

PISA: a political project and a research agenda

Svein Sjøberg ^a and Edgar Jenkins^b

^aUniversity of Oslo, Norway; ^bUniversity of Leeds, England

ABSTRACT

PISA (Programme for International Student Assessment) is one of two large scale international comparative projects of student assessment that now exert considerable influence upon school science education policy, the other being TIMSS (Trends in International Mathematics and Science Study). This paper focuses on PISA, now the most influential study. This article outlines the origins of PISA, identifies some of the challenges in its construction and the claims made for it. It argues that while the statistical and methodological aspects of PISA have received much research attention, other elements of PISA have been largely ignored. In particular, there are several outcomes of PISA testing that point towards a significant research agenda. In addition, the political, ideological and economic assumptions underpinning the PISA project have implications for school science curriculum policy that deserve closer scrutiny and debate.

ARTICLE HISTORY

Received 23 June 2020

Accepted 7 September 2020

KEYWORDS

Science education; PISA; comparative education; globalisation; education policy

PISA: origins and objectives

Large-scale international studies of educational achievement have a long history (IEA, 2018). Today, two such studies, PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) have come to dominate the field. However, the two projects differ in several important ways. Unlike TIMSS, which is basically descriptive and analytical, PISA is explicitly and intentionally normative. TIMSS is basically driven by researchers; while PISA is owned and governed by member states in the OECD (Organisation for Economic Cooperation and Development). Other differences include testing different student cohorts, the frequency of testing and the relationship of test questions to school curricula. Whereas TIMSS test items are closely related to school curricula, PISA test items are meant to address real life challenges. Both studies measure trends in test scores over time. Further details of the differences between TIMSS and PISA are summarised in the Appendix to this article.

PISA testing began in 2000, the first results being published in December of the following year. Subsequent testing has taken place every three years with science being one of the three core subjects. In each round of testing, one of these subjects is allocated 60 per cent of test time. Science was the core subject in PISA 2006 and PISA 2015. Each PISA test now includes an optional assessment of an 'innovative domain'. These range from Learning Strategies (2000) and Complex Problem Solving (2003) to

CONTACT Svein Sjøberg  svein.sjoberg@ils.uio.no

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Collaborative Problem Solving (2015) and Global Competencies (2018). Creative Thinking is meant to be the domain included in PISA 2021. The technical details of PISA, elaborated in detailed manuals and subsequent technical reports, are complex. PISA 'league tables' receive wide publicity and in many countries the data prompt policy makers to undertake educational reform (Breakspear, 2012). By late 2020, data are available from seven rounds of PISA testing, the most recent from PISA 2018 (OECD, 2019a, 2019b, 2019c, 2019d).

Originally intended for the 30+ industrialised and wealthy OECD countries, the project has expanded to include many other countries, regions and economies. This allows it to claim that participants in PISA 'make up nine tenths of the world economy' (OECD, 2010, p. 3). A PISA study is inevitably expensive to conduct. One analysis of each round of PISA testing in the USA has been estimated to cost approximately 6.7 million USD, with additional costs being incurred by individual States and in paying teachers and school coordinators to participate (Engel & Rutkowski, 2018). Key elements of developing and reporting PISA are sub-contracted to external providers, like Pearson Inc. and ETS (Educational Testing Service).

Unlike most tests, including TIMSS, PISA test items are frequently based on pieces of text designed to present students with an 'authentic situation'. These texts place a premium on reading competence, leading some commentators to suggest that PISA items test reading skills rather than science or mathematics. The fact that the correlations between individuals' PISA scores on reading, mathematics and science across all countries tested are 0.77–0.89 (OECD, 2005) lends some support to the view that testing in the different domains measures more or less the same underlying construct.

Interpreting PISA results

Given the importance attached to PISA results by legislators and others, it is important to caution against accepting some of the results at their face value. The population targeted for testing is not always what it seems. For example, in Vietnam, only 56% of 15 year olds attend schools so that it is difficult to justify a claim that Vietnamese schooling is a 'stunning success' (Schleicher, 2015; Sellar et al., 2017, p. 44). The performance of schools in China has been presented on the basis of a sample of schools and/or students in a particular region of the country. In 2015, when data from Shanghai was combined with those from other Chinese-sub-national systems, the students' performance in science was not significantly different from that of the United Kingdom, Slovenia or Australia, among others (Sellar et al., 2017, p. 32). The exclusion rate, that is the proportion of the eligible students that are excluded from taking the test, also varies considerably from country to country and can also change from one round of PISA testing to another. In Norway, for example, the exclusion rate in the first round of PISA testing was 2.7% of the 15 year old cohort; by 2018, it had risen to 7.9% (Jensen et al., 2019, p. 25).

