Chapter 2
Analysing the (Mis)Use and Consequences of International Large-Scale Assessments

Stefan Johansson

Abstract When insights are shared across borders, similarities in structures, polices, pedagogies and curricula can emerge. One global force is international large-scale assessments (ILSA), which have been criticized for spreading isomorphic ideologies. At the same time, ILSA data may have the potential to legitimize informed decisions, now covering long-term trend databases from many school-systems. Further, IEA encyclopedias, papers presented at IEA and PISA research conferences, and a growing volume of academic publications all point to numerous studies that draw on international assessment datasets to explore issues of pedagogy and classroom practice. Given the rigorous test administration of ILSA's, the data generated has the potential to provide nuanced snapshots of characteristics of different school-systems, provided that is that the data are used with caution. But are data used with caution? The current chapter discusses the use and possibilities of ILSA data and how results on ILSA's impact education and policy reforms world-wide.

Keywords Assessment · Comparative education · Education policy · Globalization · International large-scale assessments · PISA · Consequential validity · Policy impact

International Large-Scale Assessments and Their Aims

Why are we testing so many students in so many countries in so many subjects? The history of international large-scale assessments goes back more than half a century and the initial aims of the testing programme will be discussed below. The

S. Johansson (✉)
Department of Education and Special Education, University of Gothenburg, Gothenburg, Sweden
e-mail: stefan.johansson@gu.se

J. Zajda (ed.), *Globalisation, Ideology and Education Reforms*, Globalisation, Comparative Education and Policy Research 20,
https://doi.org/10.1007/978-94-024-1743-2_2

International Association for the Evaluation of Educational Achievement (IEA) was founded in 1958 with the aim of studying educational achievement and its determinants in different countries. One of the objectives was that countries could learn from the experiences of others and avoid developments that had been shown to produce unsatisfactory results. By collecting data researchers could analyze differences and similarities around the globe through an educational laboratory (Husén 1979; Walker 1976). The first IEA study, First International Mathematics Study, (FIMS64) was conducted in 1964, and in 1970–1971 the six-subject survey (SSS) was conducted, the latter including three populations and six subjects. Such an extensive project has not been carried out since then, although in recent years the number of international assessments has increased. After the SSS, there was a rather low level of activity within IEA until the 1990s. The notable exception was the Second International Study in Mathematics and Science (SIMS80 and SISS84). The TIMSS 95 study marked a new phase in the development of the international large-scale assessments (ILSA) of IEA (Gustafsson 2008). This phase was characterized by a less marked researcher scrutiny. The aim of the studies also shifted away from an explanatory focus towards descriptive purposes. This trend is evidenced in the national and international reports that are produced, where nowadays the emphasis is more on reporting descriptive outcomes than analyzing the factors behind them. Instead, large databases are made available for secondary analyses.

Another international organization carrying out large-scale studies is the Organization for Economic Co-operation and Development (OECD). Founded in 1961, OECD was initially an international economic organization comprising 34 countries with the aim of stimulating economic growth and world trade. The organization consists of highly developed countries that regularly meet to share policy experiences, seek answers to common problems, identify good practices, and to coordinate the domestic and international policies of its members. In more recent decades the aim and scope of OECD have been expanded. In 2000, OECD launched its Programme for International Student Assessment (PISA), which covers several subject domains, including mathematics, science and reading for 15-year old's. In each wave, one area forms the major domain, while the other two are minor domains, represented by a smaller number of items. PISA testing is conducted every third year in all OECD countries, along with many associate countries. In 2018, about 80 countries and economies participated in PISA.

There are many similarities between PISA and IEA studies such as samples with clearly defined populations, similar instruments, data collection processes, and psychometrical methods, and the implementation of rigorous quality control measures (Olsen 2005). Further, the studies have cyclic designs with a focus on measuring trends. Both organizations also provide country rankings, sometimes referred to as league tables, which attract substantial public attention when the results are launched. Although public attention may vary across countries, in many countries both IEA and OECD studies receive significant coverage in both the media and in policy debate.

Even though there are many similarities between the studies carried out by IEA and OECD, the organizations are different with respect to background and purpose. While the IEA, at least initially, aimed to provide research data for educational research, the OECD explicitly aims to impact policy and policy-making (Olsen 2005; Meyer et al. 2018). The OECD has developed new policy techniques to promote neoliberal ideas, performativity and management culture. According to Ball (2010) such ideas become increasingly important and vital for the governing of education in most European countries. OECD also issues data-based reports that describe how different countries can develop their school-system. Further, while the IEA studies focus on curriculum-defined knowledge and skills, the OECD studies attempt to capture competencies that are important (in the view of OECD experts) in adult life and for life-long learning (Lockheed and Wagemaker 2013).

