variables. *Journal of Research on Educational Effectiveness, 5*, 83–104.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.
Rothstein, J. (2007). Does competition among public schools benefit students and taxpayers? Comment. *American Economic Review, 97*, 2026–2037.
Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics, 113*(2), 553–602.
Sass, T. R. (2006). Charter schools and student achievement in Florida. *Education Finance and Policy, 1*(1), 91–122.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy.
Skoufias, E. (2005). *PROGRESA and its impacts on the welfare of rural households in Mexico*. Research Report No. 139. Washington, DC: International Food Policy Research Institute.
Somers, M. A., McEwan, P. J., & Willms, J. D. (2004). How effective are private schools in Latin America? *Comparative Education Review, 48*, 48–69.
Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J. R., McCaffrey, D. F., Pepper, M., & Stecher, B. M. (2010). Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1–21.
Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113*(485), F3–F33.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural Bolivia. *Review of Economics and Statistics, 88*(1), 171–177.
van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review, 43*(4), 1249–1286.
What Works Clearinghouse. (2013). *What Works Clearinghouse: Procedures and standards handbook*. (version 3.0). Washington, DC: Institute for Education Sciences. Downloaded Feb. 7, 2013 from: http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf.
Wolf, P. J., Kisida, B., Gutmann, B., Puma, M., Eissa, N., & Rizzo, L. (2013). School vouchers and student outcomes: Experimental evidence from Washington, DC. *Journal of Policy Analysis and Management, 32*, 246–270.
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT.
Zahorik, J., Molnar, A., Ehrle, K., & Halbach, A. (2000). Smaller classes, better teaching? Effective teaching in reduced-size classes. In S.W.M. Laine & J. G. Ward (Eds.), *Using what we know: A review of the research on implementing class-size reduction initiatives for state and local policymakers* (pp. 53–73). Oak Brook, IL: North Central Regional Educational Laboratory.

7

INTERNATIONAL LARGE-SCALE ASSESSMENTS

Uses and Implications

STEPHEN P. HEYNEMAN AND BOMMI LEE

INTRODUCTION

Measurements of academic achievement have a long history, but their frequency, heterogeneity, and, most importantly, scale have changed. Commencing as a pilot project in the 1960s, the international administration of tests of academic achievement has proliferated. The first tests were attempted in 12 countries with nonrepresentative samples, and the most recent tests were applied in 65 countries. International Large-Scale Assessments (ILSAs), in the future, will likely be applied in different jurisdictions—states (perhaps in Brazil, Russia, the United States, and India) as well as municipalities. The expansion of ILSA has been driven by demand for results from educational and political leaders, academics, journalists, and those who forge public policy, including economic policy. This growth has led to an increase in visibility and debates over interpretations and the implications of the results. This chapter provides an overview of the expansion of the ILSA, the impact of ILSA results on education policies in various countries, and a summary of the debates. It offers, albeit briefly, some guidelines as to their future utility.

BACKGROUND: THE EXPANSION OF ILSA

International large-scale assessments (ILSAs) have become more important as policymakers utilize their results in different ways. The relatively recent emphasis on educational accountability and the movement towards evidence-based research, as well as outcomes-based and standards-based education to inform public policy, account for the appeal of ILSA (von Davier et al., 2012; Steiner-Khamsi, 2003). International testing generally consists of large-scale assessments and surveys administered in different countries that provide comparative information both within and between countries (Wiseman, 2010). The main justification is that education in one country might be best understood by comparing it to education in other countries. Cross-national studies provide unique opportunities to understand issues in education and provide diagnostic and decision-making information about how to improve students' learning (Bradburn & Gilford, 1990; Ma, 1999; Stigler & Hiebert, 1999; Cai, 2004). Cross-national comparisons can also

provide a chance to examine educational factors outside the local context and to explore which are effective in which cultural contexts (Reynolds, 2000).

The origins of ILSA can be traced to the late 1950s, when Benjamin Bloom, C. Arnold Anderson, and colleagues at the University of Chicago agreed that the world might be seen as a "one big educational laboratory" (Husen, 1967; Heyneman & Lykins, 2008). It was noted that although the majority of research on education came from the United States, the country accounted for only 3 percent of the world's schools and school children. The idea was to see the differences in participating countries' educational systems as natural variations and to take advantage of this international variability to determine which characteristics might be more closely associated with academic achievement and under which circumstances (Husen, 1973). The first pilot survey of an international comparative assessment in reading, science, and mathematics was administered to nonrepresentative samples of 9,918 13-year-old students in 12 countries (Postlewaite, 1975). This informal group of academics founded a nongovernmental organization called the International Association for the Evaluation of Educational Achievement (IEA). The IEA raised enough money to sponsor a pilot study in 1964 called the First International Mathematics Study (FIMS). The style of organizing this first study continues a half century later. Curriculum and pedagogy specialists from different countries compare formal curricula, textbooks, and teacher training. Achievement test items are then developed out of these comparisons.

Since the FIMS study, the ILSAs have expanded in both reach and breadth (Rutkowski et al., 2013). FIMS was followed by the Second International Mathematics Study (SIMS) in 1980–1982, the Third International Mathematics and Science Study (TIMSS) in 1995 and its repeat in 1998. The IEA also administers the Progress in International Reading Literacy Study (PIRLS), which assesses reading skills of nine- and ten-year-olds and collects information on student and school environments. The field continues to expand and includes assessments and surveys covering topics such as computers in education, civics education, teacher education (Teacher Education and Development Study in Mathematics: TEDS-M; Teaching and Learning International Survey: TALIS), and adult literacy (Programme for the International Assessment of Adult Competencies: PIAAC) (Rutkowski et al., 2013).

Table 7.1 illustrates that the types of international testing were diversified over time, and the number of participating countries has increased, particularly after the 2000s. For instance, the number of countries participating in TIMSS increased from 45 in 1999 to 63 in 2011. One of the currently best-known assessment programs, the Program for International Student Assessment (PISA) was organized by the Organisation for Economic Co-operation and Development (OECD) and first administered to its member countries in 2000. Participation in this assessment of mathematics, science, and reading for 15-year-olds doubled over the next decade. While 32 countries participated in PISA 2000, 65 countries participated in 2012, the most recent assessment of PISA, because the OECD invited non-OECD member countries to participate (Kamens, 2013).

