaucrats in both individual nations and international associations have demonstrated remarkable skill and cunning in providing more of the horse-race rhetoric politicians crave, which serves as catnip. And, truth be told, in the quiet of their university and think-tank offices, many measurement experts and ILSA researchers acknowledge that, although troubled by some directions in assessment today, they feel pressure to follow the crowd and the money available to explore school systems with apparently unsatisfactory results.

Looking ahead. So, where does this leave us? Several things might be said. First, no responsible educator denies the need for national/local assessments, international assessments or of accountability policies in education. But, accountability for appropriate use of ILSA reports needs to run in both directions. It is clear that teachers and school administrators are today at the mercy of how public officials interpret ILSA results. This puts a special onus on ILSA leaders, researchers and measurement specialists to live up to the best practices of the profession. They should not be in the business of providing "catnip" for politicians.

Second, it is time for another look at the governance structures of ILSA assessment programs. The dominance of bureaucrats and politicians on ILSA governing bodies should give way to more balanced representation of measurement experts, researchers and practicing educators. Finally, we should openly acknowledge that what lies "below the waterline" at present is simply incomprehensible to policymakers and the lay audience. It is time to clarify communications among these sectors (including researchers, measurement experts, politicians and educators). After all, even in the churches of Ireland, priests long ago abandoned Latin and turned their faces to the congregation.

What can PISA and TIMSS tell us?

William H. Schmidt, *Michigan State University*[1]

Growing awareness of the crucial role of education in a country's economic competitiveness has made the results of international assessments a major public event. The Trends in International Math and Science Study (TIMSS) and the PISA have been used as fodder for political and policy debates; in particular stressing that the mediocre performance of students from the USA in mathematics on these tests is evidence that significant changes are required in the American educational system. Unfortunately, a lot of discussion of the TIMSS and PISA results involves hasty generalizations based on country-level averages and an obsessive focus on international rankings. Far too often we find simplistic imitation of a high-ranking country's educational policies, or sweeping assertions that USA's performance is determined by student poverty or the false belief that higher-performing countries only test their brightest students.

Why ILSA reports pose interpretative challenges. The TIMSS and PISA tests are often confused with one another, partly because they happen to have similar scales with averages around 500. Yet, there are important differences between the two.

The different rankings of the USA on the TIMSS and PISA are in part due to the *different countries* that take each test. There are fewer rich countries that take the TIMSS, which makes the USA look a bit better. But the tests themselves have *different content*. TIMSS assesses mathematical knowledge, but PISA assesses mathematical literacy (how math is applied). Although most countries' performance is similar on both tests, there are important exceptions where a country does quite well on the TIMSS but poorly on the PISA, or vice versa (Mullis *et al.*, 2012; OECD, 2013).

Finally, it is important to remember that the international average is the mean across *countries*, not the mean of all *students* in every country. The latter is a bit lower because of the below-average performance of some larger countries. However, we should be paying a lot less attention to country averages and rankings anyway. It is a mistake to define a country's students based on a single number. The reality is much more complicated.

In fact, most of the variation in student performance on both the TIMSS and PISA is *within* countries, not across them (OECD, 2003). Simply because Japan's students have a higher average test score does not mean that every Japanese student does better than every American student. In both nations, there is a wide variation in student outcomes. These variations exist *between schools* and also *within schools*. We are used to talking about "good schools" and "bad schools" as if every student attending them is performing at the same level. Yet the TIMSS and PISA

demonstrate that this just is not true. Research based on the TIMSS suggests there is also a wide variation in performance across classrooms (Schmidt *et al.*, 1999). Unfortunately, PISA does not sample by classroom, obscuring the contributions of between-classroom variation to inequality.

Lessons for the future. The key lesson to be drawn from international assessments is that the system of education in a given nation – the package of educational policies – has a major impact both on the average performance of students and the extent of inequality among students. Student poverty is an important contributor to both of these, but USA's performance *cannot* be attributed solely to the number or distribution of poor and disadvantaged students (Sousa and Amour, 2010). Some countries do a much better job at mitigating the effects of students' poverty, and poverty cannot explain why even affluent students trail their peers in other countries. Other nations have much greater equality in educational outcomes, and we should study their approaches and adapt some of them to our own circumstances.

There is a great deal to learn from international assessments like the PISA and TIMSS. The overall performance of a country can give us a clue as to nations on which we could focus our attention. However, we must take care in how we interpret the results of these tests if we are to avoid drawing misleading conclusions.

On international assessments (and any assessments): less is more

Richard Noonan, *Wallingford-Swarthmore School District, Pennsylvania*[1]

My school district, in a suburb of the city of Philadelphia in the USA, recently completed an ambitious strategic planning project in collaboration with the University of Pennsylvania's Penn Center for Educational Excellence (www.gse.upenn.edu/pcel). When surveyed, parents, residents, students, staff and representatives from higher education universally proclaimed as their lead recommendation that the district should *relax* the invasive grip that standardized testing programs currently hold on curriculum, teaching and learning in our schools. The periodic publication and clamor about international assessment results typically strengthens that hold of testing on our schools. Politicians of all stripes jump on the results to pillory American educators and double-down on more standardized testing.

