Do PISA data justify PISA-based education policy?

Luisa Araujo
Department of Human Capital and Employment, European Commission Joint Research Centre Ispra Sector, Ispra, Italy
Andrea Saltelli
University of Bergen, Bergen, Norway and Universitat Autonoma de Barcelona, Barcelona, Spain, and
Sylke Schnepf
Competence Centre on Microeconomic Evaluation, European Commission Joint Research Centre Ispra Sector, Ispra, Italy

Received 14 December 2016
Revised 17 February 2017
Accepted 24 February 2017

Abstract

Purpose – Since the publication of its first results in 2000, the Programme for International Student Assessment (PISA) implemented by the OECD has repeatedly been the subject of heated debate. In late 2014 controversy flared up anew, with the most severe critics going so far as to call for a halt to the programme. The purpose of this paper is to discuss the methodological design of PISA and the ideological basis of scientific and policy arguments invoked for and against it.

Design/methodology/approach – The authors examine the soundness of the survey methodology and identify the conflicting interpretations and values fuelling the debate.

Findings – The authors find that there are legitimate concerns about what PISA measures, and how. The authors conclude that the OECD should be more transparent in the documentation of the methodological choices that underlie the creation of the data and more explicit about the impact of these choices on the results. More broadly, the authors advise caution in the attempt to derive and apply evidence-based policy in the domain of education; the authors furthermore propose an alternative model of social inquiry that is sensitive and robust to the concerns of the various actors and stakeholders that may be involved in a given policy domain.

Originality/value – The issues and tensions surrounding the PISA survey can be better understood in the framework of post-normal science (PNS), the application of which to the PISA controversy offers a potential solution to a stalemate.

Keywords Comparative education, Evidence-based policy, Educational measurement, Programme for international student assessment (PISA)

Paper type Conceptual paper

Introduction

On 6 May 2014 the British daily newspaper *The Guardian* published an open letter (Meyer and Zahedi, 2014) addressed to Andreas Schleicher, Director for Education and Skills of the Organisation for Economic Cooperation and Development (OECD), expressing grave concerns about the deficiencies and pernicious effects of the international survey of students' skills known as "PISA" which is implemented by the OECD every three years in a large and growing number of countries (more than 60 in the 2012 round). Jointly signed by approximately 80 academics, public school district administrators, parents and teachers, the letter suggested skipping the 2015 round of PISA in order to take time to discuss and address the issues raised, at local, national and international levels, and ultimately to improve the assessment model.

This was not the first time that the Programme for International Student Assessment (PISA) had featured in the mainstream media. Several magazines and newspapers have

The views expressed are purely those of the writers and may not under any circumstances be regarded as stating an official position of the European Commission.

International Journal of Comparative Education and Development
Vol. 19 No. 1, 2017
pp. 1-17

2396-7404
DOI 10.1108/IJCED-12-2016-0023

commented on PISA since its launch in 2000 – titles such as *The Economist*, *The Guardian* and *Atlantic Monthly* typically run an article on PISA each time a new wave of data is released . No other international survey of students' skills attracts a comparable level of attention. The trends in mathematics and science study (TIMSS), for instance, first conducted in 1995 and implemented every four years, has been providing international data on student achievement in these subjects at fourth and eighth grade (corresponding to students of roughly 10 and 14 years, respectively) for two decades now. However, TIMSS aims to measure knowledge of mathematics and science content studied in school and is thus curriculum-based. By contrast, PISA is purportedly designed to measure students' ability to use or apply the knowledge acquired in school to solve problems encountered in everyday life. As described in the most recent PISA assessment framework, "PISA is a collaborative effort undertaken by its participants – the OECD member countries as well as over 30 non-member partner countries and economies – to measure how well students, at age 15, are prepared to meet the challenges they may encounter in future life. Age 15 is chosen because at this age students are approaching the end of compulsory education in most OECD countries" (OECD, 2013a, b, p. 13). Moreover, the OECD's ambition with PISA is to supply internationally comparable data.

Comparative international assessments can extend and enrich the national picture by providing a larger context within which to interpret national performance. They can show what is possible in education, in terms of the quality of educational outcomes as well as in terms of equity in the distribution of learning opportunities. They can support policy targets by establishing measurable goals achieved by other systems and help to build trajectories for reform. They can also help countries to work out their relative strengths and weaknesses and monitor progress (OECD, 2013a, b, p. 13).

PISA is thus not intended to measure command of any curriculum, but rather to measure so-called "life skills", to gather evidence that can inform education policies and to monitor education system performance. These more ambitious aims make PISA more vulnerable to criticism than surveys such as TIMSS, which also provides international comparative data on educational outcomes and reports country league tables of student achievement. Comparative education analysis since the advent of PISA is no longer about comparing curriculum goals and content across countries and mapping how differing emphases might be related to achievement outcomes: the analysis has shifted from studying what students have been taught and how much they have learned, to studying what students can do with what they have been taught (Hutchison and Schagen, 2007). Discussion of PISA has furthermore largely revolved around its use in policy-making – the criticism has been made that PISA may have contributed to the idealization of particular countries and their educational cultures (Waldow *et al.*, 2014). PISA league tables have presented top performers Shanghai, South Korea and especially Finland as the models to follow in education reform (Takayama *et al.*, 2013). (Although Finland saw its performance slip in 2012, it was a star performer in PISA for over a decade.) This kind of comparison, some argue, is based on a conception of excellence in education that is blind to contextual differences. For example, it has been argued that the culture of hard work, effort and persistence, together with an education-ambitious parenting style, can, to a large extent, explain the strong performance of Asian countries such as South Korea (Smithers, 2013). In the case of Finland, explanatory arguments have focused on the idea that the factors underlying high achievement are primarily of a pedagogical nature: Finland apparently has excellent teachers and high-quality teacher education (Simola, 2005). However, PISA data were up to now based principally on the relationship between achievement scores and a limited number of school characteristics reported by school principals. Since the 2015 PISA wave, which became available in December 2016, organizers made a teacher questionnaire optional. Although not all countries collected this information, the teachers' questionnaire

opens the possibility to investigate a greater variety of factors associated with students' achievement. Similarly, a parental questionnaire has been optional since the 2006 wave. This questionnaire covers important information on parental background, educational possessions as well as parental involvement in children's learning. The possibility to include these parental variables into the research design lessens the over-reliance on system and pedagogical factors for explaining educational outcomes. Unfortunately, only a small number of countries have taken up the option of administering the parental questionnaire: in 2009 only 14 countries and in 2012 only 11 countries chose to collect data from parents, out of 65 countries covered in each PISA round.