Education systems widely separated in PISA league tables often have different PISA scores that are not statistically significant. Wuttke (2007) studied the uncertainty in PISA results for Germany and concluded that the 'Statistical significance criteria of OECD PISA are misleading because the several sources of systematic bias and uncertainty are quantitatively more important than the standard errors communicated in official reports' (Wuttke, 2007).

A further issue arises in the attempt to record trends in test performance over time. In order to do this, PISA tests contain a small number of items that are unchanged from one test to another. When allied with sampling errors, this use of a small number of 'link items' leads to an unacknowledged uncertainty in reporting the estimates of achievement over time (Sellar et al., 2017, p. 51).

A PISA test consists of items that cover about ten hours testing time, but each student answers only a two-hour sample of these items. The statistical procedures that link individual test scores to the published parameters such as PISA mean scores have been seriously challenged. Soon after the publication of the results of PISA 2006, the Danish statistician Svend Kreiner presented a critique of the scaling methods used to calculate the PISA scores. By re-analysing the publicly available PISA data files, Kreiner demonstrated that the procedures used by PISA could result in placing countries very differently in the PISA rankings: the PISA scaling methods could put Denmark on anything from PISA rank 2 to 42, depending on how it was used. This critique was basically ignored by PISA. In later publications, Kreiner and his colleague Christensen developed and concretised their critique in several articles in highly respected journals. In 2014 they addressed 'some of the flaws of PISA's scaling model' and questioned the robustness of PISA's country rankings (Kreiner & Christensen, 2014). This critique was then taken seriously and was influential in changing PISA's procedures with respect to the 2015 data. This change of scaling model caused the resulting PISA scores of some countries to jump dramatically, much more than deemed educationally possible for a three year period.

Towards a research agenda

PISA scores and students' interest in, and attitudes towards, science

PISA tests include a student questionnaire that has many questions designed to probe young people's attitudes towards science. This was an important element of the PISA 2006 study, when science was the core subject for the first time. The definition of science literacy in PISA 2006 included 'willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen' (OECD, 2006). A special issue of the *International Journal of Science Education* (2011, 33, (1)) presented several interesting results from an analysis based on these data.

One finding is that many countries with the highest mean PISA science score were at the very bottom of the ranking of students' interest in science (Bybee & McCrae, 2011).

Finland and Japan are prime examples both being at the top of PISA science scores but at the very bottom on constructs such as 'interest in science', 'future-oriented motivation to learn science' as well as on 'future science job', that is inclination to see themselves as scientists in future studies and careers. In fact, the PISA science scores correlates negatively with Future science orientation ($r = -0.83$) and with Future science job ($r = -0.53$) (Kjærnsli & Lie, 2011).

It should be noted that these negative relationships occur when *countries* are the units of analysis. When individual students within each country are the units of analysis, some of the correlations are positive.

Although applying the statistical inference from differences between groups to individual differences is an ecological fallacy, the findings remain disturbing. If students in PISA top-ranking countries leave compulsory schooling with a strongly negative

orientation towards science, it is important to identify the reasons and the possible consequences. Correlation is of course not to be identified with causation but there is a clear pointer to the need for caution in countries that score highly in PISA science tests as role models for reform elsewhere.

In an analysis of the PISA 2015 data, Zhao (2017) pointed out that students in the so-called PISA-winners in East-Asia (Japan, Korea, Hong Kong, Singapore) seemed to suffer from what he called the 'side-effects' of the struggle to get good marks and tests-scores. He draws upon PISA data to show that students in these countries get high scores but have very low self-confidence and self-efficacy related to science and mathematics. Zhao points out that

There is a significant negative correlation between students' self-efficacy in science and their scores in the subject across education systems in the 2015 PISA results. Additionally, PISA scores have been found to have a significant negative correlation with entrepreneurial confidence and intentions (Zhao, 2017).

Science educators might reasonably conclude that there is a need for a deeper understanding of the relationship between PISA science scores and measures of student attitudes and interest. Attitudes are difficult to measure reliably and it may be that the perception that students have of science as a result of their school studies differs from their perception of science beyond the world of school.

It is important to remember that although the PISA definition of 'science literacy' includes interest in science and other attitudinal and affective aspects, these are not part of the actual PISA test score. They are difficult to measure, but some are partly addressed in the student questionnaire. As indicated above, these important aspects of science literacy do often not correlate positively with the scores on the basically cognitive items in the main PISA test.