Implications of ILSA differences. If we are going to talk intelligently about these international tests, let us start with the ways in which the different ILSA programs differ. PISA and TIMSS *do not provide* a single, universal standard of quality regarding student achievement by which schools around the world can be judged. PISA, which shifts subject area focus per administration, aims to assess students' ability to "apply knowledge", while TIMSS, which looks only at math and science performance in selected grades, focuses on more traditionally presented, curriculum-based knowledge and skills. The same country can do well on PISA, but not so well on TIMSS.

Finland, broadly held-up as an educational model to the world because of consistently high ILSA rankings, scores relatively less well on TIMSS than it does on PISA. In the area of math, PISA values more of a constructivist view of knowledge, while TIMSS assesses the traditionally sequenced concepts and skills in a discipline. TIMSS mirrors more closely the structure of testing in the National Assessment of Educational Progress (NAEP) model in the USA, where the performance at selected grades in Grades kindergarten-12 is examined (http://nces.ed.gov/nationsreportcard/studies), while PISA does not seem to have any American counterpart. The point is that the content, structure

and emphases of these two ILSA programs differ. So, their reports do not reveal the same things, and would not necessarily lead us to the same conclusions.

What can we learn from ILSA results. This fact is not sufficient reason for us to dismiss examining and learning from what international test results do reveal. We know that NAEP performance is widely variable across the states (http://nces.ed.gov/nationsreportcard/studies/statemapping/). TIMSS results enable us to benchmark the performance of states in relation to whole countries around the globe. NAEP typically points to Massachusetts as the state achievement leader in the USA; yet TIMSS results show a number of countries outperforming Massachusetts in math and science.

TIMSS results can offer fresh perspectives from which we can gauge our relative success in advancing selected achievement goals. Girls in selected countries around the world outperform the USA on TIMSS tests. We can examine what those countries are doing to broaden the range of strategies our own schools could use to achieve this important goal. My district, like others, has been taking successful steps to boost the participation level of girls in the most advanced math and science courses.

As Schmidt (this issue) points out, there is arguably one fair conclusion that both international assessments lead to, which is that many other countries have been more successful in establishing unifying national educational goals. We continue to make slow progress, via implementation of reforms related to the Common Core State Standards (www.corestandards.org), toward a goal that many other industrialized countries addressed decades ago. The state of Indiana's decision to pull-out (in order, as one politician put it, to establish curriculum "by Hoosiers, for Hoosiers") shows just how steep the challenge is, given the strong resistance to releasing curriculum from strict local control. Still, we need to persist if our students are to be well-prepared to perform at the higher education level in an increasingly globalized employment marketplace.

Advantages to less testing in schools. We could gain a great deal by pursuing a "less is more" testing approach in our schools. That is, reduce the scope and span of standardized testing programs while ensuring that any assessment we implement provides meaningful results. It is almost unfathomable to me that states would sign up to participate in regularly scheduled PISA or TIMSS testing programs on top of the expansive testing regimen we already have in place. Yet we have to create space, in what is today an overscheduled regimen of testing of all kinds, for international assessments.

Participating regularly in international education assessments can provide us with valuable benchmarks and insights. We can create that space by returning to the pre-No Child Left Behind (NCLB, 2001) era of state proficiency testing limited to one grade per school level, per year. Doing so would provide those of us at the local school district level with the breathing room needed for a more genuine perspective on schooling and student progress, with less frenetic attention to data gathering.

Five myths about international large scale assessments (ILSAs)

Laura Engel and Michael J. Feuer, *The George Washington University*[1]

With the recent release of the 2012 PISA results, we are once again reminded about the extent to which ILSAs have gripped the world of education. ILSAs, consist of a diverse set of assessments, ranging across math and science, reading, civic and citizenship education, teacher education, and others (for a complete list and history of ILSAs, see the

introduction and appendix of a special issue *in Research in Comparative and International Education* edited by Engel and Williams, 2013). ILSAs offer exciting insights into complex education systems and serve as invaluable tools to compare education systems internationally. Yet, with their high profile and considerable policy impact, ILSAs are also surrounded by a number of persistent myths. With an aim to shrink the distance between the widespread beliefs and the emergent evidence of ILSAs, we explore five prevailing myths.

Myth number 1: Average achievement scores provide an accurate and comprehensive record of overall quality and effectiveness of education systems. The convenience of a single score to represent a system's performance has consistently proved to be appealing to policymakers, the media, reformers and the public. But, researchers have as consistently warned against the inherent dangers in using a single average achievement score as the leading indicator of educational quality. A more accurate and useful picture comes instead from deeper explorations of statistically significant performance variations within and among participating systems (for a resource on ILSA data analyses, see the recent handbook on ILSA edited by Rutkowski *et al.*, 2013). It is also beneficial to draw on multiple data sources, including from national assessments and other mixed-methods educational research, rather than relying on performance on a single instrument.

