Can the outcomes of PISA 2015 contribute to evidence-based decision making in mathematics education?

Gerry Shiel

Can the outcomes of PISA 2015 contribute to evidence-based decision making in mathematics education?

Gerry Shiel

Drawing on data from the OECD's Programme for International Assessment (PISA), which assesses mathematical literacy and other domains among 15-year olds in over 70 countries every three years, this paper explores the extent to which PISA outcomes in 2015 can be described as 'solid' and hence contribute to evidence-based decision making. It identifies aspects of PISA that render its findings 'solid', but also points to pitfalls that arise in interpreting PISA outcomes related to achievement. The paper concludes by examining how PISA can contribute to thinking about the nature of evidence-based findings in mathematics education.

Introduction

A key feature of the educational landscape since 2000 has been the Programme for International Student Assessment (PISA), a study sponsored by the Paris-based Organisation for Economic Cooperation and Development (OECD) that assesses performance in mathematics, reading literacy and science among 15-year olds in over 70 countries every three years. In addition to administering tests to students, PISA administers questionnaires to students, their parents and their school principals. The student questionnaire asks about students' socioeconomic status, their attitudes towards mathematics and other subjects, and their instructional experiences. This paper looks at performance outcomes in the two most recent PISA cycles – 2012, when mathematics was a major assessment domain, and 2015, when mathematics was a minor domain, and PISA moved from a paper-based to computer-based testing in most participating countries.

Interest in the extent to which PISA provides 'solid' or 'evidence-based' findings arises because of the strong impact that PISA has on policy making in many participating countries. In Ireland, for example, a significant drop in performance in mathematics and reading literacy in PISA 2009 led to the implementation of a *National Strategy to Improve Literacy and Numeracy 2011-2020* (DES, 2011). The strategy set out a series of measures designed to improve performance, including plans to enhance initial teacher education, curriculum and assessment. In parallel with the Strategy, revised curricula in mathematics at post-primary level have been rolled out in a phased basis since 2010 in an initiative known as 'Project Maths'. This involves a strong focus on developing students' conceptual understanding in mathematics, and on applying mathematical knowledge in solving problems in context using a range of methods. Ní Shuilleabháin (2013) described Project Maths as 'a philosophical shift in Irish post-primary classrooms from a highly didactic approach with relatively little emphasis on problem solving towards a dialogic, investigative problem-focused approach to teaching and learning mathematics' (p. 23).

A key feature of the National Strategy is the inclusion of national targets for performance in PISA mathematics. In an interim review of the Strategy (DES, 2017), there are targets of 10.5% of students achieving below proficiency level 2 by 2020, and 12.0% achieving levels 5-6. The first of these is quite an ambitious relative to current performance (15% performed below Level 2 in 2015), while the second is more modest (11% performed at Levels 5-6 in 2015).

Efforts to ensure that PISA findings are solid

The procedures around the development of PISA survey instruments, including the mathematics test, are designed to ensure that findings can be relied on and used by participating countries to enhance teaching and learning, and raise performance standards. The development of the PISA mathematics test and scale encompasses the following:

- An assessment framework is developed and published at the outset of each PISA cycle (e.g., OECD, 2013). The framework provides a definition of mathematical literacy in PISA, and outlines the content areas (mathematical content categories) and processes to be assessed, the contexts in which items are to be embedded and the item formats to be used. Items are then developed in a way that ensures that all elements of the framework are adequately addressed. The assessment framework is a key source of evidence to support the validity of the PISA tests.
- Items based on the framework are submitted by countries, or are developed by the consortium charged by the OECD with implementing PISA. Items are vetted by countries for cultural and linguistic appropriateness and suitable items are forwarded for field trialling.
- The PISA field trial is conducted on a sample of 15-year olds in each participating country, and the performance of items is assessed within and between countries. The outcomes of both classical item analysis and item response theory scaling are taken into account in determining the suitability of items. These items, along with any trend items not field-trialled, are then used to compile test forms for the main study.
- Considerable effort goes into ensuring that items are scored accurately, using scoring guides prepared by the PISA consortium. Many items are marked by two or four scorers, and real-time indices of inter-rater reliability are used to guide the quality of scoring.
- The PISA main study is implemented. Quality control is a key aspect of the Main Study, as countries are held accountable to quality standards (see below).
- Performance on PISA is scaled using Item Response Theory models and links with performance on earlier rounds are established.