PISA and gender differences

Many of the countries whose students score highly in PISA science tests have the largest gender differences in performance. Finland is a prime example. Finnish girls strongly outperform boys on all three PISA subjects. In reading literacy, the difference in means is about 50 % of a standard deviation. In addition, a robust finding of PISA and other reading tests such as PIRLS (Progress in International Reading Literacy Study) is that girls outperform boys in all countries. However, PISA test scores in science and mathematics follow a gender pattern that is different from, for example, the results of TIMSS testing. These findings contrast with the more familiar pattern of national examinations where boys frequently outperform girls in science and mathematics. Is it possible that these differences stem, at least in part, from the nature of PISA testing which places heavy demands on reading competence?

PISA and inquiry-based teaching

The concept of science as inquiry has a long history and recent years have seen a resurgence in interest among policy-makers. IBSE (inquiry-based science education) was the key recommendation in the influential EU-document 'Science Education Now'

(EU, 2007) and it is now widely advocated. The term IBSE was adopted as the key concept in calls for EU-funding in the Horizon 2020-programme. IBSE also plays a major role in the recommendations in the International Council for Science reports to the individual science organisations world-wide (ICSU, 2011) and in the current international science education initiatives of The European Federation of National Academies of Sciences and Humanities. ALLEA (ALL European Academies) (<https://allea.org/science-education/>).

In PISA 2015, where science was for the second time the core subject, nine statements in the student questionnaire constituted an *Index of inquiry-based teaching*. These statements included: 'Students spend time in the laboratory doing practical experiments'; 'Students are required to argue about science questions'; 'Students are asked to draw conclusions from an experiments they have conducted'; 'Students are allowed to design their own experiments' and 'Students are asked to do an investigation to test ideas' (OECD, 2016c, p. 69). Among the interesting findings is that in most of the 'PISA-winners' (Japan, Korea, Taiwan, Shanghai, Finland) students report very little use of inquiry-based teaching.

In terms of the variation within a given country, PISA concludes that 'in no education system do students who reported that they are frequently exposed to enquiry based instruction [...] score higher in science.' (OECD, 2016c, p. 36)

Although the relationship between IBSE and PISA test scores is negative, it is a different story with respect to interest in science, epistemic beliefs and motivation for science-oriented future career

... across OECD countries, more frequent inquiry-based teaching is positively related to students holding stronger epistemic beliefs and being more likely to expect to work in a science-related occupation when they are 30.. (OECD, 2016c, p. 36)

One of the questions in the Inquiry Index is of particular interest. Experiments play a crucial role in science and play an important role in science teaching at all levels. But when it comes to PISA results, 'activities related to experiments and laboratory work show the strongest negative relationship with science performance' (OECD, 2016c, p. 71).

Key concepts and acronyms in current thinking in science education are well-known: science in context, inquiry-based science education (IBSE), hands on-science, active learning, NOS (nature of science), SSI (socio-scientific issues), argumentation, STS (Science, Technology and Society). There seems to be no evidence from PISA to lend support to any of these pedagogical strategies. Indeed, PISA findings seem to suggest that they hinder attainment. Sjøberg (2018a) fears that the struggle to increase PISA scores may result in neglecting experimental and inquiry-based teaching in schools. A more detailed analyses of PISA data in six countries has been undertaken by Oliver et al. (2019).

This conflict between the recommendations and priorities of scientists as well as science educators on the one hand, and PISA results on the other is highly problematic and requires investigation.

PISA and ICT

The student background questionnaire in PISA includes several questions regarding the use of Information and Communication Technology (ICT) in schools, and has two constructs based on these questions. One construct or index is related to the use of the internet at school, the other to the use of software and educational programs. In

a detailed study of the five Nordic countries, Kjærnsli et al. (2007) documented a clear negative relationship between the use of ICT and PISA score. It is also interesting to note that a PISA 'winner', Finland, is not only by far the Nordic country with the least use of ICT but its usage is also below the OECD average. In contrast, whereas Norway makes the most use of ICT in schools of all the OECD countries it has only has average PISA scores. In a special OECD/PISA report on the use of computers in teaching and learning (OECD, 2015), the highlighted conclusions are strikingly clear:

What the data tell us. Resources invested in ICT for education are not linked to improved student achievement in reading, mathematics or science. [...] Limited use of computers at school may be better than no use at all, but levels of computer use above the current OECD average are associated with significantly poorer results. (OECD, 2015, p. 146)

In spite of these clear findings, many countries, including Norway, strongly promote more ICT in schools, in order to climb the PISA rankings. While this is just one example of the selective readings of PISA results to justify reforms and initiatives, it also offers fertile ground for research.