Myth number 2: ILSA results prove that the American education system is declining. Scholars have disputed some of the more alarming accounts of a stagnating or declining US education system. Some argue that the USA has never actually been first in the world educationally, pointing to the consistency in USA's performance on international tests since the 1960s (Ravitch, 2013). Exaggerated claims about a lagging American system often draw on PISA results (Feuer, 2012). It is also significant to note that USA has ranked relatively better on the Trends in International Mathematics and Science Study (TIMSS), which assesses fourth and eighth grade in math and science, and Progress on International Reading Literacy Study (PIRLS), which assesses reading achievement of fourth graders, than on PISA. Not only is it important to look at different ILSA results, but continued discussion is also needed about what ILSAs do and do not measure (Chatterji, 2013; Kane, 2013). For example, Heckman and Kautz's (2013) recent report argues that achievement tests are unable to fully assess valuable skills such as curiosity, motivation and creativity (see also Perlman Robinson and Alexander's 2013 discussion of the importance of non-cognitive skills).

Myth number 3: ILSA results are predictive of long-term macroeconomic outcomes. Based on this assumption, there is a projected image of USA's economic decline resulting from educational stagnation. The now familiar alarmist rhetoric linking stagnating scores with a prediction of declining economic productivity is based on an assumed causal connection between PISA scores and long-term macroeconomic outcomes. Some researchers have called for greater caution in making such predictive and causal links, suggesting that "the discourse seems to run ahead of the evidence" (Feuer, 2013, p. 205).

Myth number 4: International benchmarking based on ILSA results is sufficient evidence to transfer best practices to education systems. One of the more poignant ironies is that while USA's education policymakers frequently call for borrowing "best practices" from top-performing ILSA countries, educational reforms that emphasize high-stakes testing as the principal tool of accountability represent the opposite of what

top-performing countries actually do (Engel *et al.*, 2011). International benchmarking, often based on average scores and league tables, is utilized as superficial "wake-up calls" to inspire system reforms. This practice can undercut the potential that the information in the ILSAs has to stimulate and facilitate deeper and more effective research and educational practices.

Myth number 5: Because of sampling or other methodological imperfections, ILSAs offer little or no value. Fervent critiques of ILSAs tend to overstate their limitations and obscure the more subtle inferences that can be derived from rigorous comparisons. Comparing the large and fragmented system in the USA with small and relatively homogeneous systems like Finland or Korea can be obviously fraught with complexity. But there is no question that with appropriate cautions, there is much that can be learned from well-designed and executed cross-national assessments of student achievement. This is especially true when secondary analysis of these large-scale data sets supplements the rankings of averages and if subjects that go beyond mathematics and science are considered (Torney-Purta and Amadeo, 2013b).

Beyond the myths. As is true for much of educational research and rhetoric, extreme positions limit the possibilities for evidence-informed progress. Sweeping claims that distort the evidence of achievement of students in a given nation relative to other systems promote either a kind of "sky is falling" rhetoric or become an invitation to defend an untenable status quo; and meanwhile, the inherent value of rigorous comparisons is diluted or lost. It would be a mistake, though, to dismiss comparative research on the grounds that it does not enable definitive conclusions. We believe the contrary is true: *With less defensively held positions and greater balance, cross-national comparisons of student achievement offer an important basis for educational research, policy and practice.*

Making sense of gloom-and-doom ILSA headlines in school districts

Carla Santorno, *Tacoma School District, Washington*[1]

"USA Teens Lag as China Soars on International Test" (Hechinger, 2010).

"Wake-up call: USA students trail global leaders"(Armario, 2010).

As a superintendent of a very diverse urban school district, I cannot control the media's predictable assault on public education based on the latest international assessment results; however, I do have a responsibility to make meaning of these evaluative assessments in relation to our work in Tacoma. Where and what is the true signal from these international assessments, extractable from a cluster of noisy data?

Contextualizing ILSA results. Let me begin first with some background about my district. Tacoma is a researcher's paradise. A mid-sized school district with 57 schools and free/reduced lunch rates from 13 to 95 per cent, it has some of the highest and lowest academically performing schools in the state. On the new *WaKIDS* screening tests, some of our elementary schools have 98 per cent of their students entering "kindergarten ready"; but for others, it is less than half. On-time graduation rates in our seven comprehensive high schools vary from 97 to 63 per cent. Bilingual rates range from 1 to 40 per cent. Some of our schools' students are 85 per cent ethnic minorities (defined as non-White), while others are at 20 per cent.

With our rich diversity, we are a perfect microcosm to analyze the impact of demographics on academic performance. As most analysts now realize, the No Child

Left Behind legislation (NCLB, 2001) has been problematic at many levels and its reauthorization a bipartisan disaster. However, the requirement to disaggregate data by subgroups has honed our skills in seeing how achievement is associated with ethnicity, poverty, English Language Learners (non-native speakers of English), students with disabilities and other subgroups. And that is the exact perspective one needs to understand the international assessment results.

Meaning of ILSA country ranks. What does it mean to be first? Who is first? Why are they first? The answer can easily be captured in educational blogger and national principal of the year, Mel Riddile's short tweet: "PISA: It's Poverty, Not Stupid" (Riddile, 2014).

Perhaps the most effective whistle blower on the misleading interpretations of these global rankings was the late (and very much great) Gerald Bracey. When an international study of high school science and mathematics results was about to be released, Dr. Bracey noticed that Greece was substantially above the USA in both physics and math in high school. Having lived in Greece for a while, he faxed (pre-tweet days) "Are you kidding me?" He noted that fourth and eighth grade students in Greece performed near the bottom. He remarked to *The Washington Post*, "Do you really think these Greek kids suddenly encountered Socratic teachers in their high schools and shot their advanced students beyond ours? In a pig's eye!" (Mathews, 2009).