The pragmatic or utilitarian approach inherent to PISA (that is, focussing on what students can do with what they have learned) goes hand-in-hand with the view that good educational outcomes imply commensurate economic returns to individuals and societies (Hanushek and Woessmann, 2012). PISA data have in effect been repeatedly used to make the "economic case" for education. For example, in a recent policy document prepared for the European Commission, Woessmann (2014) states that "(i)f every EU Member State achieved an improvement of 25 points in its PISA score (which is what for example Germany and Poland achieved over the last decade), the GDP of the whole EU would increase by between 4% and 6% by 2090; such an 6% increase would correspond to 35 trillion Euro" (p. 10).

Appreciation of PISA by economists of education has, however, been countered by repeated criticism from other quarters regarding the cross-sectional nature of the survey and the lack of information about the assumptions and limitations associated with the statistical modelling choices made in the survey design (Goldstein, 2004). A further source of uncertainty is the difference between correlation and causation (in the sense that, for example, improvement in PISA scores will not necessarily lead to GDP growth).

Given the relevance of the PISA results for policy formulation, the purpose of this paper is to focus on and disentangle the various strands of criticism of PISA. We see the controversy over PISA as having two main dimensions: methodological and ideological. The former concerns the statistical models and techniques adopted in the PISA survey design and the second concerns worldviews and conceptions of what education is for. In what follows we map these two areas of disagreement, furthermore situating the discussion over PISA in the current context of increased debate over the merits of evidence-based policy (Strassheim and Kettunen, 2005; Boden and Epstein, 2006; Saltelli and Giampietro, 2016).

Q1

What does PISA measure – and how?

The quality of any survey depends on several factors: the salience and relevance of the underlying reality it aims to capture; the quality of the measurement apparatus designed to capture it; and its statistical implementation, including its capacity to generate a representative sample for the target population (Groves *et al.*, 2009). We touch on each of these aspects in turn.

What is a "life skill score"?

A first criticism of the existing PISA data relates to its comparability across countries as a means of assessing "how well young adults, at age 15 and therefore approaching the end of compulsory schooling, are prepared to meet the challenges of today's knowledge societies" (OECD, 2004, p. 12). The OECD assumes the life skills needed to function in knowledge societies to be the same for all countries – a rather dubious assumption, given that the countries participating in PISA differ greatly in terms of their cultures and level of economic development, which raises legitimate questions about how appropriate it is to rank them in a single table. The skills needed by a young adult are likely to depend on the characteristics of the society in which the person is living; hence, what it means to "function" in a knowledge society will vary from country to country. While comparability of data across countries is desirable, a question

raised by critics (and discussed below) is whether comparability in PISA is only achieved by ignoring the great diversity of curricula across the participating countries – diversity which might in fact be a source of country-specific creativity and well-being.

Does PISA meet the requirements of validity?

Assuming that the construct of "the challenges of today's knowledge societies" is a reasonable one to explore, irrespective of cultural or developmental specificities, questions remain about how well the selection of items and the item response model used to summarize item answers into one overall score fit the abstract skill that PISA aims to measure. More specifically, the main criticism in the literature concerns the multidimensionality of the items being measured. Meyerhoefer (2007) discusses how the items used do not only measure, for example, "maths" ability but also a student's ability to comply with the test structure. According to Goldstein (2004), education skills are multidimensional; however, in PISA, items that demonstrate "poor psychometric characteristics in more than ten countries ('dodgy' items)" can be deleted (OECD, 2012, p. 148). Such items are those most prone to multidimensionality and to reflect cultural bias (Goldstein, 2004). Their removal from the set therefore obscures differences between countries that might otherwise demonstrate greater heterogeneity on varying educational dimensions. For Goldstein, the aim of a cross-national study should not be to neglect, but rather to obtain, information on underlying differences between countries.

It is common for surveys to measure only one aspect of a specific construct; however, the specific aspect needs to be defined. It is less than clear in PISA which dimensions are being measured and which suppressed.

Furthermore, validity is very difficult to achieve in a cross-national survey. The first challenge is to create items that are culturally neutral; the second is to translate these items into other languages. In order to achieve the desired neutrality, the OECD uses a variety of mechanisms to make sure that wording and translation do not impact on the results. Moreover, the OECD generally runs trials before implementing the final PISA questionnaire. If items perform badly during these trials, they are withdrawn. More openness and transparency on the part of the OECD about the results of trials and the consequent choice of items would help potential users of the results to judge their reliability.

Another object of criticism is the impact of the choice of item response models on league table rankings and variation within countries. Item response models use various assumptions in order to summarize an individual's answers to a battery of questions into a single score for that individual. For Brown *et al.* (2007) this approach is problematic, since the choice of the item response model has a considerable impact on the rank position of countries in terms of educational inequity (i.e. the ranking of countries in terms of the size of the gap between the 95th and 5th percentile of student scores in each country – a measure of how equitable a country's education system is). This question is often raised in relation to countries' overall achievement (e.g. Mont, 2011; Breen *et al.*, 2009). A related problem is that not all items are made available to the research community. The OECD justifies this with the need to conceal items from future cohorts of test-takers. However, this reticence makes the reproducibility of country rankings from the raw data impossible. Users of the PISA survey therefore cannot investigate whether published results are dependent on the choices made by the survey designers. According to Micklewright and Schnepf (2007), this implies that the PISA organizers should themselves provide sensitivity analyses of their choice of items, item response models and other assumptions in a clearly accessible way to the rest of the world. This information could then be sensitivity-audited by the community of interested scholars (Saltelli *et al.*, 2013; Saltelli and Funtowicz, 2014).