ILSA's, like PISA, are standardized tests. But in comparison to other national tests and SATs they have a particular focus on the nations' average test-score as well as analyses of subpopulations such as groups with differing social background. Within the PISA consortium, for example, measurement organizations such as the ETS and ACER provide their expertise. PISA has thus employed some of the most acknowledged measurement people in the world. Still, there are validity issues with the tests, which, to large degree are inevitable when constructing a common test for so many countries in the world. One issue relating to validity is that questions are claimed to be context dependent and may be interpreted differently between different countries in the world. Likewise, some questions show differences for girls and boys. It is challenging to judge whether the differences in achievement are related to students' ability levels or if they a depending on different interpretations of the test items, however, there are aids to analyze this, such as Differential Item Functioning (DIF) analysis. In the case of PISA, the focus is on OECD countries and questions are constructed with a focus on these countries. Some 30 associate countries are also taking the tests and their culture and context may be somewhat different from the OECD countries. ILSA organizers, however, tries to avoid much of the cultural bias that can occur. Representatives from all participating countries come together discussing the item pool as well as the assessment of the items. Items producing cultural bias are likely to be removed.

Even though IEA's and OECD's studies are comparative in nature, it should be noted that the main focus is not to compare *all* countries' performances. Several decades ago, one of the founding fathers of IEA concluded that the IEA-project by no means had the objective to compare students' performances in all different countries. Diverse cultures, varying economies, as well as different epistemological beliefs all make it difficult to compare achievement across the range of different countries (Husén 1979). In spite of cautions about comparing test scores and rankings PISA results are informing policies in various countries (see, Klemenčič and Mirazchiyski 2018). One such example is put forward by Gorur and Wu (2014) who asked for a more nuanced analysis of ILSA results, before these should inform policy. Their example is situated in Australia's top five ambitions. However, based on a more detailed reading of the results Gorur and Wu show that the results vary very much across different regions of Australia. The averages the rankings displayed

were consequently of little use for informing policy decisions. Average scores obscure far more than they reveal the authors concluded.

Despite increased popularity of ILSA's in terms of number of studies and public attention, they have met a fair amount of criticism, from different angles and perspectives. One line of criticism mainly concerns the issues of the measurement. This includes both conceptual and technical aspects of the tests. The item format in TIMSS has been criticized (see for example, Schoultz, Säljö & Wyndhamn) as well as the scaling procedures and reporting of test-scores in PISA (Kreiner and Christensen 2014). The ILSA's' validity is however not only challenged for issues regarding item content and construct. Indeed, it has been argued that several consequences of ILSA's are not associated with the measurement per se. One claim about ILSA's is that they steer learning in different ways and at different levels of the school system. For example, in response to low algebra results in TIMSS, a country that did not place much emphasis on algebra prior to the 2000 may introduce mathematics reforms, increase the time for math in school, and increase the importance of algebra in the mathematics syllabus. Such behavior would imply that ILSA's are 'steering at a distance'. Thus, if ILSA results are fed back into the nations' policy-making process, this may lead national educational systems to develop similar models for schooling, thereby causing a trend of convergence in educational policy and practice among different countries. While this may be desired by some, critics could argue that such convergence may hamper creativity and uniqueness of single educational systems, which often are a stated goal in the acts of human rights in different countries.

Some Views on ILSA Tests, Their Scores and Their Consequences

The number of research studies performed by academic scholars and others on ILSA data is vast, and increasing over time. ILSA data may be a powerful tool for policymakers, in that they can legitimize and delegitimize school-reforms (Pettersson 2008). Since news and research findings travel fast it is perhaps increasingly important to nuance the findings with an eye on validity. In the following, I will sketch some examples using and evaluating ILSA's. Among others, I will point to aspects of the content, the constructs, and the consequences of ILSA's. One reason for undertaking such investigation is that these aspects are crucial in modern validity theory. The current mainstream view of validity is that it is based on an integration of any evidence that bears on the interpretation or meaning of test scores. Further, the boundaries of validity go beyond the meaning of tests scores to include relevance, utility values and social consequences (Kane 2006; Messick 1989).