Distinction Between TIMSS and PISA

Because TIMSS and PISA are the most widely known international tests, we describe them in more detail. As Table 7.1 illustrates, TIMSS 1995, administered by the IEA, was the first assessment to bring together mathematics and science in a single study (Mullis & Martin, 2007). To reflect the fact that the demand for international test scores

Table 7.1 Selected International Large-Scale Assessments in Education

Sponsor	Description	Number of participating countries	Year(s) conducted
IEA	First International Mathematics Study (FIMS)	12	1964
IEA	Six subjects study		1970–1971
	Science	19 systems	
	Reading	15	
	Literature	10	
	French as a foreign language	8	
	English as a foreign language	10	
	Civic education	10	
	First International Science Study (FISS) (part of Six subjects study)	19	1970–1971
IEA	Second International Mathematics Study (SIMS)	10	1982
IEA	Second International Sciences Study (SISS)	10	1982
ETS	First International Assessment of Educational Progress (IAEP-I, Mathematics and Science)	6	1988
ETS	Second International Assessment of Educational Progress (IAEP-II, Mathematics and Science)	20	1991
IEA	Reading Literacy (RL)	32	1990–1991
IEA	Computers in Education	22	1988–1989
		12	1991–1992
Statistics Canada	International Adult Literacy Survey (IALS)	7	1994
IEA	Preprimary Project:		
	Phase I	11	1989–1991
	Phase II	15	1991–1993
	Phase III (longitudinal follow up of Phase II sample)	15	1994–1996
IEA	Language Education Study	25 interested countries	1997
IEA	Civic Education Study	28	1999
IEA	Third International Mathematics and Science Study (TIMSS)		
	Phase I	45	1994–1995
	Phase II-R (TIMSS-R)	40 (approximately)	1997–1998
IEA	Trends in International Mathematics and Science Study (TIMSS)		
		49	2003
		59	2007
		63	2011
OECD	Program for International Student Assessment (PISA)	32	2000
		41	2003
		57	2006
		65	2009
		65	2012
IEA	Teacher Education and Development Study in Mathematics (TEDS-M)	18	2007–2008
OECD	Teaching and Learning International Survey (TALIS)		
		24	2008
		33	2013
OECD	Program for the International Assessment of Adult Competences (PIAAC)		
	First round	24	2011–2012
	Second round	33	2014

Source: Chromy (2002) table modified by authors.

had become permanent, TIMSS was renamed Trends in International Mathematics and Science Study (TIMSS) in 2003, and it assesses mathematics and science achievement of fourth and eighth graders every four years.

One main feature that distinguishes TIMSS from PISA (which also tests mathematics and science) is that TIMSS is a curricula-based test. All IEA-sponsored studies, including TIMSS, collect information on the international variations in curricula, including variations in goals, intentions, and sequences of curricula (Robitaille et al., 1993). TIMSS also collects information on what the pupil learned (attained curriculum), what the teacher is expected to teach (intended curriculum), and what the teacher has in fact taught (implemented curriculum). TIMSS results in mathematics and science reflect how much students have learned within the given curriculum in each country. The relevance to the national curriculum may help determine the level of local education policy relevance, particularly important in those countries with few other sources of comparable student information (Elley, 2002).

Because of the way in which items are developed, TIMSS has in depth information on curriculum and teaching practices. But TIMSS also includes results from separate questionnaires administered to teachers and students and an assessment of physical facilities in sampled schools. The information from TIMSS provides a rich source to explore the influences on academic achievement. However such depth requires considerably more time and cost than ILSA tests without such background information.

PISA is a relative newcomer in the field of international testing, as the OECD became involved in international education testing only in the late 1990s (Wiseman, 2010). PISA was developed in 1997 at the request of the OECD member states. Their concern was to assess the degree to which students were prepared to enter a "knowledge society" (Lundgren, 2011). PISA tests are designed by test developers contracted by the OECD on the basis of what they consider should be the normal performance level of someone at a given age regardless of a nation's curriculum. Because items were not drawn from each nation's curriculum, a principal virtue of PISA was the fact that it could be administered quickly and comparatively cheaply. PISA was specifically adapted to the demands of political leaders who required a snapshot of student performance more than they required information on the influences on that performance.

Hence, PISA and TIMSS are different from one another not only in content but also in construction and purpose. PISA tests 15-year-olds' knowledge and skills (competencies) in reading, mathematics, and science that are needed in the labor market, as opposed to mastery of school curriculum. PISA is independent of national curricula whereas TIMSS is based on an assessment of each nation's curriculum. Barry McGaw (2002), the Director for Education of the OECD, described the difference this way: TIMSS is interested in "What science have you been taught and how much have you learned?" while PISA is interested in "What can you do with the science you have been taught?" (McGaw, 2002). PISA is also more policy oriented and provides recommendations in PISA reports, which makes it a strong tool for use in policymaking (Figazzolo, 2009).

IMPACT OF ILSA ON EDUCATION POLICIES

ILSAs now hold a unique place in the history of education. No previous innovation has captured the attention of a similarly broad range of interest groups, governments, unions,

academics, and representatives of political organizations. Reactions to TIMSS and earlier IEA assessments were significant, but often transitory, given that they may take a decade to develop. With PISA however, the reaction can be immediate, and because PISA can be repeated every two or three years, political reactions can be prolonged. Following the announcement of PISA results there is time to alter policies and test the degree to which alterations may make a difference over time. Though low for the test-takers, the stakes are high for the organizations and institutions whose reputation the results may affect.

A questionnaire to policymakers in 65 countries found that PISA performance was nearly universally considered as having the "potential to define" policy problems and set the agenda for policy debate at both national and local levels (Breakspear, 2012). PISA results are said to have created an "inter-dependence" across European systems of education and now constitute the *sine qua non* of education policymaking in Europe (Grek, 2012). In some instances the political stakes are so high that scores have been withheld and have been a factor in challenges to governments in power, as was the case in Hungary (Bajomi et al., 2009). Because of the political reaction created by consistently poor results, the Republic of South Africa stopped participating in ILSA (Wiseman, 2013). Japan has interpreted scores very carefully so as to fit into a local interpretation and legitimize policies. The international discussion of PISA results, for instance, caused the Ministry of Education to question its *yutori* (low-pressure) curriculum and reestablish the ministry's political centrality in a time of neoliberal state restructuring (Takayama, 2008).

The use of ILSA data to change education policy has increased partly due to the semantics of globalization (Schriewer, 2000), which has generated political and economic pressure to compare education systems across countries (Steiner-Khamsi, 2003). Whatever the reason, the tendency to use ILSA in national policymaking has exploded (DeBoer, 2010; Smith & Baker, 2001; Wiseman & Baker, 2005). With increased pressures for educational accountability and globalization, it is assumed that international testing will continue to expand (Kamens & McNeely, 2010). It is useful, therefore, to examine the impact of ILSA on national education policies. Table 7.2 summarizes the findings of recent studies that have examined this issue (e.g., Wagemaker, 2013; Heyneman & Lee, 2013). In this section, we focus on several useful themes and support each with illustrations from specific countries.