The meaning of the ratio scale used to measure ability must also be considered. Is a child who achieves a PISA score of 250 only half as well equipped for life as a child scoring 500?

Atkinson (1975) notes that "there is at present really no such thing as the distribution of ability: the distribution depends on the measuring rod used and cannot be defined independently of it. […] the fact that most IQ tests lead to a distribution of scores which follows the normal distribution does not necessarily tell us anything about the distribution of abilities: it may simply reflect the way in which the tests have been constructed" (p. 89). This caveat should preface every PISA report.

There is furthermore general agreement in academia that the main limitation of PISA is its reliance on cross-sectional data, i.e. data that refer in each round to different student cohorts. Scholars (e.g. Goldstein, 2004) have repeatedly argued that PISA should not be used for the purpose of drawing specific policy implications for improving education because the various rounds do not follow the same students over the course of their school careers. As previously mentioned, claims for causality based on cross-sectional data are problematic; nonetheless, the predominant PISA-based policy narrative, which makes the economic case for education reforms, is based on an assumption of causality running from education to economic growth.

In sum, many experts consider that the cross-sectional design of PISA limits its usefulness for policy. Furthermore, PISA results should be accompanied by detailed documentation of the choices made in terms of the dimensions measured, the items selected as well as those discarded, the item response models chosen and the resulting impact of all these choices on the ranking of countries.

Are PISA samples representative?

As discussed above, in order to make reliable cross-country comparisons, a common measuring rod is required. An equally relevant issue is whether the sample of students used to estimate the country average is indeed representative of the target population. PISA allows for the exclusion of students with special educational needs and newly arrived immigrants. This has raised concern (Wuttke, 2007), because some countries exclude more students than the 5 per cent threshold imposed by the PISA designers.

Also in question is the non-response bias of PISA scores. In order to limit non-response bias, PISA sets a threshold of 85 per cent for school response and 80 per cent for student response. Such thresholds are, however, no guarantee that non-response bias will be negligible, since, besides the non-response rate, the pattern of response is also significant. Low response may not increase bias if respondents and non-respondents are similar. On the other hand, high response can still yield non-response bias if respondents and non-respondents differ significantly. However, the PISA design prescribes that only those countries which do not meet the response threshold are required to examine the non-response pattern. Therefore, the PISA reports provide information on the level of school and student response by country, but not on cross-national differences in response patterns. However, Micklewright *et al.* (2012) show that students with lower ability were less likely to take the PISA test in England in 2000 and 2003. In both years the overall achievement score for England was therefore upwardly biased. The bias (the difference between the "true" and the "estimated" PISA score) was about two to three times greater than the published standard error. If the non-response bias is taken into account, England's position shifts downward by three places for maths and two for science (but none for reading) in the 2003 league tables. Non-response bias was clearly appreciably high in England for these years. In later PISA rounds, the response rate increased in England, which would suggest that the data became more representative. Indeed, over time, the trend in scores for England seems to show a decline in achievement relative to other countries. However, the perceived decline is mainly due to better coverage of poorly performing students in later rounds and to other methodological changes (Jerrim, 2013). This misunderstood drop in scores nevertheless aroused considerable public concern.

In sum, PISA currently uses simplified rules in order to ensure that data are representative in each country. This implies that possible non-response bias is considered acceptable and that, as in the example of England above, the non-response bias may considerably exceed the standard error estimate. In turn, this will affect a country's league table ranking and its evolution over time. This implies that any intention to use PISA data to inform policy should be mitigated by a careful study of the limitations of the data.

Putting it all together: total error of the PISA score

Each limitation discussed above contributes to what is called "total survey error", a measure of the precision of a survey estimate that takes all possible errors into account. Generally, PISA survey organizers communicate only one component of the total survey error with the tables of results: the standard error. Given the considerable impact of the PISA results on education policy, the OECD should document better choices made by the PISA designers concerning measurement and representativeness. This would increase the legitimacy of the information provided to the public.

The PISA worldview

The remarkable success of PISA has led to the dominance of a certain mode of governance in education. This is discussed in a recent article by Sellar and Lingard (2014) entitled "The OECD and the expansion of PISA: new global modes of governance in education". They note that:

With Eccleston (2011, p. 248), we recognize that the authority of an international organization such as the OECD has both "rational – legal and moral dimensions" and agree with the observation that an "international organization's political authority is at its zenith when its rational/technical agenda aligns with prevailing social values and sentiments".

The technical agenda implemented through PISA – in what it measures, and how – is entirely congruent with OECD ideals and the vision of the world it supports. Sellar and Lingard (2014) go on to note that "[…] the OECD uses a 'performative' semiotic construction of the concept of globalization, implying only a neoliberal reading, and denying other accounts in the process" (p. 922). PISA-based league tables are controversial, not chiefly because of their methodological limitations, but rather because of their close-to-hegemonic status and ideological implications, as Sellar and Lingard contend. This is clear in the letter published in the *The Guardian* in 2014, which raised a long list of normative and ethical issues, which are summarized here:

(1) the notion of knowledge as "Bildung" or emancipation is defeated by the economic logic of PISA; the goal of education becomes the preparing of young men and women for gainful employment;

(2) with PISA, the OECD has acquired enormous political power which serves the organization's neoliberal framings; this power is usurped from more democratic arrangements and organizations;

(3) to implement PISA and provide a host of follow-up services, the OECD has embraced "public-private partnerships" and entered into alliances with multinational for-profit companies, which stand to gain financially from any weaknesses in education systems – real or perceived – identified by PISA;

(4) relevant school subjects and areas of scholarship have been excluded by PISA;

(5) PISA harms or increases stress on school pupils and their teachers; and

(6) PISA fosters short-termism in education policy.