ILSAs and the poverty issue. Bracey and many that offer statistical analyses of these international rankings conclude that looking at results only from a competitive perspective does not tell the whole story. And the biggest devil in the details is called "poverty". In the USA, we have more socio-economic disparity than any other industrialized country. The difference between the "have" and "have not" in the USA is a gap unparalleled in the Western world. We are among the most powerful and wealthy societies in human history and yet have somehow tolerated the discarding of a large portion of our children to lives of poverty. What happens to international test scores when we statistically account for the fact that Finland has a far lower percentage of students living in poverty (3 per cent) compared to the USA (20 per cent)? Looking at the data from a simple regression analysis that adjusted for family income would likely re-write the newspaper headline as follows: "Factoring for Poverty, USA Soars on International Tests" (Irizarry, 2013).

How does this issue translate in my district? One of my principals at a high-poverty high school recounted that he recently took 15 freshmen in the ninth grade to an exhibit at the Pacific Science Center in Seattle, 40 minutes (or 30 miles) away. For 13 of the 15 students, it was their first visit to the city of Seattle. That is what poverty looks like. That is the reality of many of our high school students, whose families want the best for them but lack resources. That is the equity and opportunity gap issue that should be a national wake-up call. And that is what we should be thinking about when reading the alarmist headlines generated by reports of large-scale international assessments.

International test scores, economic competitiveness and STEM fields

Iris C. Rotberg, *The George Washington University*[1]

The ranking of the USA on international tests of science and mathematics continues to fuel rhetoric about economic competitiveness and shortages of scientists and engineers, despite the fact that the USA consistently ranks first, or among the top countries, in competitiveness. Moreover, there is little evidence of shortages of scientists and engineers to fill traditional Science, Technology, Engineering and Math (STEM) jobs. It

is sometimes argued, however, that these apparent strengths are fragile and we should not assume that because the numbers look good now, they will continue to look good in the future. That is a fair argument – none of us can predict long-term economic and scientific strength with any degree of certainty. But we do know, regardless of the outcome, it will not be international test-score rankings that make the difference.

Irrelevance of international rankings as economic indicators. The irrelevance of international test-score rankings is illustrated in reports of the International Institute for Management Development (IMD), a global business school in Switzerland, and the World Economic Forum, which rank countries by international competitiveness (see IMD, www.imd.org/; also see World Economic Forum, www.weforum.org/). The rankings are based on a set of variables chosen to reflect current knowledge about what is most important in determining competitiveness. These variables include, for example, the soundness of the economy and financial sector; business sophistication; innovation; the quality and fairness of governmental and private institutions; market efficiency; basic, technological and scientific infrastructure; and the overall strength of the education system (primarily capacity and access at all levels of education). International test-score rank was only one of the 113 criteria used by the IMD to measure these variables. Performance on international test-score comparisons was not even mentioned among the 114 criteria used by the World Economic Forum – and for good reason, given the sampling and measurement flaws in the rankings and their negligible role in assessing the overall quality of education systems, much less the strength of economies. Whether or not the USA continues to rank high on competitiveness, international test scores will remain virtually irrelevant.

Irrelevance of ILSA rankings to STEM human capital. The ILSA test-score rankings also have little value in predicting whether a country will produce an "adequate" supply of scientists and engineers. The USA's rank on test-score comparisons is often interpreted as a proxy for a shortage of talent in STEM fields, despite strong evidence that the USA has a large supply of students capable of going into those fields. It is true that many talented students choose not to enter STEM fields and many others who receive degrees in these fields choose not to work in them.

A study conducted by Anthony P. Carnevale and colleagues at Georgetown University, for example, found that only a fourth of high school students who score in the top quartile in mathematics choose to enter a STEM major in college; only half the students who start with a STEM major graduate with that major; and fewer than half the students who graduate with a STEM major are actually working in STEM fields ten years later (Carnevale *et al.*, 2011). These students, instead, have entered other fields, including architecture, business, finance or medicine.

The point is that the attrition from traditional STEM fields does not reflect a lack of American talent or training in these fields, but rather such factors as interests, salary differentials, a weak job market or outsourcing of jobs because of lower wages outside the USA. Apple is unlikely to hire American workers to replace the hundreds of thousands of workers outside the USA who are manufacturing and assembling component parts for its products because of more correct answers on a math test.

The USA currently has an ample supply of workers to fill traditional STEM jobs. Carnevale *et al.* (2011), however, frame the question differently and see a potential for future shortages. They ask whether the country can produce a skilled labor force large enough to fill both the traditional STEM jobs as well as the large number of other jobs

that might draw on similar skills, such as finance and medicine, taking into account projected retirement rates, possible reductions in foreign-born workers and a future growth in STEM jobs at sub-baccalaureate as well as higher levels of education.

Toward better use of international test scores. Whether or not the predicted shortages occur, the international test-score comparisons have become a diversion that detracts attention from the factors that can make a difference in scientific innovation and competitiveness. Indeed, the increasing focus on test scores has led to scripted learning and narrowing of the curriculum – trends that are inconsistent with an approach that encourages problem-solving and innovation. That focus is also inconsistent with educational approaches designed to give students a broad set of skills that will contribute to their effectiveness in the workplace and is likely to be counterproductive in both attracting and retaining students in STEM fields.