Furthermore, I want to nuance research findings, because, for example, perceived truths are sometimes shallow at best and that critiquing claims are not always so straightforward as we might be led to believe. I will begin to review some major

findings related to ILSA's which have had high policy impact, thereafter I will explore some critique against test items, some critique against test-score comparisons, and finally some critique against ILSA policy impacts.

ILSA and Their Constructs

In elaborating the construct of knowledge represented by ILSA test items, several scholars have questioned the construct validity of ILSA pointing to the content of the items. For example, it has been argued that the tests do not constitute trustworthy representations of students' knowledge (Schoultz et al. 2001; Serder and Jakobsson 2015; Serder and Ideland 2016). From a sociocultural perspective it is thus not possible to know what the students are responding to in a test situation, for example because "the only clues are pencil marks in the multiple-choice boxes or some inscriptions to open-ended questions" (Serder and Jakobsson 2015, p. 835). Sociocultural theorists believe that it is necessary to create situations in which students' meaning-making of the test questions can be observed and analyzed in order to understand how and why they answer questions as they do. Proponents of ILSA argue that cultural and other differences are 'factored out' via the large pool of questions available in a PISA or TIMSS assessment. Also, they argue that to compare performances, testing needs to be carried out under standardized conditions otherwise the differences may be largely due to different test conditions. Further, the general notion of construct validity (see, Messick 1989; Kane 2006) would not refer to characteristics of single items but rather a set of items to generalize findings to broader constructs. One should also note that in ILSA, the test-scores of the individual students are not in focus. Rather measures of centrality and spread that assume significance so groups can be compared.

Another example put forward by Berliner (2018) relates to test validity. He is concerned with the quite substantial differences between national raw scores and the scaled scores (plausible values) in PISA 2015. By comparing national raw scores and scaled scores Berliner finds that, for example, Slovenia and the US have equal raw scores (for a subset of items and students) but quite different scaled scores and rankings. But, reporting raw scores ignores the differences of ability of subgroups to which the set of items was administered. However, the trustworthiness does not, because comparing raw scores and scaled scores is in fact like comparing apples and oranges. Let me elaborate a bit on the sampling in ILSA's.

The test design in PISA is based on matrix sampling where each student is administered a subset of items from the total item pool. For example, in PISA 2018, there where nearly 250 questions in the pool for the reading domain. Each student receives a test form or booklet comprising of four 30-min clusters, assembled from two subject domains. In 2018, reading was the core subject and two clusters in every test form comprised reading items. For countries taking reading math and science there were 36 test forms and different groups of students answered these (but only one). The items in the test forms are overlapping to certain degree.

Inevitably, the different items of the forms make it inappropriate to use any statistic based on the number of correct responses in reporting the survey results. Differences in total scores, or statistics based on them, among students who took different sets of items may be due to variations in difficulty of the test forms or the level of ability of the group of test-takers. Unless strong assumptions are made, for example, that the different test forms are perfectly parallel, the performance of two groups assessed in a matrix sampling arrangement cannot be directly compared using raw scores.

The reasons for adapting a matrix sampling design are many, but mainly it is a way to enhance the validity of the outcome measure – the national test score. For example, in ILSA's, test forms are kept relatively short to minimize individuals' response burden. This is probably especially important in ILSA's because they are low-stake assessments for the students, i.e., they do not provide any feedback to the student. While the test form of an individual student could suffer from construct underrepresentation for, say, the reading domain, the few responses for each student will form a wide range of content representation when responses are aggregated.

The PISA technical report (OECD 2017) emphasizes that plausible values are not substitutes for test scores for individuals. Since they incorporate student responses to test items and information about their background characteristics of the student in an IRT model scaled scores cannot be used to compare individuals. These, so called plausible values (PVs) are combining the IRT scaling of the test items with a latent regression model using information from the student context questionnaire in a population model. PV's are constructed explicitly to provide consistent estimates of population effects.

ILSA and Consequences

Naturally, many policy makers see the value of investing in education. Education is a good investment for all kinds of reasons; not only for countries' economy but also for the personal well-being and development. However, the economic incitements play an important role as the improvements in educational achievement, as measured by PISA results, has been related to economic growth. In fact, one reason to PISA's policy impact may be the strong statistical claims showing that improvements in ILSA will lead to higher Gross Domestic Product (GDP) growth rates (see, for example, Sahlberg 2006; Hanushek and Woessmann 2007, 2015). As the scores of ILSA has been related to economic outputs the interest in raising test scores has gained much interest.