One of the signatories of the letter, Heinz-Dieter Meyer, commented further that:

> The OECD is a club. There is not even an aspiration to democratic legitimation. You become a member only by being invited. The founders were the US and UK and they are still the chief influences. Just as governments in those countries were taken over by neoliberal Thatcherism and Reaganism, so was the OECD. Except that, while there is resistance to these policies in member countries, in the OECD they reign supreme […]. It looks more and more like a global super-ministry of education […]. It owns a very big chunk of the world's educational policy research. It has greater influence on the generation and analysis of educational data than any other single institution (Cited in Wilby, 2014).

In his response to the letter in *The Guardian*, Andreas Schleicher (2014) rejected the idea that there was any bias in the selection of company partnerships collaborating with the OECD in the running of the PISA assessments, as well as the idea that PISA had caused a shift to short-term fixes in education policies. Schleicher asserted that "PISA has provided many opportunities for more strategic policy design".

The criticism of PISA as the vehicle of a prevailing neoliberal economic orthodoxy is not new: Meyer was expressing the view shared by many scholars that the PISA designers should be more open to alternative discourses on education and to the interests and concerns of stakeholders. The multi-authored book edited by Hopmann *et al.* (2009) PISA according to PISA – Does PISA keeps what it promises? collects a number of critical contributions that, to a large extent, prefigure the ideas expressed in the letter in *The Guardian*. Criticism expressed in the book targets, for example, the OECD's biased selection of company partnerships; the fact that PISA does not cover the breadth of education or schooling; and the neo-liberal brand of education it promotes. These criticisms reflect a worldview that sees PISA as a profit-making enterprise that misconceives the purpose of education.

Moreover, the signatories of the letter reject the idea that the accountability that PISA supposedly delivers is fit for purpose. In other words, holding school systems accountable for students' test results does not serve to improve education. Diane Ravitch (2010), one of the letter's signatories and author of *The Death and Life of the Great American School System: How Testing and Choice are Undermining Education*, is adamantly opposed to the concept of "added value" and to any measure of achievement, arguing that it is impossible to evaluate a school's effectiveness using student achievement as the outcome measure. Specifically, Ravitch argues that testing has many negative side effects: principally, diminished autonomy and local control by schools over the curriculum and choice of textbooks. This diminished control restricts what is taught in schools, relegates certain subject matters to the side-lines and pushes reading, mathematics and science to centre stage.

Regarding the potential harm done to students and teachers, some critics consider that PISA-oriented pedagogy entails more testing time for students and more scripted lessons for teachers, both of which are held to increase stress levels. If "teaching to the test" is considered harmful, then PISA is clearly leading in the wrong direction.

Meyer's critique of PISA, however, raises the more fundamental question of whether testing in the classroom can accurately capture changes in system-level education policies: "PISA […] has caused a shift of attention to short-term fixes designed to help a country quickly climb the rankings, despite research showing that enduring changes in education practice take decades, not a few years to come to fruition". In one of the several exchanges following the letter published in *The Guardian*, the OECD reacted to this criticism as follows.

There is nothing that suggests that PISA, or other educational comparisons, have caused a "shift to short-term fixes" in education policy. On the contrary, by opening up perspectives to a wider range of policy options that arise from international comparisons, PISA has provided many opportunities for more strategic policy design. It has also created important opportunities for policy-makers and other stakeholders to collaborate across borders. The annual International Summit of the Teaching Profession, where ministers meet with

union leaders to discuss ways to raise the status of the teaching profession, is an example. Not least, while it is undoubtedly true that some reforms take time to bear fruit, a number of countries have in fact shown that rapid progress can be made in the short term, e.g. Poland, Germany and others making observable steady progress every three years (OECD, 2014).

The OECD rests its claim on statistical evidence from PISA (OECD, 2010); as well as on research on student's performance in PISA and on the effect of system- and institutional-level factors on performance. Much of this research evidence comes from the field of the economics of education: "Student performance is higher with external exams and budget formulation, but also with school autonomy in textbook choice, hiring teachers and within-school budget allocations" (Fuchs and Woessmann, 2007, p. 1). Note that Fuchs and Woessmann use PISA data to show that school autonomy is related to higher scores, while Ravitch maintains that high-stakes tests diminish school autonomy and do nothing to improve students' scores.

Putting it all together: the PISA worldview

Comparative education analysis as pursued and promoted by the OECD suggests that it is possible to transfer education policy recipes from one education system to another. Yet some top performers, such as Finland, do not have external exams, and apparent improvements or deteriorations in performance in PISA tests have been questioned on the basis of sampling procedures. For example, the steady improvement in PISA scores in Germany could be partly related to a shift in the first- and second-generation immigrant population taking the survey, with more Russian and Eastern European immigrants participating in 2006 and 2009 than in previous rounds (Carnoy and Rothstein, 2013). As discussed in the previous section, if PISA data were longitudinal, these kinds of questions could be investigated.

This illustrates the difficulties confronting policy makers in using the results of international tests such as PISA to design or justify educational reforms. While academics analyse and question the evidence from PISA, politicians tend to use the published rankings: "These are the results that politicians look for first, and that they sometimes use, appropriately or otherwise, to initiate debate about system reforms and on occasion to justify reforms" (Baird *et al.*, 2011, p. 11).