PISA and the problem of translation

The problems associated with the translation of PISA questions from one language to another are well illustrated by an item on cloning released in 2006 and reproduced below (<https://www.oecd.org/pisa/38709385.pdf>) accessed 23 August 2020)

S128: Cloning

Read the newspaper article and answer the questions that follow.

A copying machine for living beings?

Without any doubt, if there had been elections for the animal of the year 1997, Dolly would have been the winner! Dolly is a Scottish sheep that you see in the photo. But Dolly is not just a simple sheep. She is a clone of another sheep. A clone means: a copy. Cloning means copying 'from a single master copy'. Scientists succeeded in creating a sheep (Dolly) that is identical to a sheep that functioned as a 'master copy'.

It was the Scottish scientist Ian Wilmut who designed the 'copying machine' for sheep. He took a very small piece from the udder of an adult sheep (sheep 1).

From that small piece he removed the nucleus, then he transferred the nucleus into the egg-cell of another (female) sheep (sheep 2). But first he removed from that egg-cell all the material that would have determined sheep 2 characteristics in a lamb produced from that egg-cell. Ian Wilmut implanted the manipulated egg-cell of sheep 2 into yet another (female) sheep (sheep 3). Sheep 3 became pregnant and had a lamb: Dolly.

Some scientists think that within a few years it will be possible to clone people as well. But many governments have already decided to forbid cloning of people by law.

Question 1

Which sheep is Dolly identical to?

- A Sheep 1
- B Sheep 2
- C Sheep 3
- D. Dolly's father

Question 2

The 'very small piece' is

- A. a cell

- B. a gene
- C. a cell nucleus
- D. a chromosome

The difficulties arose when the text and associated questions were translated from English into Swedish, Danish and Norwegian, three languages that are very similar and share a common literary tradition. All three Scandinavian texts changed the word 'nucleus' in the text to 'cell nucleus' thereby offering a significant hint to the correct answer to question 2. The Danish text altered question 1 to ask 'Which sheep is Dolly a copy of?' Thereby bringing the item closer to the newspaper heading. Other important changes in the wording were also made.

A more recent example recently released by PISA required a digital answer (available from <http://www.oecd.org/pisa/test/>). Entitled 'Running in Hot Weather', the item invited students to address the issues of overheating and dehydration that can arise when running in hot weather under different conditions of humidity. The key term *dehydration* is correctly translated into Norwegian and Danish as *dehydrering* but in the Swedish version of the item it appears as the much simpler, everyday word *uttorkad* the literal meaning of which is 'dried up'.

A further problem is that the need for comparability of translated items can lead a text to become clumsy and awkward, thereby reducing students' motivation to give the necessary attention. In most public examinations, questions are set upon largely prescribed curricula and there is a tacit or explicit understanding between teachers, students and examiners about what it is reasonable and acceptable to test. This is not the case in PISA so that even when students are being assessed in their first language, more needs to be known about the sensitivity of such responses to the form of words used in test questions and the context in which they are set.

PISA and its relationship to economic development

The importance of human resources as prime drivers in the modern economy is the foundation upon which the PISA project rests, a foundation known as Human Capital Theory. The human resources of a work-force in a modern economy are considered to be even more important than other forms of capital such as machines, buildings and infrastructure. The efficient development of a productive work-force thus becomes the key to economic development. From this perspective, expenditure on education is principally seen as an investment in future economic growth and competitiveness.

An important corollary of this perspective which has become something of 'a given' is that high scores on science and mathematics tests at school come to be regarded as key indicators of such growth and competitiveness. Disappointing PISA results and rankings on PISA are therefore to be avoided and appropriate corrective action needs to be taken.

The importance now attached to education and economic prosperity owes much to the work of Eric Hanushek, often considered to be the father of the field of 'school effectiveness'. He advocates the highly controversial Value Added Model for calculating the 'value added' effect that a school or a teacher has on student learning. Results from these calculations are then used in accountability-systems. In the USA, for example, the model is used to rank schools and individual teachers, to determine salaries and to dismiss teachers or principals if they don't 'deliver' satisfactory results. Hanushek's work is widely

used by the World Bank and the OECD in its analysis of the relationship between economic investment and educational quality.

In collaboration with Woessman, Hanushek authored an OECD report on 'The long run Economic Impact of Improving PISA Outcomes' (OECD, 2010). This report includes data that shows how much an individual country would gain by improvements in its PISA-score. As an example, the authors assert that an increase in 25 PISA points (a quarter of a standard deviation) over time would increase the GDP of Germany by 8,088 million USD. (OECD, 2010, p. 23). It is claimed that if Germany raised its PISA score to the level of Finland, the country 'would see a USD16 trillion improvement, or more than five times current GDP. All of these calculations are in real, or inflation-adjusted, terms.' (OECD, 2010, p. 25).