However, how scores on ILSA's relate to economic growth has recently been challenged. In a study by Komatsu and Rappleye (2017a, b) the strong correlation between improvement in ILSA and growth in GDP is questioned. In their study, Hanushek and Woessmann (2015) compared test scores and economic growth over the same period (1960–2000) and showed a strong relationship. But Komatsu and Rappleye (2017a, b) claimed that it is reasonable that it takes a few decades for

students to be a part of the work force; improving scores today would consequently not lead to improved GDP in the near future but in a couple of decades ahead. By comparing test scores for one period with economic growth in a subsequent period (1995–2014) the relationship decreased dramatically. While the results of Hanushek and Woessmann (2015) speak for high relation between GDP and PISA results, Komatsu's and Rappleye's results show a modest association between the two measures. The relationship is substantial (R2 = .57) when comparing the scores and growth under the same period while it is rather weak when using a subsequent period (R2 = .10). Though it makes sense that educational achievement should impact economic prosperity to some degree, it is reasonably more complex to determine than by comparing test scores and GDP.

There are many examples of countries that are willing to reform some parts of the school-system after ILSA results. Changes may already be 'in the air' and ILSA results may be the provocation to make the changes. ILSA may also be used to delegitimize decisions which already have been made (e.g., Pettersson 2008). It is important to emphasize that changes in curricula do not happen within a short time span. Already in the wake of the results of the first mathematics survey in 1964 in Sweden, it was discussed that different school reforms did not affect student results. Then it was pointed out that changes were hardly possible to detect from one year to another, rather impact takes decades. It is reasonable to believe that this holds true today as well. Actual knowledge patterns are also influenced by the teachers and their interpretations of the curriculum. Knowledge patterns are also strongly influenced by knowledge traditions for a long time. In the following I will give two examples of impact that ILSA nevertheless has been claimed to have.

On the Relation Between Innovativeness and ILSA Results

Zhao (2012) finds it reasonable to question the value and the significance of educational excellence measured by PISA since high-performing East Asian countries perceive their entrepreneurial capability to be low (Zhao 2012, p. 59). Zhao supports his claims of lacking innovativeness in East Asia with the results of the Global Entrepreneurship Monitor (GEM) study, which is a survey of perceived entrepreneurial capability in a wide range of countries (Bosma et al. 2012). As a comparison to the GEM scores, Zhao uses the mathematics results of PISA 2009. The comparison is striking. The comparison shows that high-performing countries in PISA, such as Japan, Korea, Singapore and Finland, all have low scores on their *perceived* entrepreneurial capability. At the same time, moderately performing countries like Sweden and USA had fairly high perceived entrepreneurial capability, while a low-performing country such as United Arab Emirates had students who estimates the entrepreneurial capability to be in the top. Zhao (2012) concluded that this relationship indeed can be causal, and that policy reforms intending to increase subject knowledge must be stopped, if not, students' creativity will be seriously harmed. Johansson (2018) thought this pattern rather had to do with the self-reported

measured used in GEM rather than with East Asians innovativeness. He studied scores in PISA with scores on academic self-concept in math and found the same pattern as Zhao. Students in low achieving countries assessed the scores to be high in both mathematical ability and entrepreneurial/creative ability, whereas the students in high-performing countries did the opposite. Probably the same pattern would emerge irrespective of the type of self-assessment. There is reason to believe that the test scores are unrelated to the entrepreneurial ability of the East Asian's. A possible explanation of this recurrent paradox may be related to countries response styles to attitude surveys in large-scale cross-national assessments, such as the PISA.

There is more evidence that PISA rankings and innovation do not correlate. Berliner (2018) concludes that entrepreneurship is a better predictor of nation's future economy than is ILSA results. It is also shown that the rank on the global innovation index 2016 and rank on PISA 2015 are poorly related. Without going into detail the global innovation index uses 82 different metrics of innovativeness to determine rankings. The comparison seems relevant at first place, however, using the same line of reasoning as Komatsu and Rappleye (2017a, b) when discussing the results of Hanushek and Woessmann, there is reason to believe that the lagging effects have been forgotten. The possible effect of innovation in a country and its PISA are not on the same time scale. First PISA effects are unlikely when people first enter the labor market. It would be unreasonable to expect any effects students are 15 years old. The global innovation index does not give a clear pattern either however. Examining the results from 2018 we can note Switzerland still in top, followed by Netherlands and Sweden. Singapore are in fifth place now passing the US who are in sixth place. Countries similar to the top countries, and also performing similar in PISA 2015 – Norway, Canada and Austria come first on around 20th place, after nations like China, Hong Kong, Korea and Japan. The impression is that high-performing countries are gaining positions on the innovation index while several rich countries are lagging behind.