In short, the problem seems to be that the effectiveness of education systems is judged by policy analysts on the basis of country rankings in tests like PISA, such that when a country's position in a league table is low, the need for system reform seems warranted. However, even when evidence from PISA is used appropriately, and sound statistical inferences are made, the ranking of countries on the basis of cognitive tests such as PISA reflects only one possible approach. As Kautz *et al.* (2014) point out in an OECD working paper, "Although non-cognitive skills are overlooked in most contemporary policy discussions and in economic models of choice behaviour, personality psychologists have studied these skills for the past century" (p. 14). The importance of non-cognitive skills such as perseverance, self-esteem and self-confidence and cooperation in predicting future life and labour market outcomes is necessarily neglected when cognitive skills alone are assessed. Heckman's studies (e.g. Heckman and Kautz, 2012; Heckman *et al.,* 2014) of the life outcomes of young people that received the certificate of high school equivalence upon passing the General Educational Development tests illustrate that, although their achievement is equivalent to that of regular high-school graduates, their rates of college completion and success in finding and holding down a job are much lower. This strongly suggests that skills not captured in cognitive tests must be taken into account in explaining life outcomes. In fact, as Heckman *et al.* (2014) state in the same OECD working paper, "This evidence should cause policymakers to think twice about relying on achievement tests to evaluate the effectiveness of educational systems" (p. 14).

Finally, the critique made in the original letter published in *The Guardian* that quantification provides the OECD with unwarranted leverage to influence policy in participating countries echoes a point made by Theodor Porter's (1995) book *Trust in Numbers*, "The appeal of numbers is especially compelling to bureaucratic officials who lack the mandate of a popular election, or divine right […]. Quantification is a way of making decisions without seeming to decide. Objectivity lends authority to officials who have very little of their own" (p. 8). In this sense, the migration of authority from national to supra-national level deserves critical reflection.

The PISA controversy

The PISA controversy, as we have seen, revolves around what PISA measures, how it measures it and how it is used for policy. From a methodological standpoint the lack of documentation on the impact of choices made to arrive at the final achievement score limits practitioners' trust in the data and detracts from the legitimacy of the related policy inferences. The OECD's plans for future PISA assessments include the measurement of some non-cognitive dimensions and the aim to enhance its exploratory power so as to provide policy-makers with better information (OECD, 2015). Nonetheless, their plans do not include the disclosure of the statistical procedures used to produce this policy-relevant knowledge.

For the pragmatic econometrician or applied statistician, standardized tests like PISA help to simplify a complex world, providing information that allows for comparison within and among countries. As Hopmann *et al.* (2009) note, PISA's admirers "praise PISA for giving us the best data base ever available for comparative research, for developing new tools of research, and for PISA's creative analysis of its data sets" (p. 10). The great advantage of PISA lies in its exploration of the many student background and school factors that affect a student's standardized test scores. Before information on student achievement was available, researchers mainly related the economic development of a country to the average number of years of schooling of its population. PISA raised awareness of the many factors that affect students' outcomes besides the number of hours they sit in a classroom. This was an undeniable advance for comparative education research.

However, the disputes over PISA derive largely from the deployment of the "evidence" it produces to inform and influence policy. At present, the use of scientific evidence to advocate in favour of virtually any policy may provoke a controversy. The very idea of "evidence-based policy" has been discredited, as scientific "evidence" itself is often hotly contested by experts (Strassheim and Kettunen, 2005; Boden and Epstein, 2006; Saltelli and Giampietro, 2016) – for example, in relation to the effect of pesticides on bees, the necessity of culling badgers, the advantages of genetically modified organisms (GMO), the outcomes of children raised by gay couples, the economic benefits of migration, and more (Saltelli and Funtowicz, 2014). In such disputes the parties often disagree on the very nature of the problem, not to mention its solution or what the scientific findings point to. For example, GMO may be presented as an issue of food safety by one party but as a question of governance and control over the food industry by another (Saltelli and Funtowicz, 2014).

In relation to PISA, the case of Finland stands out. Research results do not support the idea that systems with external national exams perform better, but rather that teacher quality is the main explanatory factor of high educational outcome (Takayama *et al.*, 2013). In different country contexts, however, the nature of the problem may be perceived differently.

Another example regards the very controversial debate about the association of school size with educational outcomes. Teachers and school administrators may favour reducing class size to better address the needs of students given their own experience within the field, while PISA findings point in the opposite direction: smaller class sizes are not related to higher achievement (Ehrenberg *et al.*, 2001; OECD, 2013b).

Like these issues, the dispute over PISA can be better understood through the lens of new epistemologies such as "post-normal science" (PNS) (Funtowicz and Ravetz, 1993) and the "co-production of knowledge" (Jasanoff, 1996). These epistemologies are pertinent "in the context of increased attention to issues of participation, legitimacy, transparency and accountability", in domains in which the interaction between experts, policy-makers and citizens needs to be rearticulated in more collaborative modes (Carrozza, 2014). This is also referred to as the "democratization of expertise". As we have shown, the scientific evidence from PISA presented in support of policy is disputed by critics from various sectors of society. In upholding their peculiar knowledge claims, all sides in this dispute may be guilty of inappropriate generalizations, hidden value judgements and misrepresentation of the other parties' arguments.

According to the tenets of PNS, these tensions should not be resolved by discarding science, but by investing more in the analysis of the quality of the process on which the evidence has been constructed. This can be achieved by a process of "extended participation", in which the investigation and analysis are open not only to experts from different disciplines and forms of scholarship (one of the demands of critics of PISA), but also to the active participation of relevant and legitimate stakeholders. PNS offers tools such as sensitivity auditing (Saltelli and Funtowicz, 2014) and "NUSAP", which could be used to gauge the reliability of PISA-based inferences.

For example, in reporting PISA results with NUSAP, N would be a country's (adimensional) score, S the standard error reported by the OECD, A would contain material from the second section of this paper and P would be a discussion of the technical reliability, work history and ideological positions behind the data.

Applying sensitivity auditing (Saltelli and Funtowicz, 2014) to PISA would result in a process that unfolded as follows:

Rule 1: "Check against rhetorical use of mathematical modelling"; are PISA results being over-interpreted?

Rule 2: "Adopt an 'assumption hunting' attitude"; this would focus on unearthing possibly implicit assumptions – as the second section of this paper illustrates.

Rule 3: "Detect pseudo-science"; this asks whether uncertainty has been downplayed; as discussed in this paper, there is reason to suppose that this is the case with PISA.