These and other findings based on Hanushek's economic modelling have been strongly rejected by a variety of scholars from different academic fields. In 2017, Komatsu and Rappleye offered a direct challenge in an article entitled 'A new global policy regime founded on invalid statistics? Hanushek, Woessman, PISA, and economic growth' (Komatsu & Rappleye, 2017). Using precisely the same data, they came to a totally different conclusion. Referring to the 'highly influential comparative studies [that] have made strong statistical claims that improvements on global learning assessments such as PISA will lead to higher GDP growth rates', they identified the consequence of the continued utilisation and citation of such claims as 'a growing aura of scientific truth and concrete policy reforms'. For Komatsu and Rappleye 'the new global policy is founded on flawed statistics' and they urged a more rigorous global discussion of education policy'. (Komatsu & Rappleye, 2017, p. 1) It is a discussion to which science educators have an important contribution to make.

PISA and student motivation

Reliable test data assume that respondents take the test seriously and do their best. In contrast to many other tests and exams, PISA is a 'low-stakes' test: it is anonymous, and no data are reported back to the student, the teacher, school or school district. Only national data are reported; PISA is only 'high-stakes' for the national ministries of education. In this test situation, some students may not put all their efforts into answering the questions presented to them. Educators are well aware that 'school culture' and respect for authority differs strongly between countries. One might expect that pupils in some countries are more loyal and willing to do what they are asked to do than pupils in other countries. PISA has two questions that shed light on this issue. In one question students were asked to rank their effort on the PISA test on a scale from 1 to 10. Another question asked students to rank their effort when they sit an examination. The difference between these two rankings can be seen as a measure of how serious the students are when they take the PISA test. The data, as revealed by the Swedish Newspaper Dagens Nyheter (16. June 2014), showed that the Swedish students have the largest difference. Norway and Denmark had similar numbers. Asian PISA-winners had small differences; the students reporting maximum effort on both questions.

PISA: a political and economic project

As a project of the OECD, PISA reflects the desire to promote the economic development that gives the organisation its *raison d'être*. Such promotion might be achieved in several different ways, for example, by investing in a science education designed to foster human development. Equally, it might be achieved by adopting a more instrumental approach to education that emphasises the development of a skilled labour force for a free market economy. In the 1980s, the OECD adopted essentially conservative ideas that prioritised the latter view embodying the economic function of schooling.

The Norwegian economist Kjell Eide was central in the development of the educational involvement of the OECD in period from the early 1960s up to the beginning of 1990s. Reviewing the political debates that took place within the OECD in that decade, Eide concluded in 1995 that if the ambition of the OECD was to assume 'responsibility for arranging international examinations on behalf of governments ... it will make the OECD a strong instrument of power and contribute to a harmonization that will exceed everything we have feared ...' (Eide, 1995: 104, author's translation). Four years later, PISA made clear that it constituted a commitment by all the governments of OECD countries to 'monitor the outcomes of education systems in terms of student achievement, within a common framework that is internationally agreed' (OECD, 1999, p. 11). In 2013, Andreas Schleicher, the Director of PISA, claimed that the project was 'really a story of how international comparisons have globalized the field of education that we usually treat as an affair of domestic policy' (Schleicher, 2013). The following quotation from an OECD report confirms this normative effect of PISA.

PISA has now become an almost global standard, and is now used in over 65 countries and economies [...] PISA has become accepted as a reliable instrument for benchmarking student performance worldwide ... (Breakspear, 2012)

Such a claim presents a significant difficulty. In acknowledging that PISA supplants education as 'an affair of domestic policy', it ignores the great diversity of social, political and economic contexts within which school systems are established and function. Such inherent diversity is overridden by using PISA as a normative instrument of educational policy and governance. In some respects, therefore, the response of legislators to PISA results that are found wanting is not only predictable, but inevitable (Alexander, 2012.)

The claim also does not fit comfortably with other statements about the precise aims of the PISA initiative. In 1999, a year before the first round of testing, PISA asked the following questions.

'How well are young adults prepared to meet the challenges of the future? Are they able to analyse, reason and communicate their ideas effectively? Do they have the capacity to continue learning throughout life? Parents, students, the public and those who run education system need to know (OECD, 1999:11).