On the Changes of Curricula and ILSA Results

Currently, individual countries' strivings to come out at the top of ranking tables seem to have created a homogenizing of school-systems. When insights are shared across borders, similarities in structures, polices, pedagogies and curricula can emerge. Such homogenization may not necessarily be positive. Consequently the 'borrowing' of policies has become one of the main criticisms levelled against ILSA's. According to Spring (2008), world cultural theorists claim that a combination of international tests and the sharing of international research are resulting in a global homogeneity of instructional practices. In this case, and in spite of obvious national differences, such as language, culture and religion, it is concluded that countries become more similar over time—a process often referred to as 'isomorphism' (Wiseman et al. 2013). Borrowing and lending of educational policies are mechanisms of change that are subtle and difficult to chart and analyze. One reason

for this is that official descriptions of national educational policy for strategic reasons need not reflect actual practices. Thus, what has been written on paper, does not imply change in actual teaching or achievement. For example, it has been problematic to establish any reliable evidence for that global influences (e.g., ILSA's) would lead to actual convergence in student achievement, or attained curricula. Researchers within the organization of IEA developed a three-level model of the context and components of the school curriculum in order to shed light on the different levels of curricula, which are important for students' learning opportunities (e.g., Keeves 1972; Robitaille and Garden 1989).

At the level of an educational system (the school-system, the educational region, the school district) there is a set of *intentions for the curriculum*. There are goals and traditions. There are impulses from the community of educators that help shape the character of the curriculum. This collection of intended outcomes, together with course outlines, official syllabi, and textbooks forms an intended curriculum. The second level deals with the classroom, the setting in which the content becomes *implemented* or translated into reality by the teacher. The classroom is central to the educational process, because in the classroom that children are introduced to the study of mathematics, for example, and it is where their concepts and attitudes are formed, and it is the teacher who has the responsibility for transmitting this knowledge to students. The final level of this model represents the *attained curriculum***.** After a given period of time at school, the student has acquired a body of mathematical knowledge, and acquired certain attitudes toward the subject. Thus, what aspects of the curriculum as intended, say, by a national agency of education and taught by the teacher, are actually learned by the student?

If in fact nations are borrowing educational policies, comparing educational systems, and setting educational benchmarks based on recommendations from an international agenda, one would expect to see an increased similarity in students' responses on international educational assessments over time. In other words, an increasingly similar curriculum should produce an enhanced similarity between international responses to assessment items that are intended to measure curriculum. In that country-level trend data is available in ILSA's these seem therefore well suited to investigate whether there is a trend towards convergence or curricular harmonization. As the effects of global processes previously have been hard to study empirically, the ILSA's provide a unique approach to study how educational policies developed across countries and over time.

Johansson and Strietholt (2016) investigated if countries converge with respect to their patterns of knowledge in mathematics subdomains, aiming at framing the larger question about a trend towards isomorphism in countries' curricula. The results showed little evidence for a convergence at global level, however, there was compelling evidence that tradition and culture is a strong force when it comes to students' content knowledge. Similarities in culture and language seem to have great impact on the knowledge patterns since countries within same region/culture/language clustered in many occasions. Furthermore, there seem to be a general trend emerging, in that many countries do not have any pronounced strengths or weaknesses in later TIMSS; they perform fairly similar in all subdomains. This

might express that students in these countries had the opportunity to learn the tested content, and that the link between the intended, implemented and attained curricula is strong. However, the countries without relative strengths and weaknesses perform quite differently in absolute terms. Some of the countries are among the top performers and some are among the low performers. If the intended curricula overlap substantially across all these countries, there is likely something else that differs (teacher instruction, school resources, etc.) to a great deal since the variation is significant among these countries.