ILSA "Shock"

In some countries the results of the ILSA served as a wake-up call for the education sector. The shock value was strongest in countries that believed their education system had been of high quality. Traditionally the German education system had an underpinning philosophy that focused on individuals' desire to develop within themselves (Sorkin, 1983; Neumann et al., 2010). Local states had authority over education governance; teachers had authority over student assessment (Fensham, 2009). There was no standardized testing.

The belief that German education system was superior to others was shaken by the unexpected results from TIMSS and PISA. The mediocre German performance in TIMSS 1995 was followed by an even worse performance in PISA 2000 (Lehmann, 2011; Neumann et al., 2010). The shock in Germany had an impact similar to that in the United States with the launch of *Sputnik* and the publication of the *Nation at Risk* report (Gruber, 2006). The mediocre test performance of German students led to several significant

Table 7.2 Impact of ILSA in Selected Countries

	Initiated reforms due to ILSA results							Used for expediting existing reforms
Countries	**PISA**	**PIRLS**	**TIMSS**	**Changed curriculum reflecting ILSA test items**	**Established national education standards or standardized examination**	**Became aware of subgroups (immigrants, low-income)**	**Decentralized education finance and strengthened school autonomy**	
Australia	×							×
Chile			×	×	×			
Denmark	×					×		
England		×						×
France	×						×	
Germany	×	×			×	×	×	
Iran			×	×				
Ireland	×			×				
Italy	×							×
Japan	×			×				
Kyrgyzstan	×			×	×		×	
Latvia			×	×	×			
Macedonia			×	×				
New Zealand	×							×
Romania			×	×				
Russia			×	×	×			
Slovenia			×	×				
Switzerland	×					×		×

changes. Policymakers increased the funding of schools, took measures to improve instructional quality, and introduced National Education Standards (NES) (Neumann et al., 2010). The PISA framework influenced the NES and related curricular reforms. German states also adopted a concept of German schools as self-managing organizations and strengthened their autonomy (Grek, 2009). ILSA results also shifted the academic debate, which had been didactic and normative, to one that is more empirical and practice focused (Bohl, 2004).

At first, France ignored its mediocre performance in PISA (Dobbins & Martens, 2012), but after the deterioration in performance between 2003 and 2006, the government attempted to institutionalize a results- and evaluation-based approach emphasizing international comparisons (Mons & Pons, 2009; Pons, 2011). PISA seems to have provided a point of convergence for center-right and center-left points of view, including the priority of transferring Finnish pedagogical methods and school autonomy policies (Dobbins & Martens, 2012).

Kyrgyzstan experienced similar consequences as a result of participating in PISA 2006. Kyrgyzstan took the last place among the 57 participating countries, and the poor results came as a shock (Shamatov & Sainazarov, 2010). The PISA results indicated a shortage of school resources and highlighted issues of equity in access. The results were disputed and precipitated a national sample-based assessment (NSBA) that validated PISA results (Wagemaker, 2013). Despite the dispute over the results, the government of

Kyrgyzstan accelerated reforms in several areas, including development of new standards and curricula, a reduction of the number of subjects and education load per teacher, and the introduction of per-student financing (Shamatov & Sainazarov, 2010; Briller, 2009; Silova & Steiner-Khamsi, 2009).

ILSA and Low-Performing Subpopulations

In some instances ILSA results helped focus attention on specific low-scoring subpopulations. PISA and PIRLS results raised concerns about immigrant populations in Germany (Schwippert, 2007). PISA results revealed that students in Switzerland had shortcomings in reading competencies and significant gaps based on socioeconomic background (OECD, 2002). On the basis of PISA 2000, Denmark implemented a range of reforms focusing on students who were disadvantaged socioeconomically and immigrants (Egelund, 2008).

ILSA and the Support of Policies Already in Place

In some instances the results were used to justify policies already in place and over which there had been controversy. Switzerland, for example, was already making efforts to harmonize the different educational standards in each canton. After the PISA results, this trend moved to the top of the political agenda (Bieber & Martens, 2011). In Ireland, PISA 2006 results speeded up changes already planned for the curriculum in lower secondary education (Figazzolo, 2009). In England, PIRLS 2001 revealed that their National Literacy Project (a large-scale pilot study of a new way of teaching literacy at the primary level) proved to be quite successful (Twist, 2007). Similarly, students in New Zealand showed high performance in PISA, reinforcing existing policies (Dobbins, 2010).

ILSA Items and Curriculum

In some instances the items on ILSA tests, which often emphasize higher order cognitive skills of assessment and evaluation, have served to influence those emphases in new curriculum. Educators in the Russian Federation analyzed the TIMSS and PIRLS frameworks and developed recommendations for new educational standards for primary schools (Kovaleva, 2011). Macedonia and Iran used TIMSS test items to develop their national test items (Elley, 2005; Kiamanesh & Kheirieh, 2000). Slovenia and Romania also reflected the content of TIMSS curriculum framework and assessment items in their national curricula (Klemencic, 2010; Noveanu & Noveanu, 2000). PISA 2003 has led Japan to revise its national curriculum to incorporate competencies tested in PISA (Breakspear, 2012).

In case of Chile, TIMSS 1999 prompted curricular reform, new content standards in 2009, and reform in teacher education (Cariola et al., 2011). Singapore used TIMSS as one of several sources to inform annual reviews of their national curriculum (Poon, 2012). Iran used TIMSS to develop its first national item bank for primary education and a national research project to develop new test items (Kiamanesh & Kheirieh, 2000).

ILSA and National Assessments

In some instances ILSA tests have stimulated countries to design and implement their own national assessments. This included the Czech Republic (Strakova et al., 2000) and

Latvia, which established a centralized examination system for secondary education (Geske & Kangro, 2000). Russia did not have a tradition of using standardized assessment in schools, but began to use standardized tests after the TIMSS results were released (Kovalyova, 2000). Whether establishing national assessments will lead to better student performance is a separate question. Finland, for instance, one of the highest-achieving countries in PISA, does not implement a national assessment until the end of basic education (ninth grade) (Kupiainen et al., 2009).

Countries Less Affected by ILSA Results

Not every country is affected equally by the results of ILSA. If a country was already aware of its poor results, new below-average performance reiterates what had already been expected. For instance, the United States was not heavily affected by the ILSA results because Americans were already aware of their performance problems since the first *Sputnik* and *A Nation at Risk* (Wiseman, 2013; Bieber & Martens, 2011).