Rule 4: "Find sensitive assumptions before these find you"; this is a reminder that before publishing results the analysis of sensitivity should be done and made accessible to researchers. PISA fails this step.

Rule 5: "Aim for transparency"; see discussion in the second section. The fact that analysis of the sensitivity of the achievement scores to the choice of the models and items used is hampered by lack of data availability is a major limitation of PISA.

Rule 6: "Do the right sums"; the analysis should not solve the wrong problem – doing the right sums is more important than doing the sums right. In PISA this step would ask if it is acceptable that education be investigated as an input to growth. It would question whether the components measured correspond to what is desirable in the field of education, as discussed in the third section.

Rule 7: "Focus the analysis on the key question answered by the model, exploring holistically the entire space of the assumptions". To assess the total uncertainty in PISA scores, all sources of uncertainty should be activated simultaneously – not just one at a time. As mentioned above, only the standard error is reported by the survey organizers. This should be communicated clearly to the lay audience with a statement such as "uncertainties being taken into consideration, the rank of country X could vary between five and twenty". If PISA country ranks were to prove volatile in a sensitivity audit, the league tables would be received and treated with greater caution.

This checklist may seem exaggerated; however, when a scientific analysis is destined to inform an important policy process, it is reasonable to ask that methodological standards be set high.

Regardless of the potential of both NUSAP and sensitivity auditing to strengthen PISA-based statistics, the most critical improvement that could be made by the PISA organizers would be to adopt the PNS model of extended participation. We find support for this idea in a recent book by Kristina Rizga (2016), discussing what are called "failing schools". Rizga notes:

> Some of the most important things that matter in a quality education – critical thinking, intrinsic motivation, resilience, self-management, resourcefulness, and relationship skills – exist in the realms that can't be easily measured by statistical measures and computer algorithms, but they can be detected by teachers using human judgment. America's business-inspired obsession with prioritizing "metrics" in a complex world that deals with the development of individual minds has become the primary cause of mediocrity in American schools. […]. Educational reforms won't succeed unless there is greater inclusion of the voices of students and teachers […]. (Rizga, 2016, p. xiii)

We find the tension between impersonal metrics and principles of participation and inclusion as described by Rizga a good instantiation of the tensions discussed in the present paper. Note that a similar discussion of over-reliance on metrics is taking place in relation to the evaluation of science and scientists (see e.g. Wilsdon, 2015), where the possible "gaming" of metrics and their potential perverse incentives come into play.

Concluding remarks

The debate over PISA is characterized by conflicting values and divergent interpretations of facts – as our discussion of non-reporting has shown. The letter published in *The Guardian* (2014) shows how various stakeholders could come together to oppose PISA. The specific criticisms listed in the letter motivated us to review the methodological and ideological issues raised in the debate over PISA and to discuss the PISA-based policy discourse. We further illustrate how invoking science in support of policy is anything but unproblematic. In short, this review of the arguments for and against PISA reveals a rich spectrum of methodological and ideological positions which justifies some kind of social action with regard to PISA – that is, an effort to create awareness that PISA is not a neutral statistical exercise producing objective numbers and hard facts.

Given that PISA has considerable impact on policy, the survey designers should make an effort to be more transparent in the documentation of the choices they make in the generation of data. This would include information on the representativeness of data and how the modelling choices impact on the results. The addition of a NUSAP analysis or sensitivity audit of these choices would increase the legitimacy of the data and of the league tables which have attracted so much public attention.

Ideally, PISA should be used to support policies that are beneficial and not harmful to students. But there are those who object to a single view of what constitutes the "good" in education. The "anti-PISA" voices that united to produce the letter to *The Guardian* reflect a plurality of worldviews of education that include the "anti-accountability" movement and a rejection of the economic case for education. The criticisms we reviewed strongly suggest that the case for education should be made in broader terms. Moreover, we show that even when such a worldview is assumed, the aforementioned methodological issues would need to be resolved in order for PISA to have more scientific credibility. In the opinion of the authors the measurement of cognitive skills as a proxy for fitness for a policy goal – including economic growth – is a legitimate approach, provided one is clear as to the aims and limitations of such an exercise. An open and vigilant stance should be maintained against possible undesired effects and on the wider issue of governance. Questions such as "who has the right and the power to produce and use such measures?" and "are they desirable?" are also legitimate.

John Dewey's definition of social science as social inquiry includes a science capable of exploring and tackling existing problems to bring about an improvement in the social life of a community. For Dewey such an inquiry should involve the same actors of the community concerned (Boydston, 1985) and take into account the various sets of norms and values at play in the production of socially robust knowledge. That is, knowledge that is respectful of and robust to the concerns of different actors and stakeholders (Nowotny, 2003). To achieve this we advocate looking at the issue through the lens of PNS. While this might appear a bizarre application of an epistemological approach developed to deal with environmental issues, PNS approaches are today recommended for a host of problems including "eradication of exogenous pests […], offshore oil prospecting, legalization of recreational psychotropic drugs, water quality, family violence, obesity, teenage morbidity and suicide, the ageing population, the prioritization of early childhood education, reduction of agricultural greenhouse gases, and balancing economic growth and environmental sustainability" (Gluckman, 2014).

PNS was suggested by Silvio Funtowicz and Jerome R. Ravetz (1991, 1992, 1993) to address issues where "facts are uncertain, values are in dispute, stakes are high and decisions urgent". Several tensions may be seen as leading to PNS. One more evident in PNS programmatic mantra (stakes-uncertainty-conflict-urgency) is a reaction to reductionist risk and cost-benefit analyses when applied to complex environmental problems (Funtowicz and Ravetz, 1994). This tension is linked to the realization that one thing is to use science to discover and domesticate nature, quite another to use it to fix the damage done to nature by man and his technologies. Another important tension, as laid out in Ravetz's (1971) work , is that one cannot substitute the restricted communities of little science, and their moral and morale, with the societies of big- or mega- or techno-science while pretending to keep the pristine quality assurance arrangements. Both tensions need for their resolution a new type of science, where extended peer communities are called in to keep in check both what science does and how its quality can be maintained.