Another interpretation of the trend towards less pronounced strengths and weaknesses may be that certain countries are "teaching to the test" (see, for example, Biggs 1999) to higher degree than others. The concept of teaching to the test could subsume desirable as well as undesirable behaviors, and the tests in TIMSS should be aligned to the curriculum goals. It seems anyway reasonable that the focus on the tests in ILSA's, such as TIMSS varies across countries, thus the ILSA results are more high-stake for certain countries, than for others (c.f., Grek 2009). In recent past, some countries' performances, including Singapore's, have become more focused on algebra and geometry, previously performing evenly in the four subdomains.

Conclusion

Based on performances on ILSA's countries act in different ways. One action may be school-reforms like curriculum change and decentralization of school-systems. Thereby, the consequential validity of the ILSA's has been questioned. Within a validity framework, it is merely not enough to evaluate the validity of, for example, single items in ILSA's. Their subject constructs, relation to curricula in the assessments of IEA, their purpose and impact are factors on different levels that have to be considered when approaching the unified concept of validity. From the literature it can be concluded that several researchers have identified threats to the validity of ILSAs (e.g., Berliner 2018; Baker and LeTendre 2005; Zhao 2012). The literature is both vast and diverse, and ILSAs have been criticized for various having types impact – from the simplification of the school-debate and concept of knowledge, to the travelling of policies across nations. Without doubt, the results of ILSAs have a substantial impact in the media, in discussions of policy educational policy, as well as in public debate (Carvalho and Costa 2015; Nóvoa and Yariv-Mashal 2003). However, there seems to be a need for empirical studies addressing ILSA's long-term effects on globalization in education.

The ways in which global processes impact on educational systems around the world is a contentious issue, and there is no clear resolution in sight. The arrival of the Internet a couple of decades back, increased global traveling, and international trade all have influence on us and bring with them an enormous transformative potential on national cultures. Even though criticized, perhaps one of the greatest

benefits as regards ILSAs are the benefits associated with the production of data generated from the numerous studies undertaken, constituting a comparative element comprising up to 60 educational entities. Data have a longitudinal component at the country level, which facilitates opportunities to investigate causal effects of the impact of different reforms in different countries. One might wonder, therefore, if not analyses based on ILSA data should guide policy initiatives, what else should?

References

Baker, D., & LeTendre, G. (2005). *National differences, global similarities: World culture and the future of schooling*. Stanford: Stanford University Press.

Ball, S. J. (2010). New voices, new Knowledges and the new politics of education research: The gathering of a perfect storm? *European Educational Research Journal, 9*(2), 124–137. https://doi.org/10.2304/eerj.2010.9.2.124.

Berliner, D. (2018). PISA is simply another standardized test. In S. Lindblad, D. Pettersson, & T. S. Popkewitz (Eds.), *Education by the numbers and the making of society. The expertise of international assessments*. New York/London: Routledge.

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development, 18*(1), 57–75. https://doi.org/10.1080/0729436990180105.

Bosma, N., Wennekers, S., & Amorós, J. (2012). *Global entrepreneurship monitor: 2011 extended report: Entrepreneurs and entrepreneurial employees across the globe*. London: Global Entrepreneurship Research Association.

Carvalho, L. M., & Costa, E. (2015). Seeing education with one's own eyes and through PISA lenses: Considerations of the reception of PISA in European countries. *Discourse: Studies in the Cultural Politics of Education, 36*(5), 638–646. https://doi.org/10.1080/01596306.2013.871449.

Gorur, R., & Wu, M. (2014). Leaning too far? PISA, policy and Australia' 'top five' ambitions. *Discourse: Studies in the Cultural Politics of Education, 36*(5), 1–18. https://doi.org/10.1080/01596306.2014.930020.

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24*(1), 23–37. https://doi.org/10.1080/02680930802412669.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal, 7*(1), 1–17. https://doi.org/10.2304/eerj.2008.7.1.1.

Hanushek, E. A., & Woessmann, L. (2007). *Education quality and economic growth*. Washington, DC: The World Bank.

Hanushek, E. A., & Woessmann, L. (2015). *The knowledge capital of nations: Education and the economics of growth*. Cambridge, MA: MIT Press.

Husén, T. (1979). An international research venture in retrospect: The IEA surveys. *Comparative Education Review, 23*, 371–385.

Johansson, S. (2018). Do students' high scores on international assessments translate to low levels of creativity? *Phi Delta Kappan, 99*(7), 57–61. https://doi.org/10.1177/0031721718767863.