A document, *PISA Technical Standards* (e.g., OECD, 2014), is issued in each cycle to guide countries in ensuring that their samples, response rates, security procedures, translation and coding practices are of a sufficiently high standard that their data warrants inclusion in international reports. For example, the 2015 Technical Standards indicate that response rates of 85% at school level and 80% at the student level are required. The achieved samples of countries failing to meet these criteria are examined in detail for potential bias. In some cases, countries have not been included in international reports because of low response rates (e.g., the Netherlands in 2000, and the UK in 2003).

At the end of each PISA cycle, a technical report is prepared by the PISA consortium and is issued by the OECD (e.g., OECD, 2017). It details the procedures used in each aspect of the implementation of PISA, including sample design, field operations, quality control, survey weighting, scaling, proficiency scale construction, and coding reliability.

The consortium charged with implementing PISA establishes expert groups for mathematics, science and reading literacy, and there is also a Technical Advisory Group, which advises the Consortium on its use of scaling and other procedures, and a Questionnaire Expert Group. These groups act as a further check on the quality of the PISA instruments and outcomes.

Hence, PISA has taken several precautions to ensure the quality and solidity of its findings. Notwithstanding the fact that PISA assesses the mathematics that students require for life after they leave school (or mathematical literacy) and for future study, rather than mathematics based on school curricula, the steps taken to ensure that findings are solid are extensive.

The introduction of computer-based assessment as a threat to the solidity of PISA findings

Prior to 2015, PISA implemented computer-based testing in subsets of countries on an optional basis. In 2012, for example, mathematics was assessed on paper in all 65 participating countries, and on computer on an experimental basis in a subsample of 32 countries. In 2015, however, there was a shift to computer-based assessment in most participating countries, with 56 of 73 countries, including all 34 OECD member countries, administering PISA in this format. The remaining countries administered PISA on paper.

The transition to computer-based testing in PISA presented some significant challenges for the OECD. A key component of PISA is the availability of trend data – that is, performance from one PISA cycle to the next must be placed on the same underlying scale so that average performance and performance across proficiency levels in each country and on average across OECD countries can be tracked from cycle to cycle. The task facing the OECD and its contractors[1] was to establish the feasibility of linking performance on the 2015 computer-based tests to scales based on performance on paper-based tests in earlier cycles. This was further complicated by a requirement to continue to provide trend data for countries that administered PISA in paper-based form in 2015.

There were several ways in which the transition to computer-based testing could have been managed, given the imperative to maintain trends. For example, all students (or equivalent samples of students) taking PISA 2015 could have been given paper-based and computer-based tests. Then trends could have been established with reference to performance on the paper-based measures and new computer-based scales could have been devised, based on the computer-based items, and used for trend analysis in the future. This would have eliminated any concerns about mode effects (an advantage or disadvantage arising from implementing PISA on computer).

The approach taken by the OECD and its contractors was to make adjustments in 2015 based on how the same items performed on paper and on computer in the PISA 2015 Field Trial, which took place in all participating countries in spring or autumn 2014. In the case of mathematics, which was a minor domain, items used in earlier PISA cycles (i.e., trend items) were transferred from paper to computer, and equivalent representative samples of students from each country took the paper- and computer-based tests. Hence, the purpose of the mode study was to ascertain whether tasks or items

[1] The lead contractor in PISA 2015 was the Educational Testing Service in the US. The lead contractor in all earlier cycles of PISA was the Australian Council for Educational Research.

presented in one mode (i.e., paper) functioned differently when presented in another mode (i.e., computer) and vice versa. For the purpose of analysis, items were pooled across countries, as individual countries did not have sufficiently large samples of students to allow for reliable comparisons of individual items across modes, or for an analysis of item-by-country interactions. Where item parameters were judged to be 'strongly invariant' (that is, similar on paper and computer), item parameters were constrained to be the same in the 2015 Main Study (OECD, 2017). In the course of the Field Trial analysis, a subset of items showed mode effects. To account for these effects in the Main Survey, different item parameters were estimated for paired paper- and computer-based items. According to the OECD (2017, Chp. 7, p. 53), 'this established an invariance model that assumes scalar or strong invariance for the majority of items and metric invariance for a minority of items for which difficulty differences were detected'. A correlation of .95 was found between paper-based and computer-based item parameters for mathematics in the Field Trial, further supporting a link between performance on computer-based tests in 2015 and paper-based tests in earlier cycles, as well as between computer- and paper-based tests administered in 2015.