Impact of Regional Assessments

Perhaps the most significant reluctance to participate in ILSA has come from countries that do not expect to perform well. [1] Their resistance has been modified by the introduction of region-based ILSA in sub-Saharan Africa and Latin America. The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) began in 1995 with the agreement of 15 anglophone countries. Data on mathematics and reading literacy were gathered in 2000 and 2007. The Program d'analyse des systems educatifs des pays dela CONFEMEN (PASEC) was begun in 17 francophone countries with information gathered on French, mathematics, and occasionally a national language. UNESCO, UNICEF and the Inter-American Development Bank helped establish the Laboratorio latin americano de evaluation de la calidad de la educacion and sponsor tests to Monitor Learning Achievements (MLA) in the 1990s (Heyneman & Lee, 2013, p. 49).

Like PISA and TIMSS, regional assessments have had significant impact in some countries, whereas in other countries they were used to reinforce the *status quo* of current education policy. In Togo and Guinea teachers were being hired on contract with short periods of pedagogical training. PASEC results suggested that the teachers employed on a contractual basis were not doing any worse than teachers who are regularly employed, and current policies were continued (Bernard & Michaelowa, 2006).

As with global surveys, the impact of regional assessments tends to be larger when results are unexpected (Murimba, 2005). For instance, unexpected results led to policy change in Senegal where it was widely believed that a Senegalese primary student who repeats his/her grade would perform academically better than those who do not; however, the initial results of PASEC in 1998 showed no evidence to support that (Bernard & Michaelowa, 2006). After the continued negative result of grade repetition in 2003, the government prohibited grade repetition (Bernard & Michaelowa, 2006).

Perhaps the most important impact in Latin America has not been the ILSA tests directly but findings summarized, interpreted, and disseminated by an organization called the Partnership for Educational Revitalization in the Americas (PREAL). Since 1995, PREAL has published report cards that give grades to countries for their educational performance. Published in French, Spanish, Portuguese, and English, each report

card has stimulated changes in education policy and practice. These have included new national education standards in Peru and Honduras and a guide for reform in standards in Chile. Recommendations from PREAL's report cards have been directly implemented in El Salvador, Panama, and the Dominican Republic. PREAL's demand for greater accountability has influenced changes in Jamaica, Brazil, and Mexico City; and although countries in Latin America have traditionally been reluctant to participate in global ILSA, PREAL's recommendation to participate influenced countries to break with tradition and join PISA or TIMSS. The argument made by PREAL concerns the shift away from dictatorships toward democracy. PREAL views the results of ILSA as an illustration of the power voters have over the existence of information on which to judge their public school systems. Without these assessments, political leaders might claim excellence without reference to the facts; with assessments, claims of excellence can be compared to international standards of evidence (Heyneman & Lee, 2013, p. 50).

THE FUTURE OF ILSA

Technical Debates

Though carefully designed, ILSA tests are not without controversy. In spite of the efforts put into their construction, some argue that they continue to lack validity or reliability (Prais, 2003; Bracey, 2004, 2009; Dohn, 2007). Others argue that test items should be constructed and used differently. The purpose of ILSA tests is abstract and distant from the test-taker. Instead, they suggest the purpose should focus on having students achieve concrete accomplishments, which certify important qualifications. Test-takers differ dramatically in their desire to perform well. Korean students believe their performance will reflect the prestige of their nation; American students are not similarly motivated (Baker, 2007). If tests could be restructured to certify concrete accomplishments, students and schools might regain much needed balance in the differences on the part of the test-taker to do well on them. Interpretations of test results usually concentrate on the differences between schools in accounting for differences in academic achievement. Some suggest that the more valuable approach might be to concentrate on the differences within rather than between schools. For instance the OECD PISA report points out that some countries (Finland, Iceland, and Norway) have lower portions of variance explained by the differences between schools (considered an accomplishment); while in other countries (Hungary, Germany, Bulgaria, Austria, Belgium, Slovenia, and the Netherlands) the portion of variance explained by between-school differences is considerably greater. The distinction may matter. Policy recommendations generated by concerns over between-school variance might differ from those (such as to address immigrant differences) over within-school variance (Gaber et al., 2012).

The implications of research findings that use ILSA data are sometimes hampered because of problems in sampling, coverage, and administration (Wiseman, 2010). Sampling frames of ILSA tests are not the same across countries. Allowing for different sampling frames makes sense as each country has its unique context, but it also makes it difficult to compare the findings across countries. For example, sampling schools by location (rural, urban) might be important in countries like China where the achievement gap between rural and urban school students is large, but would not make much sense for countries like Singapore, where all schools are located in the urban area.

In addition to differing sampling frames, some countries may not fully adhere to the IEA or OECD's regulations related to sampling representativeness. If a country samples schools limited to elite students it can present problems of international validity and may generate objections from countries whose samples adhere to the standards recommended in a report of the National Academy of Sciences (Bradburn & Gilford, 1990). PISA was applied in many regions of China, but only scores from Shanghai were reported publicly. In addition there have been questions about whether the *hukou system*[2] in China has prevented a representative sample of 15-year-olds in the Shanghai student population from taking the test (Loveless, 2014). Thus, the high score of Shanghai in PISA 2009 is subject to criticism that it does not necessarily represent the student population in Shanghai.

Another limitation of ILSA is the absence of prior test scores. ILSA provides only cross-sectional data, making it difficult to draw causal inferences with these assessments. However, Carnoy (2006) suggests that PISA and TIMSS provide sufficient cross-sectional data to challenge important assumptions, and with their repeated applications, they can be used to approach causal inference on the effects of particular interventions.

The findings from ILSAs may also be hampered because of the cultural heterogeneity of participating countries. Can one generalize a major finding across countries with such diversity? Hanushek et al. (2013) address this issue through the use of country-fixed effects. Yet fixed effects may cause researchers to miss how country-specific characteristics affect student outcomes. This may be important for instance when comparing East Asian countries (where education is uniformly highly valued) with other parts of the world. The lesson is that drawing policy implications requires caution.

Academic Debates

Technical debates concern the nature of test design; academic debates concern the nature of cultural explanations, political effects, and institutional implications of the results. Some observers note the sudden dominance of the ILSA, organized by international organizations, and conclude that the nation-state has begun a withdrawal of sovereignty over education policy. Their concern is that

> the very meaning of public education is being recast from a project aimed at forming national citizens and nurturing social solidarity to a project driven by economic demands and labor market orientations.
>
> (Meyer & Benavot, 2013, p. 10)

These concerns are underscored by some in the academic community who bemoan the dominance of neoliberalism and its association with demand-based reforms that favor school choice, privatization, centralized goals, and competition as ways to improve efficiency and effectiveness (Ravitch, 2011; Steiner-Khamsi, 2003; Daun, 2005). They worry that school systems may be reduced to bureaucratic objects subject to the controls of external auditors (Apple, 2005; Power, 1999). And they seem concerned that the rise of ILSA represents a power shift away from the nation-state toward international non-educational organizations.