The focus on test scores also detracts attention from the serious underrepresentation of low-income populations in STEM fields and points to the larger problem that underrepresentation illustrates – the growing gap in income and access. The gap will not be narrowed by rhetoric about international test-score rankings.

OECD: Poverty explains 46 per cent of the variation in PISA scores

Paul Ash, *Lexington School District, Massachusetts*[1]

Iris Rotberg (this issue) argues that international tests of science and mathematics, such as the PISA, are not related to international economic competitiveness. She argues:

> International test-score rank was only one of the 113 criteria used by the IMD to measure these variables. Performance on international test-score comparisons was not even mentioned among the 114 criteria used by the World Economic Forum.

Indeed, this should come as no surprise, because the PISA was never designed to be a predictor of economic competitiveness. Such tests may measure student knowledge and skills on an international basis, but they were not designed to measure the factors that contribute to economic success within a nation. In 2013, in an international study of 1,700 worldwide CEOs by IBM, the following four traits were deemed as critical for an employee's future success in the world of work and life: collaborative, creative, flexible and communicative. These four traits are not even measured on the PISA tests.

As a school superintendent for 16 years in Massachusetts, I have seen the value of state, national or international standardized tests when they provide practitioners with actionable data they can use to assess the overall effectiveness of their school or school district's curriculum and instruction. Tests such as the PISA are useful tools if used for their intended purpose – to measure content knowledge and skills that students need to master for work and academic study after high school.

ILSA misuses in political contexts. While it is true that the most recent test scores of the USA were average among 62 education systems, the data do not tell us *why* one nation scored higher or lower than another. Unfortunately, US Secretary of Education, Arne Duncan recently misused the student results and blamed the public schools entirely for the nation's average PISA results in 2012. Duncan stated:

> PISA is an important, comparative snapshot of US performance because the assessment is taken by 15 year-olds in high schools around the globe. The big picture of US performance on the 2012 PISA is straightforward and stark: It is a picture of educational stagnation […]

educational complacency and low expectations. (www.ed.gov/news/speeches/threat-educational-stagnation/).

Even the OECD authors of the PISA test reports acknowledge that PISA results can be attributed to a combination of variables, including but not limited to schooling, life experiences/home environment, poverty, access to early childhood programs, school attendance and health. In 2013, the OECD wrote in one of their reports that poverty explains up to 46 per cent of the variance in PISA mathematics score in OECD countries. At no time did OECD claim, as Arne Duncan stated, that schools' performance on the test can be blamed on low expectations and complacency.

While it may be the case that OECD has never claimed that PISA is a direct proxy for economic competitiveness, still, *it has come awfully close, providing politicians here and elsewhere with an irresistible opportunity to make that leap in logic*. In releasing the 2013 PISA results, for example, OECD Secretary-General Angel Gurria linked schools and education to high levels of youth unemployment, rising inequality and a pressing need to boost economic growth and national competitiveness. Not even a footnote pointed to the considerable contributions of Wall Street and the City of London to these very problems (www.oecd.org/unitedstates/a-usa-and-international perspective on 2012 PISA/).

Making appropriate interpretations of ILSAs. PISA and other international standardized tests are useful in some contexts and not in others. These tests are neither entirely predictive of every outcome we value in society, nor entirely useless. Let us not overstate the usefulness of international standardized tests, but let us not condemn them completely. If American education policymakers are really concerned about student success after high school and college, then I suggest our nation establish effective policies that will reduce childhood poverty and ensure all students a high-quality public education. I also recommend Grade kindergarten-12 educators listen to the 1,700 worldwide CEOs who need graduates who are collaborative, creative, flexible and communicative.

Merits of international assessments

Henry Braun, *Boston College*[1]

In discussions of the future, the term *globalization* is ubiquitous, typically referring to the breakdown of national barriers to the movement of goods, services and people. Paralleling the emergence of a one-world economy, ILSAs have also risen to prominence, with acronyms such as TIMSS, PIRLS, PISA and SAS (PIAAC) now broadly recognized. ILSAs are garnering heightened media attention and, in many countries, exerting increasing influence on education policy. Not surprisingly, this trend has occasioned considerable criticism.

Acknowledging criticisms of ILSAs. Some of the criticism is methodological, principally questioning the comparability of results, given variable sample quality, differential coverage of the implicit curriculum and the need to adapt the assessment instrument to dozens of different cultures and languages. Other criticisms are directed at the excessive attention paid to country rankings and the tendency to over-interpret the results in the search for productive policy strategies. One particular concern is that setting national education goals in terms of improving a nation's ranking on one or more

ILSAs could lead to a global homogenization of education that does a disservice to the nation's distinctive culture and educational needs.

These critiques have merit. Indeed, ILSA sponsors and the organizations that actually conduct the assessments have worked to address methodological deficiencies, though much remains to be done. On the policy side, criticism of the pernicious impact of (naïve) country comparisons is certainly in order. At the same time, we should not lose sight of the many positive contributions that ILSAs make to education policy. The goal should be to strengthen their capacity to do good while working assiduously to minimize negative consequences, unintended or not. Let us look at some of these contributions.