The PISA 2015 Field Trial yielded other interesting findings that applied to mathematics as well as other domains. For example, across countries, students taking the Field Trial tests on computer had significantly fewer omitted responses than students taking the paper versions. Furthermore, there were fewer effects of cluster positon on performance when tests were administered on computer (that is, items administered by computer were more likely to perform in the same way regardless of whether they appeared early or late in the test). However, as Jerrim et al. (in press) note, while the Field Trial did not yield large differences across modes for male and female students, no analyses were conducted to examine potential interactions with variables such as ethnicity or socioeconomic status. They also questioned the representativeness of the samples used in the Field Trial, which, in some countries, could be described as convenience samples. They viewed this as weakening the external validity of the results, given the implications for the adjustments made within Main Study scaling to enhance cross-mode comparability.

Overall performance on PISA 2015 mathematics

The PISA main study took place in all participating countries in 2015. The OECD issued two volumes of findings in December 2016 that included country mean scores in mathematics, and comparisons with performance in earlier cycles. The mean score of students in Ireland in 2015 was 503.7 (OECD, 2016). This was significantly above the average across OECD countries (490.2), and was about the same as in 2012 (501.5), 2006 (501.5) and 2003 (502.8). Indeed, the only year in which average performance moved outside the 501-504 range was in 2009 (487.1).

While the mean mathematics score of students in Ireland was stable in the transition to computer-based assessment, a number of countries saw large declines in performance between 2012 and 2015. These included Korea (down 29.7 score points, though still well above Ireland at 517.4), Chinese Taipei (17.5), Hong Kong (13.3), Poland (13 points), and the Netherlands (10.7 points). On the other hand, a small number of countries experienced increases in achievement, including Sweden (15.7 points), Norway (12.4), the Russian Federation (11.9), and Denmark (11.1).

It is noteworthy, however, that Norway, Denmark and the Russian Federation were among the countries with the highest use of computers by students in mathematics classes in PISA 2012 for

purposes such as entering data on a spreadsheet, drawing a graph of a function, constructing geometric figures, re-writing algebraic expressions and solving equations (OECD, 2015). In contrast, Korea, Hong-Kong China and Ireland were among the countries with the lowest usage of ICTs by students in mathematics classes.

The fact that Ireland's overall performance on PISA 2015 is similar to 2012 can be interpreted in a number of ways:

- It suggests that students in Ireland are equally adept as solving mathematical problems in paper and computer-based formats; indeed, this would suggest that the mode of assessment does not matter, at least for students in Ireland.
- It suggests that students in Ireland improved in their mathematics between 2012 and 2015, but this improvement was largely hidden because of the transition to computer-based testing.

The second of these seems the most likely. PISA 2015 was the first PISA cycle in which all students in Ireland's sample had studied under the Project Maths syllabus. This interpretation is consistent with a finding that students in initial Project Maths schools (24 schools that had implemented Project Maths first) outperformed students in non-initial schools in PISA 2012 mathematics (see Merriman et al. 2013), though the difference was relatively small (4 score points) and not statistically significant.

A further relevant finding relates the optional computer-based assessment of mathematics administered as part of PISA 2012. In that assessment, students in Ireland had a mean score that was not significantly different from the corresponding OECD average score, despite achieving a mean score on paper-based mathematics that was significantly above the corresponding OECD average in the same year (Perkins et al., 2013). Hence, performance on PISA 2015 can be interpreted as being indicative of a possible improvement.

Interestingly, the OECD has continued to hold the positon that mode effects in PISA 2015 mathematics were small and did not impact on the performance of participating countries (OECD, 2016, 2017). Implicit in this is the view that performance on computer-based assessment in 2015 can be linked back to performance on paper-based assessment in earlier PISA cycles.

Other threats to the solidity of PISA 2015 findings

The transition to computer-based assessment in PISA is clearly one threat to the validity of scores reported by the OECD for PISA 2015 mathematics. However, there were several other changes to PISA 2015 which could also impact on the interpretation of outcomes, and hence the solidity of PISA findings. The changes – several of which occurred because a new consortium was contracted by the OECD to gather and analyse PISA data – include:

- Changes in the assessment design – the design of PISA 2015 was modified to reduce or eliminate differences in construct coverage for major and minor assessment domains for test takers. In practice, this meant that fewer students took mathematics in PISA 2015, compared with earlier cycles, but more mathematics items were included in the assessment, thereby allowing for broader construct coverage.