Johansson, S., & Strietholt, R. (2016). Konvergieren Leistungsprofile in Mathematik? Evidenz aus fünf IEA Studien. In R. Strietholt, W. Bos, H.-G. Holtappels, & N. McElvany (Eds.), *Jahrbuch der Schulentwicklung. Band 19 – Daten, Beispiele und Perspektiven*. Weinheim Basel: Beltz Juventa.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Washington, DC: American Council on Education and National Council on Measurement in Education.

Keeves, J. P. (1972). *Educational environment and student achievement*. Stockholm: Almqvist and Wiksell.

Klemenčič, E., & Mirazchiyski, P. V. (2018). League tables in educational evidence-based policy-making: Can we stop the horse race, please? *Comparative Education, 54*(3), 309–324. https://doi.org/10.1080/03050068.2017.1383082.

Komatsu, H., & Rappleye, J. (2017a). A new global policy regime founded on invalid statistics? Hanushek, Woessmann, PISA, and economic growth. *Comparative Education, 53*(2), 166–191. https://doi.org/10.1080/03050068.2017.1300008.

Komatsu, H., & Rappleye, J. (2017b). A PISA paradox? An alternative theory of learning as a possible solution for variations in PISA scores. *Comparative Education Review, 61*(2), 269–297. https://doi.org/10.1086/690809.

Kreiner, S., & Christensen, K. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to Reading literacy. *Psychometrika, 79*(2), 210–231. https://doi.org/10.1007/s11336-013-9347-z.

Lockheed, M., & Wagemaker, H. (2013). International large-scale assessments: Thermometers, whips or useful policy tools? *Research in Comparative and International Education, 8*(3), 296–306.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillian.

Meyer, H. D., Strietholt, R., & Epstein, D. Y. (2018). Three models of global education quality and the emerging democratic deficit in global education governance. In M. Akiba & G. K. LeTendre (Eds.), *Routledge international handbook of teacher quality and policy*. New York: Routledge.

Novóa, A., & Yariv-Mashal, T. (2003). Comparative research in education: A mode of governance or a historical journey? *Comparative Education, 39*(4), 423–438.

OECD. (2017). *PISA 2015 technical report*. Paris: OECD.

Olsen, R. V. (2005). *Achievement tests from an item perspective. An exploration of single item data from PISA and TIMSS studies, and how such data can informm us about students' knowledge and thinking in science*. PhD dissertation, University of Oslo.

Pettersson, D. (2008). *Internationell kunskapsbedömning som inslag i nationell styrning av skolan* [International knowledge assessments: an element of national educational steering]. PhD dissertation, Uppsala University.

Robitaille, D. F., & Garden, R. A. (1989). *The IEA study of mathematics II. Context and outcomes of school mathematics*. Pergamon: Oxford.

Sahlberg, P. (2006). Education reform for raising economic competitiveness. *Journal of Educational Change, 7*(4), 259–287. https://doi.org/10.1007/s10833-005-4884-6.

Schoultz, J., Säljö, R., & Wyndhamn, J. (2001). Conceptual knowledge in talk and text: What does it take to understand a science question? *Instructional Science, 29*(3), 213–236. https://doi.org/10.1023/A:1017586614763.

Serder, M., & Ideland, M. (2016). PISA truth effects: The construction of low performance. *Discourse: Studies in the Cultural Politics of Education, 37*(3), 341–357. https://doi.org/10.1080/01596306.2015.1025039.

Serder, M., & Jakobsson, A. (2015). "Why bother so incredibly much?": Student perspectives on PISA science assignments. *Cultural Studies of Science Education, 10*(3), 833–853. https://doi.org/10.1007/s11422-013-9550-3.

Spring, J. (2008). Research on globalization and education. *Review of Educational Research, 78*(2), 330–363. https://doi.org/10.3102/0034654308317846.

Walker, D. A. (1976). *The IEA six subject survey: An empirical study of education in twenty-one countries*. Uppsala: Almqvist & Wiksell International.

Wiseman, A. W., Astiz, M. F., & Baker, D. P. (2013). Comparative education research framed by neo-institutional theory: A review of diverse approaches and conflicting assumptions. *Compare: A Journal of Comparative and International Education, 44*(5), 688–709. https://doi.org/10.1080/03057925.2013.800783.

Zhao, Y. (2012). Flunking innovation and creativity. *Phi Delta Kappan, 94*(1), 56–61. https://doi.org/10.1177/003172171209400111.