The letter to *The Guardian* which occasioned this paper reminds us of the importance of reflecting on critical questions such as what constitutes knowledge, who decides what constitutes knowledge (Lyotard, 1979/1984, p. 8), and the relationship between knowledge and social order (Shapin and Schaffer, 1985, p. 15). The reflections offered in this paper are intended to help inform the debate over PISA, which is very likely to reignite after the publication of the 2015 wave of PISA results in December 2016.

References

Atkinson, A.B. (1975), *The Economics of Inequality*, Clarendon Press, Oxford.

Baird, J.-A., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Terra, S. and Daugherty, R. (2011), "Policy effects of PISA", Oxford University Centre for Educational Assessment.

Boden, R. and Epstein, D. (2006), "Managing the research imagination? Globalisation and research in higher education", *Globalisation, Societies and Education*, Vol. 4 No. 2, pp. 223-236.

Q9 Boydston, J.A. (1985), *The Later Works of John Dewey, 1925-1935*, Vol. 6, Southern Illinois University Press, Carbondale, IL,pp. 64-65.

Q3 Breen, R., Luijkx, R., Mueller, W. and Pollak, R. (2009), "Nonpersistent inequality in educational attainment: evidence form eight European countries", *American Journal of Sociology*, Vol. 114, pp. 1475-1521.

Brown, G., Micklewright, J., Schnepf, S.V. and Waldmann, R. (2007), "International surveys of educational achievement: how robust are the findings?", *Journal of the Royal Statistical Society Series A*, Vol. 170, pp. 623-646.

Carnoy, M. and Rothstein, R. (2013), "What do international tests really show about US student Q4 performance?", Economic Policy Institute Report.

Carrozza, C. (2014), "Democratizing expertise and environmental governance: different approaches to the politics of science and their relevance for policy analysis", *Journal of Environmental Policy and Planning*, Vol. 17, pp. 108-126.

Ehrenberg, R., Brewer, D., Gamoran, A. and Willm, D. (2001), "Class size and student achievement", *Psychological Science in the Public Interest*, Vol. 2 No. 1, pp. 1-30.

Fuchs, T. and Woessmann, L. (2007), "What accounts for international differences in student performance? A re-examination using PISA data", *Empirical Economics*, Vol. 32, pp. 433-464.

Funtowicz, S.O. and Ravetz, J.R. (1991), "A new scientific methodology for global environmental issues", in Costanza, R. (Ed.), *Ecological Economics: The Science and Management of Sustainability*, Columbia University Press, New York, NY, pp. 137-152.

Funtowicz, S.O. and Ravetz, J.R. (1992), "Three types of risk assessment and the emergence of postnormal science", in Krimsky, S. and Golding, D. (Eds), *Social Theories of Risk*, Westport, CT, Greenwood, pp. 251-273.

Funtowicz, S.O. and Ravetz, J.R. (1993), "Science for the post-normal age", *Futures*, Vol. 25, pp. 739-755.

Funtowicz, S.O. and Ravetz, J.R. (1994), "The worth of a songbird: ecological economics as a post-normal science", *Ecological Economics*, Vol. 10 No. 3, pp. 197-207.

Gluckman, P. (2014), "Policy: the art of science advice to government", *Nature*, Vol. 507, pp. 163-165.

Goldstein, H. (2004), "International comparison of student attainment: some issues arising from the PISA study", *Assessment in Education*, Vol. 11, pp. 319-330.

Groves, R., Floyd, J., Couper, P., Lepkowski, J., Singer, E. and Tourangeau, R. (2009), *Survey Methodology*, Wiley, Hoboken, NJ.

Hanushek, E. and Woessmann, L. (2012), "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation", *Journal of Economic Growth*, Vol. 17, pp. 267-321.

Heckman, J.J. and Kautz, T. (2012), "Hard evidence on soft skills", *Labour Economics*, Vol. 19, pp. 451-464.

Heckman, J.J., Humphries, J.E. and Kautz, T. (2014), "Who are the GEDs?", in Heckman, J.J., Humphries, J.E. and Kautz, T. (Eds), *The Myth of Achievement Tests: The GED and the Role of Character in American Life*, University of Chicago Press, Chicago, IL.

Hopmann, S., Brinek, G. and Retzl, M. (2009), *PISA According to PISA – Does PISA Keep What it Promises?*, University of Vienna Press, Vienna.

Hutchison, D. and Schagen, I. (2007), "Comparisons Between PISA and TIMSS – Are We the Man with Two Watches?", National Foundation for Educational Research, pp. 1-32.

Jasanoff, S. (1996), "Beyond epistemology: relativism and engagement in the politics of science", *Social Studies of Science*, Vol. 26 No. 2, pp. 393-418.

Jerrim, J. (2013), "The reliability of trends over time in international education test scores: is the performance of England's secondary school pupils really in relative decline?", *Journal of Social Policy*, Vol. 42 No. 2, pp. 259-279.

Kautz, T., Heckman, J.J., Diris, R., Weel, B.T. and Borghans, L. (2014), "Fostering and measuring skills: improving cognitive and non-cognitive skills to promote lifetime success", Working Paper No. 110, OECD Education .

Lyotard, J.F. (1979/1984), *The Postmodern Condition: A Report on Knowledge*, Manchester University Press.

Meyer, H.-D. and Zahedi, K. (2014), "An open letter: to Andreas Schleicher", OECD, Paris, Global Policy Institute, *Guardian*, 5-6 May, available at: www.globalpolicyjournal.com/blog/05/05/2014/open-letter-andreas-schleicher-oecd-paris; www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics (accessed 20 June 2016).

Micklewright, J. and Schnepf, S.V. (2007), "Inequality of learning in industrialised countries", in Atkinson, T. (Ed.), *Inequality and Poverty Re-examined*, Oxford University Press, Oxford.