Acknowledging contributions of ILSAs. Before the advent of ILSAs in the 1960s, each country's educational system was hermetically sealed – there was no way to make meaningful comparisons among them in productive and meaningful ways at all. At the state level in the USA, this was the situation in the USA before NAEP's Trial State Assessment began (www.nces.ed.gov/nationasreportcard/studies). One problem was that the claims made by those in charge of the system, often exaggerated and self-serving, could not be easily refuted. Perhaps the most well-known example was the common assertion that the state's students were performing above the national average (Cannell, 1988).

Today, the burden of proof lies on those making claims that run counter to the evidence provided by ILSAs, particularly if the divergence is substantial. Although country "league tables" play an out-size role in the minds of many policymakers, more useful comparisons are possible. For example, comparisons across jurisdictions of the variances in test scores, of the gradients of test scores on socio- economic status or of gaps between immigrants and native born students can be informative and even a spur to action. Sophisticated statistical analyses are not needed to extract useful information from ILSAs, as is attested by the information-rich almanacs produced in conjunction with the release of the basic data (OECD, 2013).

Making justifiable ILSA comparisons. A striking example is provided by the results of the latest Survey of Adult Skills (SAS), conducted under the auspices of the OECD's Programme for the International Assessment of Adult Competencies. SAS is a household survey that assesses adults aged 16 to 65 in literacy, numeracy and (under current federal administration administration) in problem-solving in technology-rich environments. Thus, it is possible to compare both skill levels and the relationships of skills to background characteristics across age groups within a country, as well as within age groups across countries. Of course, the cautions above regarding over-interpretation of the findings apply here as well.

What do we find? With regard to the oldest cohort, adults aged 55-65, the USA is a leader in literacy among OECD countries. Further, in the USA, as in other OECD countries, the youngest cohort, aged 16-24, has stronger literacy skills than the oldest cohort. However, in comparison to their age peers across the OECD, young adults in the USA are, at best, in the middle of the pack. With regard to problem-solving, the oldest cohort leads the OECD, but the youngest cohort places last. The concern is not simply the precipitous drop in the rankings; rather, it is that the score gap between the USA and other country leaders is substantively meaningful and serves as one of many possible indicators of global competitiveness. In the absence of the SAS, it would be nearly impossible to draw such a policy relevant conclusion.

Bearing in mind the limitations. As cross-sectional studies, ILSAs are limited in offering evidence to support directly the kinds of causal conclusions desired by policymakers and score differences of a few points at the national level are not particularly meaningful. But used wisely, the rich data generated by ILSAs, in conjunction with other relevant evidence, provide unique insights that can challenge unmerited complacency and establish worthy benchmarks for educators and policymakers to aim for.

Conclusion and recommendations

Judith Torney-Purta, *University of Maryland*[2]

This moderated discussion has provided an excellent starting point for formulating future steps to enhance the quality and contributions of ILSAs. We should capitalize on its momentum. Several directions for action emerge from this wide-ranging discussion. Issues raised by school leaders deserve attention (see Ash; Harvey; Noonan; and Santorno; this issue), as do the points made by scholars and measurement experts on the myths and merits of ILSA programs (see Braun; Engel and Feuer; Pizmony-Levy; Rotberg; and Schmidt, this issue). To conclude, I offer a few recommendations for the future.

Four recommendations. First, this moderated discussion suggests a need for *greater transparency* about the research processes of ILSA programs, particularly on decisions currently viewed as invisible or "below the waterline". Educational leaders and the public could benefit from better understandings of how questions on ILSAs are constructed (and reviewed cross-nationally) and how students' responses are treated in analytic models to produce comparable scores. These stakeholders do not know enough about sample weighting and how differences between countries are tested for significance or presented in tables. More productive discussions could result from more transparency.

Second, a *broadened dialogue* including more stakeholders, especially school district leaders and journalists who write about ILSAs and education (in-print and on-line), seems warranted. This dialogue should be structured on a long-term basis. It should not simply be reactive to questions surrounding participation of schools and students in a particular ILSA program, or in the use, dissemination and interpretation of ILSA results. A *committee of ILSA program advisors* with representatives from all relevant constituencies could be established, modeled on the Board on International and Comparative Studies in Education, to guide future ILSA studies (BICSE; see details in Torney-Purta and Amadeo, 2013a, p. 110).

The BICSE was established at the National Academy of Sciences in the 1990s funded by the federal Department of Education and National Science Foundation in the USA. It provided oversight and guidance to the rapidly developing field of ILSAs (National Research Council, 2003). An interchange of information and opinion should also be encouraged more broadly.

Third, researchers (especially early career researchers) should be encouraged to *invest time and develop expertise* in doing ILSA research, particularly in designing secondary analysis of the vast wealth of data that have been collected by ILSAs, and in communicating findings to stakeholders in user-friendly terms (Torney-Purta and Amadeo, 2013b; Chatterji, 2013). Efforts to bring together databases such as the

Cross-Time Cross-System Project (XTXS) are a helpful step and should be better known. Mixed-method research in which analysis of ILSAs form a part could be productive.