- Changes in the calibration sample – prior to 2015, item difficulty in PISA was estimated using the responses of students in the most recent cycle (e.g., in 2012, this comprised data from students who took PISA in 2009). Moreover, the calibration sample in earlier cycles comprised a random sample of 500 students per participating country. In 2015, item parameters were re-estimated using all students in all participating countries in the previous four PISA cycles. This change was implemented to reduce the uncertainty around estimates of the item parameters used in calibration.
- Changes to the scaling model – in earlier PISA cycles, a one-parameter Item Response Theory (IRT) model (with adjustment for partial credit) was used to scale performance. In 2015, item functions based on a two-parameter logistic IRT model for dichotomous data, and a generalized partial-credit model for polytomous data were used in scaling data in the case of new items, while functions based on a one-parameter model were used (as previously) with trend items. Unlike its predecessor, the new approach does not give equal weighting to all items when constructing a score, but assigns optimal weights to tasks based on their capacity to distinguish between high- and low-achieving students.
- Changes in the treatment of differential item functioning across countries – where items performed unexpectedly differently across countries, the calibration in 2015 allowed for a number of country-by-cycle-specific item parameters. In previous cycles, items that showed differential item functioning (e.g., because of differences across languages) were dropped from scaling. The change in 2015 was intended to reduce the dependency of country rankings on the selection of items included in the assessment (for a country) and hence improve fairness (OECD, 2016).
- Changes in the treatment of not-reached items – in PISA 2015, not-reached items (unanswered items at the end of a section, such as at the end of the first and second hour of testing) were treated as not administered when estimating proficiency (i.e., scoring student responses), whereas in previous PISA cycles they were treated as incorrect. A reason for this change was to eliminate the opportunity for countries and test takers to randomly guess answers to multiple-choice questions at the end of a section of the test. As in previous cycles, not-reached items were treated as not administered when computing item parameters (i.e., during scaling).

The OECD (2016) acknowledges that improvements to the PISA test design and to scaling in PISA 2015 can be expected to result in reductions in link error (the error associated with particular sets of items used in a particular cycle) between 2015 and future cycles. However, it also acknowledges that the changes described above may result in increased link error between PISA 2015 and earlier cycles, as past cycles used a different design (paper-based assessment) and used different scaling procedures. Furthermore, the OECD (2016) acknowledges that the change in the treatment of not-reached items could result in higher scores than would have been estimated in earlier PISA cycles for countries with many unanswered items.

Conclusion

The problem in terms of interpreting trend scores is that any of the changes implemented by the OECD and their contractors in relation to the design and scaling of PISA in 2015 could have impacted on the scale scores achieved by students. Interpretation becomes even more difficult when multiple changes are implemented, as these may interact with one another in complex ways. The OECD has sought to address this in a limited way by rescaling data from earlier PISA cycles using the methods implemented in 2015. Thus, in the case of Ireland, performance on PISA mathematics changed by +2 score points between 2012 and 2015 (see above), but, the change was 6.0 score points when newer scaling methods were applied to the 2012 mathematics data. On average across OECD countries, the impact of changes to scaling procedures was also reported to be small (a published drop of 3.7 score points between 2012 and 2015, and a drop of 2.5 score points following rescaling of the 2012 data) (OECD, 2016). For most countries, differences arising from re-scaling are within the error margins of the original difference scores reported by the OECD.

While the readjustment of scores from PISA 2012 using the new scaling procedures implemented in 2015 may go some way towards reassuring users that PISA outcomes are comparable over time, the sheer number of changes implemented in PISA 2015, including the change to computer-based testing, indicates that particular care should be exercised in interpreting PISA 2015 data.

Efforts to improve the design and scaling of PISA 2015 also contain some lessons for efforts to generate solid data in mathematics education. On the one hand, solid findings can be obtained by implementing the same testing procedures and methodologies on multiple occasions (e.g., pre- and post-intervention). In the words of Beaton (1990), 'when measuring change, do not change the measure' (p. 165). On the other hand, at least in the case of longitudinal, multi-year surveys such as PISA, there is an ongoing need to build innovation into all aspects of the project to maintain relevance and deliver more robust measures for the future. One clear danger is that, when mathematics becomes a major assessment domain in PISA 2012, the construct measured will also change, as new items specifically designed to take advantage of the affordances computers, will be introduced for the first time.

Acknowledgements

Participants who sent us questions during and after the discussion debate are warmly thanked, with the hope that they will see their issues, concerns and interrogations included in the lines above.

References

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science*, *11*(5), 181–185.

Ashcraft, M. H., & Moore, A. M. (2009). Mathematics anxiety and the affective drop in performance. *Journal of Psychoeducational Assessment*, *27*(3), 197-205.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *553* (7604), 452-454.

Baumgartner, H., & Steenkamp, J. B. E. M. (1998). Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. *Marketing Letters*, *9*, 21-35.

