

## **Editorial**

### ***The Power of PISA - Limitations and possibilities for educational research***

On 6<sup>th</sup> December 2016, the Programme for International Student Assessment (PISA) releases its report on the achievements of 15-year-olds from 72 countries and economies around the world. This triennial international survey aims to evaluate education systems across 72 contexts by testing skills in Mathematics, Science and Reading Literacy. This is the sixth cycle of PISA and the OECD suggests countries and economies now have the capability to compare the results over time to 'assess the impact of education policy decisions'<sup>1</sup>. Compared to other education studies, the media coverage of PISA must be described as massive (Meyer and Benavot, 2013, Baird et al., 2016) and, as with previous years, it is expected that PISA will attract considerable discussion among policy makers, educators and researchers (Wiseman, 2014). It is therefore timely to present a thematic issue of *Assessment in Education*, where we publish four articles that have analysed previous datasets from the PISA studies each commenting upon the challenges, limitations and potential future assessment research on the PISA data.

The articles touch upon issues regarding sampling, language, item difficulty and demands, as well as the secondary analyses of students' reported experiences of formative assessment in the classroom. One important message from the authors in this thematic Special Issue is the need for a more complex discussion around the use and misuse of PISA data, and the importance of pointing to the limitations of how the results are presented to policy makers and the public. In an area where the media produces narratives on schools and education systems based upon rankings in PISA, researchers in the field of large-scale assessment studies have a particularly important role in stepping up and advising on how to interpret and understand these studies, while warning against potential misuse.

In 2014, Yasmine El Masri gave a key-note at the *Association for Educational Assessment-Europe* conference in Tallinn, Estonia, following her Kathleen Tattersall New Researcher Award. We are pleased to publish the paper based upon her DPhil research: *Language effects in international testing: the case of PISA 2006 items*. Together with Jo-Anne Baird and Art Graesser, El Masri investigates the extent to which language versions of the PISA test in Arabic, English and French are comparable in terms of item difficulty and demand. As there is an ongoing discussion on whether it is possible to assess in a fair manner and compare science, mathematics and reading performances across countries and cultures, this present study offers important findings for future research. By using released PISA items El Masri *et al.* show how language demands vary when comparing Arabic, English and French versions of the same item, and hence could impose different cognitive demands on the students participating in the PISA test in different countries. With the expansion of PISA to other countries through *PISA for Development* and the need for fair comparisons across countries, El Masri *et al.* suggest that subsequent research could explore the possibility of investigating computational linguistics approaches in test

---

<sup>1</sup> <http://www.oecd.org/pisa/aboutpisa/> (downloaded September 30<sup>th</sup> 2016).

transadaptation as an alternative to the use of expert judgment which is the current practice in international test development.

The next article in this issue by Freitas, Nunes Balcao Reis, Seabra and Ferro (2016), *Correcting for sample problems in PISA and the improvement in Portuguese students' performance*, report a study conducted in Portugal where the authors uncovered considerable deviation between the population represented in the PISA samples in 2006, 2009 and 2012 and the effective Portuguese population. The research team therefore addressed problems of representativeness of the PISA samples and recalculated scores using post-stratified weights on the PISA samples in 2006, 2009 and 2012 for Portugal, concluding that the recalculated scores were lower than the ones officially reported by PISA in 2006 and 2009, but in line with the reported findings from 2012. In the PISA 2012 report from Portugal, it is claimed that there is stagnation in school performance from 2009 to 2012, whilst the recalculated scores obtained by the authors of this article actually show an improvement. The authors note that Portugal has always had lower participation rates than the OECD average in PISA, and that they also have a high retention rate among students, which influences the sample. Countries with similar problems of representativeness could, such as England, which did not meet the PISA response rate standard in PISA 2000 and 2003 (Baird et al. 2013, UK Statistics Authority, 2012) can benefit from the strategy suggested in this article, and re-analyse their data using post-stratified weights. As the authors note, it would also strengthen the political value of its reports' conclusion.

In the article *How is formative assessment related to students' reading achievement? Findings from PISA 2009*, Li Hongli reports from the 2009 US PISA study US data set, and provides an analysis of a total of 5233 students from 165 schools. Formative assessment was measured by using nine student questionnaire items, where students had to report on a four-point Likert scale whether they agreed or disagreed with statements such as *The teacher explains beforehand what is expected of the students*, and *The teacher tells students in advance how their work is going to be judged*. Hongli found that formative assessment was positively related to students' reading achievement both directly and indirectly; formative assessment had a positive relationship with students' reading achievement via teacher-student relationships and also with attitudes towards reading. Hongli argues that the nationally representative data-set from PISA confirms previous research claims that formative assessment can improve student learning (Black and Wiliam, 1998, Shepard, 2005) and offers evidence in a research field where there is a significant lack of empirical evidence.

Hongli's study also represents an interesting example of how PISA can be used for secondary analysis in areas such as formative assessment, but there still needs to be a critical review of the student questionnaire in PISA: What are the possibilities and limitations of using student self-report questionnaire data, in addition to the test performance data, from PISA and other International Large Scale Assessment studies? The questionnaire instrument has previously been criticised for not giving reliable results (Hopfenbeck and Maul, 2011,

Samuelstuen, 2007), and OECD has acknowledged such limitations (OECD, 2009, Lie and Turmo, 2005). Still, as in much survey research, PISA has continued to use student questionnaires to assess students' approaches to learning, use of formative assessment, motivation and interest. Particular cautions should be given to the interpretation of the student questionnaire.