These questions have subsequently appeared in many subsequent PISA reports and other documents. However, these stress that the skills and knowledge tested by PISA are not primarily defined in terms of the common denominators of national curricula but in terms of what skills are deemed essential for future life (OECD, 2009: 11). As a result, PISA *does not measure according to national school curricula* but according to an assessment framework made by OECD-appointed PISA experts (OECD, 2016a).

There would seem to be a degree of tension between statements such as these and offering PISA results as valid measures of the quality of national school systems.

The impact of PISA on national curriculum policies

The attention given to PISA results in national media varies from country to country but in most cases, it is substantial and has increased with each round of PISA testing (Breakspear, 2012, 2014). In some countries, the media coverage has been highly dramatic. In Norway, for example, the PISA 2000 and 2003 results provoked headlines such as 'Norway is a school loser' across two pages of a national newspaper (Dagbladet, December 5th, 2001). (Norway was actually above the middle of the OECD countries). For the Conservative Prime Minister of that country, the PISA 2000 outcome was 'like coming home from the Winter Olympics [in which Norway normally excels] without a medal'. Historians and educators have examined in detail how successive Norwegian governments have used the country's PISA results to 'legitimize school reforms'. (Helsvig, 2017; Sjøberg, 2018b). Curiously, some of the curriculum reforms introduced to enhance the PISA results of students in Norway, Denmark and Sweden are at odds with those that characterise the science curriculum in a high-scoring country like neighbouring Finland.

Norway is by no means alone in giving PISA results an unwarranted significance. In the USA the 2018 results headlines claimed 'It isn't just working: PISA test Scores Cast Doubt on U.S. Education Efforts' (New York Times, 3 December 2019). The decline in PISA scores in 37 countries, including those in high performing countries like Finland, Japan and Korea, was blamed on students who were 'Sleepless, distracted and glued to devices: no wonder students' results are in decline' (Sydney Morning Herald, December 5th, 2019). Unsurprisingly, PISA results judged positive prompted headlines like 'Mainland Chinese Students Best in World as Singapore, Hong Kong slip down the rankings' (South China Morning Post, December 3rd, 2019). In the UK, differences in PISA data from different parts of the Kingdom have received particular attention. The 2016 test results in Scotland caused a political row in that country despite the fact that the PISA scores were 'similar to the OECD average' (BBC News, 3 December 2019).

The response of legislators to PISA results and the attendant publicity has been to propose ways in which school curricula can be modified in order to maximise PISA performance.

The PISA results from the first round of PISA testing placed Germany below the middle of the 'league table' of participating countries and they became an important issue in the German election in the following year (Ertl, 2006) They also led to major initiatives to improve the quality of school science and mathematics education. The German National Institute for Science Education, IPN (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik), which had the contract to run PISA in Germany, received substantial funding to improve school science education. By 2014, Steffen and Hössle could conclude that 'Germany finally introduced national standards for science education as one reaction following the results of the PISA studies' (Steffen & Hößle, 2014, p. 343).

Science educators, curriculum developers and policy makers perhaps ought to give greater scrutiny to the relationship that has developed in many countries between PISA as an assessment instrument and its consequences for the school science curriculum.

Conclusion

As a major international comparative study, PISA differs from much earlier work in the field of comparative education. It is quantitative rather than qualitative and is underpinned by *a priori* assumptions about the relationship between science and mathematics test scores and economic development. As noted above, those assumptions and the calculations derived from them are open to challenge.

Moreover, as a quantitative survey, PISA data can take no account of the many different beliefs, assumptions, pedagogical practices, and cultural, social, economic and political contexts within which schooling takes place and which, among much else, influence student performance and attitudes. The fact that PISA tests take no account of these factors means that its globalising influence runs the risk of reducing school curricula to a narrow norm the outcomes of which that can be measured. In addition, if, as PISA asserts, the project seeks to assess how well students' scientific education equips them to respond to the problems they are likely to face in their future lives, any attempt to do so that ignores these variables seems unlikely to constitute a valid basis upon which to compare and rank countries, regions and economies.

Despite such severe limitations, the PISA initiative has raised the profile of science and mathematics education, although in doing so, it may also have had the effect of devaluing the importance of other school subjects and the curriculum as a whole. It has also unquestionably opened up a variety of research perspectives, and, as noted above, a number of issues that deserve investigation. These benefits of PISA are not inconsiderable but they need to be set alongside the difficulties in measuring what the testing program claims to measure. PISA scores and rankings are not facts, nor are they objective or neutral outcomes of the project. There is therefore an important task facing the science education community, namely to give the PISA project the rigorous scholarly examination community it deserves.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Svein Sjøberg is Emeritus Professor of Science Education at Department of Teacher Education and school Research at the University of Oslo, Norway. He has worked with children's conceptual development, with gender and science education and education in developing countries. His current research interests are the political, social, ethical and cultural aspects of science education, in particular the impacts and influence of large scale assessment studies like PISA and TIMSS.