Finally, we should focus on improving the processes of *international collaboration* within ILSA projects. This could enhance the quality of the assessments, the transparency of the analyses and reporting and potentials for secondary analysis and other forms of research. Suggestions can be found in the report from a Workshop on Building Infrastructure for International Collaborative Research in the Social and Behavioral Sciences issued by the National Research Council (in press). This Workshop convened 50 individuals ranging from university administrators, to leaders of international collaborations, to funders, to leaders of professional organizations.

Notes

1. Authors, listed in presentation order. Biographical information follows at the end of the article.
2. Moderators, listed in presentation order. Biographical information follows at the end of the article.
3. Catnip is a herb of the mint family that cats like, producing an euphoric, sometimes hallucinogenic effect on them; also called "catmint" and "catswort".

References

Armario, C. (2010), "Wake-up call": US students trail global leaders. *NBC News*, available at: www.nbcnews.com/id/40544897/ns/us_news-life/t/wake-up-call-us-students-trail-global-leaders/#.U5duHnajN8F

Backhoff, E. (2013), "Validity issues in international large scale assessment (ILSA) programs: thoughts for developing countries", in Chatterji, M. (Ed), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing Limited, Bingley, pp. 233-251.

Cannell, J.J. (1988), "Nationally normed elementary achievement testing in America's public schools: how all 50 states are above the national average", *Educational Measurement: Issues and Practice*, Vol. 7 Nos 5/9.

Carnevale, A.P., Smith, N. and Melton, M. (2011), *STEM: Science, Technology, Engineering, Mathematics*, Georgetown University Center on Education and the Workforce, Washington, DC.

Chatterji, M. (2013), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing Limited, Bingley.

Engel, L.C. and Williams, J.H. (2013), "The globalization of assessment: a forum on international tests of student performance", *Research in Comparative International Education*, available at: www.wwwords.co.uk/rcie/content/pdfs/8/issue8_3.asp

Engel, L.C., Williams, J.H. and Feuer, M.J. (2011), "The global context of practice and preaching: do high-scoring countries practice what US discourse preaches?", paper presented at the World Educational Research Association, Taiwan.

Feuer, M.J. (2012), *No Country Left Behind: Rhetoric and Reality of International Large-Scale Assessment*, Educational Testing Service, Princeton, NJ.

Feuer, M.J. (2013), "Validity issues in international large scale assessment (ILSA) programs: 'truth and consequences'", in Chatterji, M. (Ed), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing Limited, Bingley, pp. 197-217.

Foshay, A.W. (1962), *Educational Achievements of Thirteen-year Olds in Twelve Countries: Results of an International Research Project, 1959-61*, Vol. 4, International Studies in Education, UNESCO, New York, NY.

Harvey, J. (2014), "Response to 'validity, test use, and consequences: pre-empting a persistent problem'", *Education Week*, available at: http://blogs.edweek.org/edweek/assessing_the_assessments/2014/03/response_to_validity_test_use_and_consequencespre-empting_a_persistent_problem.html

Hechinger, J. (2010), "Stratifying PISA scores by poverty rates suggests imitating Finland is not necessarily the way to go for US schools", simple statistics, available at: http://simplystatistics.org/2013/08/23/stratifying-pisa-scores-by-poverty-rates-suggests-imitating-finland-is-not-necessarily-the-way-to-go-for-us-schools/

Heckman, J.J. and Kautz, T. (2013), "Fostering and measuring skills: Interventions that improve character and cognition", Research Paper No. 19656, National Bureau of Economic Research, Cambridge, MA.

Husén, T. (1982), *A Cross-National Perspective on Assessing the Quality of Learning*, US Department of Education, Washington, DC.

International Association for the Evaluation of Educational Achievement (IEA) (1968), Pamphlet, IEA, The Netherlands.

Irizarry, R. (2013), "US teens lag as China soars on international test", *Bloomberg News*, available at: www.bloomberg.com/news/2010-12-07/teens-in-u-s-rank-25th-on-math-test-trail-in-science-reading.html

Kane, M. (2013), "Validity and fairness in the testing of individuals", in Chatterji, M. (Ed), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing Limited, Bingley, pp. 17-55.

Laurie, R. (2013), "Applying Feuer's validation framework in a Canadian context: a look at international large scale assessment programs", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing Limited, Bingley, pp. 251-263.

Mathews, J. (2009), "Gerald Bracey, 69, dead, acidic critic of education policy", *The Washington Post*, available at: www.washingtonpost.com/wp-dyn/content/article/2009/10/22/AR2009102204549.html

Mullis, I.V.S., Matin, M.O., Foy, P. and Arora, A. (2012), *TIMSS 2011 International Results in Mathematics*, TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

National Commission of Excellence in Education (1983), *A Nation at Risk: The Imperative for Educational Reform*, US Government Printing Office, Washington, DC.

National Research Council, Board on International Comparative Research in Education (2003), *Understanding Others, Educating Ourselves*, The National Academies Press, Washington, DC.

No Child Left Behind (NCLB) Act of 2001 (2001), Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Organisation for Economic Co-operation and Development (2003), "Education at a glance: OECD indicators", *OECD Indicators*, Centre for Educational Research and Innovation, Paris.