This opinion has been common to those interested in the World Bank, but is new with respect to other international organizations. Traditionally most international organizations have been viewed as harmless. Some suggest that they are now able to pressure

national governments to adopt new objectives and values (Finnemore, 1996; Barnett & Finnemore, 1999). Meyer and Benavot put it this way:

> Like all bureaucracies, they (international organizations) make rules, and, in so doing, they create social classifications that frame and reframe our understanding of social practice. International Organizations define tasks, create new social actors, and transfer new models of political organization around the world.
>
> (2013, p. 12)

These new concerns about the dominance of international organizations may overestimate their degree of autonomy. OECD, for instance, has little room to maneuver outside of the directives and financing of its member states. Its priorities, methods, and reports directly reflect their interests. Projects such as PISA are financed by voluntary contributions from those states that wish to participate. And though there are significant differences in the level of financing by which OECD member states are assessed, each has an equivalent vote over OECD's direction and products. Similarly, the World Bank does essentially what its member states demand. Although member states own quite different levels of equity shares, it is rarely the case that countries with more shares line up against those with fewer shares (Heyneman, 2003 and 2012).

Perhaps more compelling are the arguments that come from those who suggest that the position of dominance of ILSA represents a general shift toward what is seen as a rational audit explosion gaining ascendency in all public organizations—public hospitals, utilities, social welfare, and a move toward comparing similar organizations across national borders. To them, this audit culture appears to have emerged in a general search for best practices and magic bullets (Kamens, 2013). The underlying rationale however has been the connection between the performance of education and other public institutions and economic growth, prosperity, and what is described as the knowledge economy. It seems simplistic to suggest that this connection exists, but some might go further and portray test scores in particular subjects such as mathematics as being able to influence economic growth (Hanushek & Woessmann, 2009). If the economic fate of the nation rests on the ease of entry to the knowledge economy, and if entrance is influenced by the spread of geometry knowledge, the importance of PISA results rises to a new level.

In contrast to these conceptions of the economic importance of TIMSS and PISA results some explain performance on the basis of unshakable local cultural tradition. These observations do not deny the importance of ILSA but rather focus on the cultural background to testing. Korean, Japanese, and Chinese cultures for instance, have depended upon examination performance for centuries. Examinations have represented one of the few means of open access to social mobility, and examination results have taken a place of singular importance within the family for acquiring status assurance of the mother. Tests have also been used not only to gain access to universities but to allocate students to particular institutions and programs of study (Tucker, 2011). Similarly, while it has been common to note that Finland's success on PISA has come without the many demand-side reforms typical of North America—for example, teacher performance tests and frequent individual, school, state, and national assessments—what is less often noted is that Finland's economy had virtually collapsed in the 1990s, and national education performance was widely believed to be necessary for national survival (Sahlberg, 2011). Such extreme consensus is not simple to replicate elsewhere.

Differing test cultures have also led to questions about what is meant by schooling. In many Asian countries families are so preoccupied with examination performance that they consider the public school only one of several necessary mechanisms to augment performance. About 8 in 10 South Korean high school students utilize private tutors in addition to the time spent in public schooling. Universal private tutoring may double the amount of time devoted to test preparation. This suggests that Korean test performance may be high but also inefficient. When time out of school preparing for tests is combined with the time spent in school in both South Korea and the United States, it turns out American pupils are more efficient than pupils in South Korea. PISA performance of South Korean students come at about 30 percent greater time cost than the typical student in North America (Heyneman, 2013). This result underscores the fact that schools and education more generally comprise only a portion of the explanation for differences in ILSA results. Particularly important may be the result of differences in poverty and the existence of social safety nets associated with lowering the effects of poverty (Ladd, 2012); although the impact of poverty may differ from one country to the next (Heyneman & Loxley, 1983; Gamaron & Long, 2007), perhaps because of differences in income inequality (Chudgar & Luschei, 2009).

Official interpretations of PISA results have led to many new and constructive ideas well aside from the traditional Olympic race to the top. For instance, new interpretations have emphasized international comparisons of achievement gaps based on social background or performances changes over time sometimes associated with shifts in education policies (OECD, 2011; Schleicher, 2009).

Debates therefore have raised important issues but have not led to calls for the elimination of ILSA. The question then becomes what should we do with ILSA results? Is there an appropriate and constructive manner to use them?

CONCLUSIONS

ILSA performance will remain an important component of education in the foreseeable future. The question is how the education community should treat the results. Should it rail against the tests and their visibility in political debates (Meyer & Benevot, 2013, p. 14; Meyer, 2013)? Should the results be used to reinforce education policy positions held *ex ante*? Should a case be advanced for communities not to participate?

Systems of large-scale testing concentrate on particular content areas that are required and measureable. These can include knowledge that is expected to be used on a frequent basis, such as economic principles, scientific evidence, and the skills of synthesis and evaluation. They can also include knowledge of information that is not expected to be used on a frequent basis, such as the periodic table of elements (Feuer, 2012, p. 11).

But schools are expected to accomplish many goals other than in particular content areas. These may include the incorporation of characteristics such as diligence, empathy, social responsibility, and the normalcy of performing manual labor. These may include skills of leadership, cultural awareness, and the ability to care for animals. They may include proficiency in a second language or actions thought to foster particular outcomes such as community service and cross-cultural experiences. One problem of ILSAs is that they are not able to reflect the degree to which school systems accomplish these other goals well (Heyneman, 2005).

In addition to narrow coverage, ILSA results are associated with other weaknesses. On the basis of ILSA results, it is common to infer trends from snapshots in time, which may lead to generalizations that are premature. For example, in the 1980s many assumed that Japan's achievement scores were responsible for its superior economic performance prior to the stagnation in economic performance in the 1990s. ILSA may be biased in other ways. For instance, there may be a tendency for scores to be biased upwards in countries experiencing population declines (Feuer, 2012, p. 11).

Attributing changes in economic performance to scores in mathematics achievement (Hanushek & Woessmann, 2009) may be tenuous given that economic performance depends on a wide variety of influences. Current students would not have an impact on labor market productivity without a time lag of a decade or more, but the association with economic performance rarely accounts for this. Even in high-performing countries, there are persisting internal variations such as gender gaps in Finland and high-performing states within the United States. Moreover, it may be the case that the direction of the influence is the opposite from what is assumed. Economic performance, for example, may have an influence on school performance (Feuer, 2012, pp. 17–18).