Q5 Micklewright, J., Schnepf, S.V. and Skinner, C.J. (2012), "Non-response biases in surveys of school children: the case of the English PISA samples", *Journal of the Royal Statistical Society Series A*, Vol. 175, pp. 915-938.

Mont, G. (2011), "Cross-national differences in educational achievement inequality", *Sociology of Education*, Vol. 84, pp. 49-68.

Nowotny, H. (2003), "Democratising expertise and socially robust knowledge", *Science and Public Policy*, Vol. 30, pp. 151-156.

OECD (2004), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003*, OECD Publishing, Paris.

OECD (2010), *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science*, OECD Publishing, Paris.

OECD (2012), *PISA 2012 Technical Report*, OECD Publishing, Paris.

OECD (2013a), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing, Paris.

OECD (2013b), *PISA 2012 Results: What Makes Schools Successful? Ressources, Policies and Practices*, OECD Publishing, Paris.

OECD (2014), "Response to points raised in Heinz-Dieter Meyer 'open letter", available at: www.oecd.org/pisa/aboutpisa/OECD-response-to-Heinz-Dieter-Meyer-Open-Letter.pdf (accessed 20 June 2016).

Porter, T.M. (1995), *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*, Princeton University Press, Princeton, NJ.

Q6 Ravetz, J.R. (1971), *Scientific Knowledge and its Social Problems*, Oxford University Press.

Ravitch, D. (2010), *The Death and Life of the Great American School System: How Testing and Choice are Undermining Education*, Basic Books, New York, NY.

Rizga, K. (2016), *Mission High: One School, How Experts Tried to Fail It, and the Students and Teachers Who Made It Triumph*, Nation Books, New York, NY.

Saltelli, A. and Funtowicz, S. (2014), "When all models are wrong: more stringent quality criteria are needed for models used at the science-policy interface", *Issues in Science and Technology*, Vol. 17, pp. 108-126.

Saltelli, A. and Giampietro, M. (2016), "What is wrong with evidence based policy, and how can it be improved?", Special issue on PNS in FUTURES (forthcoming),available at: www.andreasaltelli.eu/file/repository/FUTURES_Saltelli_Giampietro_6.pdf (accessed 17 June 2016).

Saltelli, A., Pereira, Â.G., Van der Sluijs, J.P. and Funtowicz, S. (2013), "What do I make of your latinorum? Sensitivity auditing of mathematical modelling", *International Journal of Foresight and Innovation Policy*, Vol. 9, pp. 213-234.

Schleicher, A. (2014), "Letter", *Guardian*, 8 May, available at: www.theguardian.com/education/2014/may/08/pisa-programme-short-term-fixes (accessed 20 June 2016).

Sellar, S. and Lingard, B. (2014), "The OECD and the expansion of PISA: new global modes of governance in education", *British Educational Research Journal*, Vol. 40 No. 6, pp. 917-936, doi: 10.1002/berj.3120.

Simola, H. (2005), "The finnish miracle of PISA: historical and sociological remarks on teaching and teacher education", *Comparative Education*, Vol. 4, pp. 455-470.

Smithers, A. (2013), "Confusion in the ranks: how good are England's schools?", The Sutton Trust, available at: www.suttontrust.com/wp-content/uploads/2013/02/CONFUSION-IN-THE-RANKS-SMITHERS-LEAGUE-TABLES-FINAL.pdf (accessed 12 June 2016).

Takayama, K., Waldow, F. and Sung, Y.K. (2013), "Finland has it All? examining the media accentuation of 'finnish education' in Australia, Germany, and South Korea", *Research in Comparative and International Education*, Vol. 8, pp. 307-325.

Waldow, F., Takayama, K. and Sung, Y.K. (2014), "Rethinking the pattern of external policy referencing: media discourses over the 'Asian Tigers', PISA success in Australia, Germany and South Korea", *Comparative Education*, Vol. 50, pp. 302-321.

Q7 Wilby, P. (2014), "Academics warn international school league tables are killing 'joy of learning", *Guardian*, 6 May, available at: www.theguardian.com/education/2014/may/06/academics-international-school-league-tables-killing-joy-of-learning (accessed 20 June 2016).

Wilsdon, J. (2015), "We need a measured approach to metrics", *Nature*, Vol. 523, pp. 129.

Woessmann, L. (2014), "The economic case for education", EENEE Analytical Report 20, European Expert Network on Economics of Education (EENEE), Institute and University of Munich.

Wuttke, J. (2007), "Uncertainties and Bias in PISA", in Hopmann, S., Brinek, G. and Retzl, M. (Eds), *PISA According to PISA*, University of Vienna Press, Vienna.

Q8

Further reading

Funtowicz, S.O. and Ravetz, J.R. (1990), *Uncertainty and Quality in Policy for Science*, Kluwer, Dordrecht.

Meyerhoefer, W. (2009), "Testfähigkeit – was ist das? [Does PISA keep what it promises?]", in Hopmann, S., Brinek, G. and Retzl, M. (Eds), *PISA according to PISA*, University of Vienna Press, Vienna.

Ravitch, D. (2016), "Solving the mystery of the schools, the New York review of books", 24 March, available at: www.nybooks.com/articles/2016/03/24/solving-the-mystery-of-the-schools/ (accessed 12 June 2016).

Shapin, S. and Schaffer, S. (2011), *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*, Princeton University Press, Princeton, NJ.

Strassheim, H. and Kettunen, P. (2014), "When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms", *Evidence and Policy*, Vol. 10 No. 2, pp. 259-277.

Van der Sluijs, J., Craye, M., Funtowicz, S., Kloprogge, P., Ravetz, J. and Risbey, J. (2005), "Experiences with the NUSAP system for multidimensional uncertainty assessment", *Water Science and Technology*, Vol. 52, pp. 133-144.

Corresponding author
Luisa Araujo can be contacted at: luisa.borgesaraujo@ec.europa.eu