Edgar Jenkins is Emeritus Professor of Science Education Policy at the University of Leeds, UK, where he was Head of the School of Education and Director of the Centre for Studies in science and Mathematics Education. His most recent book *Science for All: The struggle to establish school science in England* was published in 2019.

ORCID

Svein Sjøberg  <http://orcid.org/0000-0001-9638-0498>

References

- Alexander, R. (2012). Moral panic, miracle cures and educational policy: What can we really learn from international comparison? *Scottish Educational Review*, 44(1), 4–21. <http://eprints.whiterose.ac.uk/76276/>.
- 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*, No. 71, OECD Publishing. <https://doi.org/10.1787/5k9fdqffr28-en>
- Breakspear, S. (2014). *How does PISA shape education policy making? Why how we measure learning determines what counts in education*. Centre for Strategic Education. <http://simonbreakspear.com/wp-content/uploads/2015/09/Breakspear-PISA-Paper.pdf>
- Bybee, R., & McCrae, B. J. (2011). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education*, 33(1), 7–26. <https://doi.org/10.1080/09500693.2010.518644>
- Eide, K. (1995). *OECD og norsk utdanningspolitikk. En studie av internasjonalt samspill, (OECD and Norwegian education policy. A study of international interaction)*. NAVFs Utredningsinstitutt.
- Engel, L. C., & Rutkowski, D. (2018). Pay to play: What does PISA participation cost in the US? *Discourse: Studies in the Cultural Politics of Education*. 484–496. <https://doi.org/10.1080/01596306.2018.1503591>
- Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, 32(5), 619–634. <https://doi.org/10.1080/03054980600976320>
- EU. (2007). *Science Education Now: A renewed pedagogy for the future of Europe, (The Rocard report)*.
- Helsvig, K. (2017). *Reform og rutine. Kunnskapsdepartementet historie (Reform and routine. The history of the Ministry of Education) (1945–2017)*. Pax.
- ICSU. (2011). *Report of the ICSU Ad-hoc review panel on science education*. International Council for Science. https://www.mathunion.org/fileadmin/ICMI/files/Other_activities/Reports/Report_on_Science_Education_final_pdf.pdf
- IEA (2018). *Sixty years of IEA – (1958–2018)*. IEA, Amsterdam. <https://indd.adobe.com/view/da338b4a-5e60-492e-b325-2b8c8f88cf42>
- Jensen, F., Pettersen, A., Frønes, T. S., Kjærnsli, M., Rohatgi, A., Eriksen, A., & Narvhus, E. K. (2019). *PISA 2018. Norske elevers kompetanse i lesing, matematikk og naturfag. (Norwegian Report PISA 2018)*. Universitetsforlaget.
- Kjærnsli, M., & Lie, S. (2011). Students' preference for science careers: International comparisons based on PISA 2006. *International Journal of Science Education*, 33(1), 121–144. <https://doi.org/10.1080/09500693.2010.518642>
- Kjærnsli, M., Lie, S., Olsen, R. V. & Roe, A. (2007). *Tid for tunge løft. Norske elevers kompetanse i naturfag, lesing og matematikk i PISA 2006*. Oslo: Universitetsforlaget
- Komatsu, H., & Rappleye, J. (2017). 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>
- Kreiner, S., & Christensen, K. B. (2014, April). 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>
- OECD. (1999). *Measuring student knowledge and skills. A new framework for assessment*.
- OECD. (2005). *PISA 2003 technical report*.
- OECD. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*.
- OECD. (2009). *PISA 2006 technical report*.