Regardless of economic performance, it may be dangerous to rely on ILSA results to determine education policy. When applied properly, international comparisons are used to inform; when applied improperly they are used to mimic. For instance, unlike students in the United States, students in Finland are rarely subjected to quantitative evaluations in their schools, yet they perform well on PISA (Sahlberg, 2011). But Finnish children from both rich and poor families have similar values with respect to education. Teachers may not have to face the same problems of classroom discipline as do teachers in the United States. Raising the stakes for American pupils is one method of instilling a desire to perform well among students who have problems of understanding the importance of their education. Where that importance is already well understood and already present across socioeconomic strata, frequent performance tests may not be necessary. The key is to not use the presence or absence of a sector policy in a high-performing country to dictate a transfer of that policy elsewhere. For instance, when referring to the use and interpretation of National Assessment of Educational Progress (NAEP) scores Feuer says:

> The basic idea that the program can promote dialogue, rather than issue summative comparative judgments of quality of teaching or schooling in various locales, remains one of its distinguishing characteristics. By analogy, then, rather than view the results of ILSA programs as prima facie evidence of comparative success or failure, and by extension as the clinching argument for reforms that imitate characteristics of schools systems where students perform better, it might make more sense to explore how different types of assessments reflect values and expectations of schooling and to use the results as catalyst for public conversation and debate.
>
> (2012, p. 20)

One illustration of how tenuous it may be to extrapolate from one successful environment to another is the prevalence of shadow education among the Asian nations so successful in PISA. While it is true that PISA scores are high among pupils in South Korea, it is also the case that the typical Korean pupil spends the entire period of adolescence preparing for their tests, with little room for other development experiences or goals.

Scores may be higher in Korea, but Korean adolescents have little experience locating employment or participating in sporting events, activities that may be as common an expected experience among American adolescents as participating in shadow education is for adolescents in Asia. In predicting future economic performance, which is more relevant? In determining balance among well-adjusted adults, which is more important? These are the questions that ILSA raise but cannot answer.

NOTES

1. Until recently, state-level performance on the National Assessment of Educational Progress, the national assessment within the United States, was resisted by the southeastern states which anticipated their low performance by comparison to states in other regions.
2. The *hukou* system in China limits a family's permanent residence to their place of birth. Work in urban areas must be on the basis of a temporary migrant. Children of migrants are not allowed to attend regular public schools outside their place of birth. This significantly limits the representativeness of children in public schools in urban areas. The *hukou* system in China is parallel to the systems in North Korea, Vietnam, and the former Soviet Union.

REFERENCES

Apple, M. W. (2005). Education, markets and an audit culture. *Critical Quarterly, 47*(1–2), 11–29.

Bajomi, I., Berényi, E., Neumann, E., & Vida, J. (2009). The reception of PISA in Hungary. *Knowledge and Policy ORIENTATION, 3* (Supra-national Instruments Working Paper 12). Retrieved September 16, 2013, from http://www.knowandpol.eu/IMG/pdf/pisa.wp12.hungary.pdf.

Baker, E. L. (2007). 2007 presidential address: The ends of testing. *Educational Researcher, 36*(6), 309–317.

Barnett, M. N., & Finnemore, M. (1999). The politics of power and pathologies of international organizations. *International Organizations, 53*(4), 699–732.

Bernard, J. M., & Michaelowa, K. (2006). How can countries use cross-national research results to address "the big policy issues"? (Case studies from Francophone Africa). In K. N. Ross, & I. Jurgens-Genevois (Eds.), *Cross-national studies of the quality of education: planning their design and managing their impact* (pp. 229–240). Paris, France: International Institute for Educational Planning, UNESCO.

Bieber, T., & Martens, K. (2011). The OECD PISA study as a soft power in education? Lessons from Switzerland and the U.S. *European Journal of Education, 46*(1), 101–116.

Bohl, T. (2004). Empirische Unterrichtsforschung und Allgemeine Didaktik, Entstehung, Situation und Konsequenzen eines prekaren Spannungsverhaltnisses im Kontext der PISA-Studie. *Die Deutsche Schule, 96*(Ig/Heft 4), 414–425.

Bracey, G. W. (2004). International comparisons: Less than meets the eye? *Phi Delta Kappan, 85*(6), 477–78.

Bracey, G. W. (2009). PISA: not leaning hard on US economy. *Phi Delta Kappan, 90*(6), 450–51.

Bradburn, M. B., & Gilford, D. M. (1990). *A framework and principles for international comparative studies in education.* Washington, DC: National Academies.

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (OECD Education Working Papers 71). Paris, France: Organisation for Economic Co-operation and Development. Retrieved from http://dx.doi.org/10.1787/5k9fdfqffr28-en.

Briller, V. (2009). *Learning achievement in CEE/CIES region: An analysis of 2006 PISA results.* Presentation made at Seventh Central Asian Forum on Education Organized by UNICEF (September 15–17), Bishkek, Kyrgyzstan.

Cai, J. (2004). Why do US and Chinese students think differently in mathematical problem solving? Impact of early algebra learning and teachers' beliefs. *Journal of Mathematical Behavior, 23*(2), 135–167.

Cariola, L., Covacevich, C., Gubler, J., et al. (2011). Chilean participation in IEA studies. In C. Papanastasiou, T. Plomp, & E. Papanastasiou (Eds.), *IEA 1958–2008: 50 years of experience and memories* (pp. 373–388). Nicosia, Italy: Cultural Center of the Kykkos Monastery.

Carnoy, M. (2006). Rethinking the comparative—and the international. *Comparative Education Review, 50*(4), 551–570.

Chromy, R. R. (2002). Sampling issues in design, conduct and interpretation of international comparative studies of school achievement. In National Research Council (Ed.), *Methodological advances in cross-national surveys of educational achievement* (pp. 80–117). Washington, DC: National Academies.

Chudgar, A., & Luschei, T. (2009). National income, income inequality, and the importance of schools: A hierarchical cross-national comparison. *American Education Research Journal, 46*, 626–658.

Daun, H. (2005). Globalization and the governance of national education systems. In J. Zahda (Ed.), *International handbook on globalization, education and policy research* (pp. 93–107). Dordrecht, Netherlands: Springer.

DeBoer, J. (2010). Why the fireworks? Theoretical perspectives on the explosion in international assessments. *International Perspectives on Education and Society, 13*, 297–330.

Dobbins, M. (2010). Education policy in New Zealand—Successfully navigating the international market for education. In K. Martens, N. A. Kenneth, M. Windzio, & A. Weymann (Eds.), *Transformation of Education Policy*. Basingstoke: Palgrave.