John Jerrim in this issue has used data from PISA 2012 to investigate whether students' skills in Mathematics differ between paper and computer versions of the PISA mathematics test. Analysing data from 200,000 students in 32 countries, Jerrim found a substantial drop of more than 50 PISA test points (half a standard deviation) in the average performance of children in Shanghai China when comparing the computer test with the paper and pencil test. Jerrim points out that, although Shanghai is a high performing jurisdiction, it is only on the paper and pencil PISA test that it is exceptional. Students performed better on the paper and pencil tests in eleven other countries, of which some were also high performing jurisdictions, such as Chinese-Tapei, Hong Kong and Singapore. Examining the percentages of students who reached the highest proficiency level in PISA, it was revealed that the decline in achievement was driven for the most part by fewer students being able to reach the top level when administering the computer version of the test in Shanghai-China and Chinese-Tapei. In contrast, Jerrim found that 13 countries performed better on the computer test, among them countries such as Brazil, Columbia and Chile. He further suggests that his findings could have implications for how we should interpret the PISA 2015 results, particularly since the vast majority of countries used computer-based assessments. In a third of the economies, Jerrim has found a difference of more than 10 PISA points between the two versions. As Jerrim rightly points out, first, the OECD has previously claimed changes in the magnitude of 10 points (0.1 standard deviation) as substantial (OECD, 2011, p. 201) and second, since this study has identified patterns where results differ on the two versions of the PISA test, the two modalities for offering PISA – pencil and paper and computer - should be monitored carefully. Jerrim further advises academics, policy makers and journalists to take great care when interpreting results from PISA 2015. I echo this advice.

This thematic issue brings to the readership's attention four key issues requiring further investigation, namely issues of language, secondary analysis of PISA data, the reliability of the self-report in student questionnaire and the two different modalities of PISA – computer based and pencil and paper version - and how these might impact on attained performance levels. The power and influence of PISA on educational research and policy (Hopfenbeck et al 2016) makes it of seminal importance to further monitor and examine these key very real issues that carry weighty implications for consequential validity (Messick 1989) not only for systems but also for individuals.

## References

Baird, J., Johnson, S., Hopfenbeck, T.N., Isaacs, T., Sprague, T., Stobart, G. & Yu, G (2016) [On the supranational spell of PISA in policy](#), *Educational Research*, 58 (2),

121-138. Special Issue: International Policy Borrowing and Evidence-based Educational Policy Making: Relationships and Tensions.

Baird, J., Ahmed, A., Hopfenbeck, T.N., Brown, C. & Elliott, V. (2013) [Research evidence relating to proposals for reform of the GCSE](#). OUCEA Report.

Black, P. and D. Wiliam (1998) Assessment and Classroom Learning, *Assessment in Education: Principles, Policy & Practice*, 5:1, 7 – 74.

El Masri, Y., Baird, J-A., & A. Graesser (2016) Language effects in international testing: the case of PISA 206 science items. *Assessment in Education: Principles, Policy & Practice*, 23:4, p. xx

Freitas, Nunes Balcao Reis, Seabra and Ferro (2016), *Correcting for sample problems in PISA and the improvement in Portuguese students' performance*, *Assessment in Education: Principles, Policy & Practice*, 23:4, p. xx

Hopfenbeck, T.N. & Maul, A. (2011) Examining Evidence for the Validity of PISA Learning Strategy Scales Based on Student Response Processes, *International Journal of Testing*, 11 (2), 95-121.

Hopfenbeck, T.N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J. and J.A. Baird (*accepted for publication*) Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment, *Scandinavian Journal of Educational Research*.

Jerrim, J. (2016) PISA 2012: how do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, 23:4, p. xx

Lie, S. & A. Turmo (2005) Cross-Country Comparability of Students' Self-Reports \_ Evidence from PISA 2003, Internal Working OECD/PISA document, TAG (0505)11.

Meyer, H.-D., and A. Benavot (2013) PISA, Power and Policy the emergence of global educational governance, Oxford Studies in Comparative Education, Symposium Books

Messick, S. (1989) Validity, in R. L. Linn (Ed), *Educational measurement* (3<sup>rd</sup> ed., pp. 13 – 103). New York, Maxmillan.

OECD (2010) *PISA 2009 results: Learning to Learn Student engagement, strategies and practices*, Vol III, Paris, OECD.

OECD (2011) *Lessons from PISA for the United States, strong performers and successful reformers in education*. Paris: OECD Publishing.

Samuelstuen, M, Bråten, I. and Valås, H. (2007) Context Effects in Norwegian 10<sup>th</sup>-Grade Students' Reports on Learning Strategies using the Cross-Curricular

Competencies Instrument, *Scandinavian Journal of Educational Research*, 51 (5): 511 – 529.

Shepard, L.A. (2005) Linking formative assessment to scaffolding. *Educational Leadership*, 63, 66 – 71.

UK Statistics Authority (2012) Letter from Andrew Dilnot to David Miliband. Available online at: <http://www.statisticsauthority.gov.uk/reports---correspondence/correspondence/index.html>, last accessed January 2013.

Wiseman, A.W. (2014) *Policy Responses to PISA in Comparative Perspective*. In Meyer, H.-D., and A. Benavot (2013) *PISA, Power and Policy the emergence of global educational governance*, Oxford Studies in Comparative Education, Symposium Books