Organisation for Economic Co-operation and Development (OECD) (2013), "Asian countries top OECD's latest PISA survey on state of global", OECD, available at: www.oecd.org/newsroom/asian-countries-top-oecd-s-latest-pisa-survey-on-state-of-global-education.htm

Pizmony-Levy, O. (2013), "Testing for all: the emergence and development of international assessments of student achievement, 1958-2012", *unpublished doctoral dissertation*, Indiana University, Bloomington, IN.

Plisko, V.W. (2013), "Validity and international large scale assessment programs: a reaction to Feuer's 'truth' and 'consequences'", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing Limited, Bingley, pp. 251-263.

Ravitch, D. (2013), "My view of the PISA scores", Diane Ravitch, available at: http://dianeravitch.net/2013/12/03/my-view-of-the-pisa-scores/

Riddile, M. (2014), "PISA: it's still 'poverty not stupid'", National Association of Secondary School Principals, available at: http://nasspblogs.org/principaldifference/2014/02/pisa-its-still-poverty-not-stupid/

Robinson, J.P. and Alexander, J. (2013), "Three lessons from the latest PISA scores", Brookings brief, available at: www.brookings.edu/blogs/education-plus-development/posts/2013/12/11-lessons-pisa-scores-perlman-robinson

Rutkowski, L., von Davier, M. and Rutkowski, D. (Eds) (2013), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL.

Schmidt, W.H., McKnight, C.C., Cogan, L.S., Jakwerth, P.M. and Houang, R.T. (1999), *Facing the Consequences: Using TIMSS for a Closer Look at US Mathematics and Science Education*, Kluwer Academic Publishers, USA.

Sousa, S. and Armor, D.J. (2010), "Impact of family vs. school factors on cross-national disparities in academic achievement: evidence from the 2006 PISA survey", Research Paper, No. 2010-25, School of Public Policy, George Mason University, Fairfax, VA.

The Onion (2013), "Report: Chinese third-graders falling behind US high school students in math, science", *The Onion Report*, available at: www.theonion.com/articles/report-chinese-thirdgraders-falling-behind-us-high,31464/

Torney-Purta, J. and Amadeo, J. (2013a), "The contributions of international large-scale studies in civic education and engagement", in von Davier, M., Gonzalez, E., Kirsch, I., Yamamoto, K. (Eds), *The Role of International Large-Scale Assessments*, Spring, New York, NY, pp. 87-114.

Torney-Purta, J. and Amadeo, J. (2013b), "International large-scale assessments: challenges in reporting and potentials for secondary analysis", *Research in Comparative and International Education*, Vol. 8 No. 3, pp. 248-258.

Wagemaker, H. (2013), "International large scale assessment (ILSA) programs and the challenges of consequential validity", in Chatterji, M. (Ed.), *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing Limited, Bingley, pp. 217-233.

Wandall, J. (2013), "Education, testing, and validity: a Nordic comparative perspective", in Chatterji, M. (Ed.) *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*, Emerald Group Publishing Limited, Bingley, pp. 137-163.

Author Affiliations

Oren Pizmony-Levy, Teachers College, Columbia University, New York, NY, USA

James Harvey, National Superintendents Roundtable, Seattle, WA, USA

William H. Schmidt, Michigan State University, East Lansing, MI, USA

Richard Noonan, Wallingford-Swarthmore School District, Wallingford, PA, USA

Laura Engel, The George Washington University, Washington, DC, USA

Michael J. Feuer, The George Washington University, Washington, DC, USA

Henry Braun, Boston College, Chestnut Hill, MA, USA

Carla Santorno, Tacoma Public Schools, Tacoma, WA, USA

Iris C. Rotberg, The George Washington University, Washington, DC, USA

Paul Ash, Lexington Public Schools, Lexington, MA, USA

Madhabi Chatterji, Teachers College, Columbia University, New York, NY, USA

Judith Torney-Purta, University of Maryland, College Park, MD, USA

About the authors

Oren Pizmony-Levy is an Assistant Professor in the Department of International and Transcultural Studies at Teachers College, Columbia University, with research interests in the intersection between education and social movements, such as accountability, environmentalism and human rights.

James Harvey is the Executive Director of the National Superintendents Roundtable, USA. Follow at www.superintendentsforum.org or on Twitter: @natlsuperntndnt.

William H. Schmidt is University Distinguished Professor and Co-Director of the Education Policy Center at Michigan State University.
Richard Noonan is the Superintendent of schools at the Wallingford-Swarthmore School District, Pennsylvania.

Laura Engel is an Assistant Professor of international education and international affairs at The George Washington University.

Michael J. Feuer is Dean of the Graduate School of Education and Human Development at The George Washington University.

Henry Braun is the Boisi Chair in Education and Public Policy in the Lynch School of Education at Boston College and Director of The Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP).

Carla Santorno is the Superintendent of schools at the Tacoma Public School District, Washington.

Iris C. Rotberg is a Research Professor of education policy affairs at The George Washington University.

Paul Ash is the Superintendent of schools at the Lexington School District, Massachusetts.

Madhabi Chatterji is Associate Professor of Measurement, Evaluation, and Education, and the Founding Director of the Assessment and Evaluation Research Initiative at Teachers College, Columbia University (AERI@TC). Madhabi Chatterji is the corresponding author and can be contacted at: mb1434@tc.columbia.edu

Judith Torney-Purta is Professor Emerita of Human Development and Quantitative Research Methods at the University of Maryland at College Park.