- OECD. (2010). (*Hanushek and Woessman*) *The high cost of low educational performance: The long run economic impact of improving PISA outcomes*. Retrieved August 23, 2020, from <https://www.oecd.org/pisa/44417824.pdf>
- OECD. (2015). *Students, computers and learning: Making the connection*. Retrieved 23, 2020, from <https://doi.org/10.1787/9789264239555-en>
- OECD. (2016a). *PISA 2015 assessment and analytical framework: Science, reading, mathematics and financial literacy*.
- OECD. (2016b). *PISA 2015 results (Volume I): Excellence and equity in education*.
- OECD. (2016c). *PISA 2015 results (Volume II): Policies and practices for successful schools*.
- OECD. (2019a). *PISA 2018 assessment and analytical framework*.
- OECD. (2019b). *PISA 2018 results. What students know and can do (Vol. I)*.
- OECD. (2019c). *PISA 2018 results. Where all students can succeed (Vol. II)*.
- OECD. (2019d). *PISA 2018 results. What school life means for students' lives (Vol. III)*.
- Oliver, M., McConney, A., & Woods-McConney, A. (2019, December). The efficacy of inquiry-based instruction in science: a comparative analysis of six countries using PISA 2015. *Research in Science Education*. <https://doi.org/10.1007/s11165-019-09901-0>
- Schleicher, A. (2013). *Use data to build better schools*. TEDGlobal, video. http://www.ted.com/talks/andreas_schleicher_use_data_to_build_better_schools?language=en
- Schleicher, A. (2015, June 17). Vietnam's 'stunning' rise in school standards. *BBC News*. <http://www.bbc.com/news/business-33047924>
- Sellar, S., Thompson, G., & Rutkowski, D. (2017). *The global education race: Taking the measure of PISA and international testing*. Brush Education Inc.
- Sjøberg, S. (2018a). The power and paradoxes of PISA: Should we sacrifice inquiry-based science education (IBSE) to climb on the rankings? *NorDiNa*, 14(2), 186–202. <https://doi.org/10.5617/nordina.6185>
- Sjøberg, S. (2018b). PISA – Oraklet i Paris? Global styring af skole og uddannelse. (PISA – The oracle in Paris? Global governance of schooling and education). In D. Sommer & J. Klitmøller (Eds.), *Fremtidsparat - Hindsides PISA: Nordiske perspektiver på uddannelse*. Hans Reitzels Forlag. 73-105.
- Steffen, B., & Hößle, C. (2014). Decision-making Competence in Biology Education: Implementation into German Curricula in Relation to International Approaches. *Eurasia Journal of Mathematics, Science & Technology Education*, 10(4), 343–355. <https://doi.org/10.12973/eurasia.2014.1089a>.
- Wuttke, J. (2007). Uncertainty and bias in PISA. In S. Hopman, T. G. Brinek, & M. Retzl (Eds.), *PISA according to PISA — Does PISA keep what it promises?* (pp. 241–263). Lit Verlag.
- Zhao, Y. (2017). What works may hurt: Side effects in education. *Journal of Educational Change*, 18(1), 1–19. <https://doi.org/10.1007/s10833-016-9294-4>

Appendix

The basic features of PISA and TIMMS

Below are similarities and differences between PISA and TIMSS in simplified form.

- TIMSS was initiated and is (to a certain degree) governed by academics and researchers, while PISA was established by the OECD and is governed by representatives for governments in OECD member states.
- TIMSS is basically descriptive and analytical, while PISA is explicitly and intentionally normative.
- Both studies tests are survey studies, testing a representable sample from their target population. Typical sample sizes are 5–7000 students.
- TIMSS tests students in a particular school grade (4. and 8.), while PISA tests students at a particular age (15).
- TIMSS selects whole classes (and their teachers), while PISA samples individual students from selected schools.
- TIMSS tests every 4th year, PISA every 3rd year.

- TIMSS is 'curriculum based'. The test is meant to be close to the school science and mathematics curriculum, while the PISA testing is based on an assessment framework that is made by appointed experts.
- TIMSS items are typical 'school exam' questions in science and mathematics, while PISA items usually have a substantial amount of text, and are meant to address authentic, real life challenges.
- Testing time is about two hours for both studies. In addition, both studies have student background questionnaires of about half an hour. Additional data are also collected from school principals and teachers.
- The total testing time for both studies is about 10 hours, but each student answer only a selection of the items. This enables a broader sampling of contents to be covered by the tests.
- In recent rounds of TIMSS and PISA the testing is done on a computer.
- TIMSS has two subjects, while PISA has three core domains: science, mathematics and reading plus an optional domain: 'financial literacy'.
- TIMSS has equal testing time on science and mathematics, while PISA has one of its three subjects in focus in each round. Only the main subject provides reliable data. Science was the focus in PISA 2006 and PISA 2015.
- The research design allows TIMSS and PISA to track trends over time. Data for trends are made possible by maintaining some items from one test round to the next.
- TIMSS and PISA calculate and publish data that are statistically normalised, with a mean population score of 500 and a standard deviation of 100. These parameters are calculated based on the results at one particular year, in order to be seen as an 'absolute' scale.
- TIMSS and PISA are anonymous and 'low-stakes' tests for the student, their teacher and their school. Only population results are reported.