Dobbins, M., & Martens, K. (2012) Towards an education approach a la finlandaise? French education policy after PISA. *Journal of Education Policy, 27*(1), 23–43.

Dohn, N. B. (2007). Knowledge and skills for PISA: Assessing the assessment. *Journal of Institutional and Theoretical Economics, 154*(4), 696–705.

Egelund, N. (2008). The value of international comparative studies of achievement—A Danish perspective. *Assessment in Education: Principles, Policy & Practice, 15*(3), 245–251.

Elley, W. (2002). *Evaluating the impact of TIMSS-R in low-and middle-income countries: An independent report on the value of World Bank support for an international survey of achievement in mathematics and science.* Amsterdam: International Association for the Evaluation of Educational Achievement.

Elley, W. B. (2005). How TIMSS-R contributed to education in eighteen developing countries. *Prospects, 35*(2), 199–212.

Fensham, P. J. (2009). The link between policy and practice in science education: The role of research. *Science Education, 93*(5), 1076–1095.

Feuer, M. J. (2012). *No country left behind: Rhetoric and reality of international large scale assessment.* 13th William H. Angoff Memorial Lecture. Princeton, NJ: Education Testing Service.

Figazzolo, L. (2009). *Impact of PISA 2006 on the education policy debate.* Brussels, Belgium: Education International. Retrieved from http://download.ei-ie.org/docs/IRISDocuments/Research%20Website%20Documents/2009–00036–01-E.pdf.

Finnemore, M. (1996). *National interest in international society.* Ithaca, NY: Cornell University.

Gaber, S., Cankar, G., Umek, L. M., & Tasner, V. (2012). The danger of inadequate conceptualization in PISA for education policy. *Compare, 42*(4), 647.

Gameron, A., & Long, D. (2007). Equality of educational opportunity: A 40-year retrospective. In R. Teese, S. Lamb, & M. Duru-Bellat (Eds.), *International studies in educational inequality, theory, and policy* (pp. 23–48). Dordrecht, Netherlands: Springer.

Geske, A., & Kangro, A. (2000). Latvia. In D. F. Robitaille, A. E. Beaton, & T. Plomp (Eds.), *The impact of TIMSS on the teaching & learning of mathematics & science* (pp. 85–88). Vancouver, Canada: Pacific Educational.

Grek, S. (2009). Governing by numbers: The PISA "effect" in Europe. *Journal of Education Policy, 24*(1), 23–37.

Grek, S. (2012). What PISA knows and can do: Studying the role of national actors in the making of PISA. *European Educational Research Journal, 11*(2), 243–254.

Gruber, K. H. (2006). The German "PISA-shock": Some aspects of the extraordinary impact of the OECD's PISA study on the German education system. In E. Ertl (Ed.), *Cross-national attraction in education: Accounts from England and Germany.* Providence, RI: Symposium.

Hanushek, E. A., Link, S., & Woessmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics, 104*, 212–232.

Hanushek, E. A., & Woessmann, L. (2009). *Do better schools lead to more growth? Cognitive skills, economic outcomes and causation* (NBER Working Paper No. 14633). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14633.

Heyneman, S. P. (2003). The history and problems in the making of education policy at the World Bank: 1960–2000. *International Journal of Education Development, 23*, 315–337.

Heyneman, S. P. (2005). Student background and school achievement: What is the right question? *American Journal of Education, 12*(1), 1–9.

Heyneman, S. P. (2012). Education policy and the World Bank: When models become monopolies. In A. Wiseman & C. Collins (Eds.), *Education strategy in the developing world: Understanding the World Bank's education policy revision* (pp. 43–62). Bingley, UK: Emerald Group.

Heyneman, S. P. (2013), The international efficiency of American education: The bad and the not-so-bad news. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power and policy: The emergence of global educational governance* (pp. 279–302). Providence, RI: Symposium.

Heyneman, S., & Lee, B. (2013). The impact of international studies of academic achievement on policy and research. In D. Rutkowski, L. Rutkowski, & M. von Davier (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 37–72). Boca Raton, FL: CRC.

Heyneman, S. P., & Loxley, W. (1983). The effect of primary school quality on academic achievement

across twenty-nine high- and low-income countries. *American Journal of Sociology, 88*(6), 1162–94.

Heyneman, S. P., & Lykins, C. (2008). The evolution of comparative and international education statistics. In H. F. Ladd, & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 105–127). New York: Routledge.

Husen, T. (Ed.). (1967). *International study of achievement in mathematics: a comparison of twelve countries,* Vols. 1–2. Stockholm, Sweden: Almqvist and Wiksell.

Husen, T. (1973). Foreword. In L. C. Comber & J. P. Keeves, *Science achievement in nineteen countries* (pp. 13–24). New York: Wiley.

Kamens, D. (2013). Globalization and the emergence of an audit culture: PISA and the search for "best practice" and magic bullets. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power and policy: The emergence of global educational governance.* Providence, RI: Symposium.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review, 54*(1), 5–25.

Kiamanesh, A. R., & Kheirieh, M. (2000). Iran. In D. F. Robitaille, A. E. Beaton, & T. Plomp (Eds.), *The impact of TIMSS on the teaching & learning of mathematics & science* (pp. 61–65). Vancouver, Canada: Pacific Educational.

Klemencic, E. (2010). The impact of international achievement studies on national education policymaking: The case of Slovenia. How many watches do we need? In C. Wiseman (Ed.), *The impact of international achievement studies on national education policymaking* (pp. 239–268). Bingley, UK: Emerald Group.

Kovaleva, G. (2011). Use and impacts of TIMSS and PIRLS data in the Russian Federation, *IEA Newsletter,* September.

Kovalyova, G. (2000). Russia. In D. F. Robitaille, A. E. Beaton, & T. Plomp (Eds.), *The impact of TIMSS on the teaching & learning of mathematics & science* (pp. 120–124). Vancouver, Canada: Pacific Educational.

Kupiainen, S., Hautamäki, J., & Karjalainen, T. (2009). *The Finnish education system and PISA.* Helsinki, Finland: Ministry of Education.

Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management, 31*(2), 1–25.

Lehmann, R. (2011). The significance of IEA studies in education in East Asia. In C. Papanastasiou, T. Plomp, & E. Papanstasiou (Eds.), *IEA 1958–2008: 50 years of experiences and memories* (pp. 389–410). Nicosia, Italy: Cultural Center of the Kykkos Monastery.

Loveless, T. (2014, January 8). *PISA's China problem continues: A response to Schleicher, Zhang, and Tucker.* Blog posted to http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2014/01/08-shanghai-pisa-loveless.

Lundgren, U. P. (2011). PISA as a political instrument. In M. A. Pereyra, H.-G. Kotthoff, & R. Cowen (Eds.), *PISA under examination: Changing knowledge, changing tests, and changing schools* (pp. 17–30). Rotterdam, Netherlands: Sense.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.* Mahwah, NJ: Erlbaum.

McGaw, B. (2002). Paragraph 1, Examination of Witnesses, Select Committee on education and Skills, 20 March 2002. Retrieved from http://www.publications.parliament.uk/pa/cm200102/cmselect.

Meyer, H.-D. (2013, December 19). OECD's PISA: A tale of flaws and hubris. *Teachers College Record.* Available at http://www.tcrecord.org/Content.asp?ContentId=17371.

Meyer, H.-D., & Benavot, A. (Eds.). (2013). *PISA, power and policy: The emergence of global educational governance.* Providence, RI: Symposium.

Mons, N., & Pons, X. (2009). La reception de PISA en France: Connaissances et regulation du systeme educative [The reception of PISA in France: Knowledge and the regulation of the education system] (Knowledge and Policy in Education and Health Sectors Working Paper 12).

Mullis, I. V., & Martin, M. O. (2007). Lessons learned: What international assessments tell us about math achievement. In T. Lovelace (Ed.), *TIMSS in perspective: Lessons learned from IEA's four decades of international mathematics assessments* (pp. 9–36). Washington, DC: Brookings Institution Press.

Murimba, S. (2005). The impact of southern and eastern Africa consortium for monitoring educational quality (SACMEQ). *Prospects, 35*(1), 91–208.

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education, 8,* 545–563.

Noveanu, G., & Noveanu, D. (2000). Romania. In D. F. Robitaille, A. E. Beaton, & T. Plomp (Eds.), *The impact of TIMSS on the teaching & learning of mathematics & science* (pp. 117–119). Vancouver, Canada: Pacific Educational.

OECD. (2002). *Education at a glance 2002. Paris,* France: Organisation for Economic Co-operation and Development.

OECD. (2011), *Lessons from PISA for the United States.* Paris, France: Organisation for Economic Co-operation and Development. Retrieved from *http://dx.doi.org/10.1787/9789264096660-en.*

Pons, X. (2011). L'evaluation des politiques éducatives [The evaluation of education policies]. Paris: PUF.

Poon, C. L. (2012). Singapore. *IEA Newsletter,* N39 April 2012.

Postlewaite, N. (1975). The surveys of the International Association for the Evaluation of Educational Achievement (IES). In A. C. Purves & D. U. Levine (Eds.), *Educational policy and international assessment: Implications of the IEA surveys of achievement* (pp. 1–33). Berkeley, CA: McCutchan.

Power, M. (1999). *The audit society: Rituals of verification.* Oxford, UK: Oxford University.

Prais, S. J. (2003). Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education, 29*(2), 139–163.

Ravitch, D. (2011). *The death and life of the great American school system: How testing and choice are undermining education.* New York: Basic.

Reynolds, D. (2000). School effectiveness: The international dimension. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 232–256). New York: Routledge.

Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science* (TIMSS Monograph 1). Vancouver, Canada: Pacific Educational.

Rutkowski, D., Rutkowski, L., & von Davier, M. (2013). A brief introduction to modern international large-scale assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 3–10). Boca Raton, FL: CRC.

Sahlberg, P. (2011). *Finnish lessons: What the world can learn from educational change in Finland.* New York: Teachers College.

Schleicher, A. (2009). Securing quality and equity in education: Lessons from PISA. *Prospects, 39,* 251–263.

Schriewer, J. (2000). World system and interrelationship networks: The internationalization of education and the role of comparative inquiry. In T. S. Popkewitz (Ed.), *Educational knowledge* (pp. 305–343). Albany, NY: SUNY.

Schwippert, K. (2007). The impact of PIRLS in Germany. In K. Schwippert (Ed.), *Progress in reading literacy: The impact of PIRLS 2001 in 13 countries* (pp. 93–108). New York, NY: Munster Waxmann.

Shamatov, D., & Sainazarov, K. (2010). The impact of standardized testing on education quality in Kyrgyzstan: The case of the Program for International Student Assessment (PISA) 2006. *International Perspectives on Education and Society, 13,* 145–179.

Silova, I., & Steiner-Khamsi, G. (Eds.). (2009). *How NGOs react: Globalization and education reform in the Caucasus, Central Asia and Mongolia.* Bloomfield, CT: Kumarian.

Smith, T. M., & Baker, D. P. (2001). Worldwide growth and institutionalization of statistical indicators for education policy-making. *Peabody Journal of Education, 76*(3/4), 141–152.

Sorkin, D. (1983). Wilhelm von Humboldt: The theory and practice of self-formation (Bildung). *Journal of the History of Ideas, 44,* 55–73.

Steiner-Khamsi, G. (2003). The politics of league tables. *Journal of Social Sciences, 1,* 1–6.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* New York: Free Press.

Strakova, J., Paleckova, J., & Tomasek, V. (2000). Czech Republic. In D. F. Robitaille, A. E. Beaton, & T. Plomp (Eds.), *The impact of TIMSS on the teaching & learning of mathematics & science* (pp. 41–44). Vancouver, Canada: Pacific Educational.

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*(4), 387–407. doi:10.1080/03050060802481413.

Tucker, M. (Ed.). (2011). *Surpassing Shanghai: An agenda for American education built on the world's leading systems.* Cambridge, MA: Harvard University.

Twist, L. (2007). The impact of PIRLS in England. In K. Schwippert (Ed.), *Progress in reading literacy: The impact of PIRLS 2001 in 13 countries.* New York: Munster Waxmann.

Von Davier, M., Gonzalez, E., Kirsch, I., & Yamamoto, K. (Eds.). (2012). *The role of international large-scale assessments: Perspectives from technology, economy, and educational research.* New York: Springer.

Wagemaker, H. (2013). International large-scale assessments: From research to policy. In D. Rutkowski, L. Rutkowski, & M. von Davier (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 11–36). Boca Raton, FL: CRC.

Wiseman, A. W. (2010). Introduction: The advantages and disadvantages of national education policy making informed by international achievement studies. In A. W. Wiseman (Ed.), *International perspectives on education and society* (pp. xi-xxii). Bingley, UK: Emerald Group.

Wiseman, A. W. (2013). Policy responses to PISA in comparative perspective. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power and policy: The emergence of global educational governance.* Providence, RI: Symposium.

Wiseman, A. W., & Baker, D. P. (2005). The worldwide explosion of internationalized education policy. In D. P. Baker & A. W. Wiseman (Eds.), *Global trends in educational policy* (pp. 1–21). London: Elsevier Science.