Beaton, A.E. (1900). Epilogue. In A.E. Beaton & R. Zwick (Eds.), *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (Tech Rep. No. 17-TR-21). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Bell, A.W. (1976). A study of pupil's proof-explanations in mathematical situations. *Educational Studies in Mathematics*, *7*, 23-40.

Carey, E., Hill, F., Devine, A., & Szűcs, D. (2017). The modified abbreviated math anxiety scale: A valid and reliable instrument for use with children. *Frontiers in Psychology: Developmental Psychology*, 8, 11. http://dx.doi.org/10.3389/fpsyg.2017.00011.

Caviola, S., Primi, C., Chiesi, F., & Mammarella, I. C. (2017). Psychometric properties of the Abbreviated Math Anxiety Scale (AMAS) in Italian primary school children. *Learning and Individual Differences*, *55*, 174-182.

Cipora, K., Szczygieł, M., Willmes, K., & Nuerk, H. C. (2015). Math anxiety assessment with the abbreviated math anxiety scale: Applicability and usefulness. Insights from the Polish Adaptation. *Frontiers in Psychology*, 6.

DES. (2017). *National Strategy. Literacy and numeracy for learning and life 2011-2020. Interim review: 2011-2016. New targets: 2017-2020*. Dublin: Author.

DES. Depart of Education and Skills (Ireland). (2011). *Literacy and numeracy for learning and life. The National Strategy to improve literacy and numeracy among children and young people 2011-20*. Dublin: Author.

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics Anxiety: What Have We Learned in 60 Years? *Frontiers in Psychology*, *7*.

Dreyfus, T. on behalf of the Education Committee of the EMS (2014). Solid findings: concept images in students' mathematical reasoning. *Newsletter of the European Mathematical Society, 93*, 50-52.

Dreyfus, T., & Becker, J. (1998). What are the results of research in mathematics education? Report of Working Group 4. In A. Sierpinska & J. Kilpatrick (Eds.), *Mathematics education as a research domain: a search for identity* (pp. 23-28). Dordrecht, The Netherlands: Kluwer.

Eden, C., Heine, A., & Jacobs, A. M. (2013). Mathematics anxiety and its development in the course of formal schooling—a review. *Psychology*, *4*(06), 27

Education Committee of the European Mathematical Society (2011a). "Solid findings" in mathematics education. *Newsletter of the European Mathematical Society, 81*, 46-48.

Education Committee of the European Mathematical Society (2011b). Do theorems admit exceptions? Solid findings in mathematics education on empirical proof schemes. *Newsletter of the European Mathematical Society*, *82*, 50-53.

Even, R., & Kvatinsky, T. (2010). What mathematics do teachers with contrasting teaching approaches address in probability lessons? *Educational Studies in Mathematics*, *74*, 207–222.

Fischbein, E. (1982). Intuition and proof. *For the Learning of Mathematics*, *3*(2), 9-18.

Galla, B. M., & Wood, J. J. (2012). Emotional self-efficacy moderates anxiety-related impairments in math performance in elementary school-age youth. *Personality and Individual Differences*, *52*(2), 118–122.

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological Science*, *24*(10), 2079–2087.

Green, C., Taylor, C., Buckley, S., & Hean, S. (2016). Beyond synthesis: augmenting systematic review procedures with practical principles to optimise impact and uptake in educational policy and practice. *International Journal of Research & Method in Education, 39*(3), 329-344.

Hammersley, M. (2011). *Methodology. Who needs it?* London: Sage.

Harari, R. R., Vukovic, R. K., & Bailey, S. P. (2013). Mathematics anxiety in young children: An exploratory study. *The Journal of Experimental Education, 81*(4), 538–555.

Healy, L., & Hoyles, C. (2000). A Study of proof conceptions in algebra. *Journal for Research in Mathematics Education 31*, 396-428.

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, 33-46.

Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS): Construction, validity, and reliability. *Assessment, 10*(2), 178–182.

Jameson, M. M. (2013). The development and validation of the Children's Anxiety in Math Scale. *Journal of Psychoeducational Assessment, 31*(4), 391-395.

Jerrim, J., Micklewright, J., Heine, J-H, Salzer, C., & McKeown, C. (in press). *PISA 2015: How big is the 'mode effect and what has been done about it?*

Karasel, N., Ayda, O., & Tezer, M. (2010). The relationship between mathematics anxiety and mathematical problem solving skills among primary school students. *Procedia - Social and Behavioral Sciences*, 2(2), 5804–5807.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543

Merriman, B., Shiel, G., Cosgrove, J., & Perkins, R. (2014). *Project Maths and PISA 2012. Performance in initial Project Maths schools and non-initial schools on PISA 2012 mathematics and problem solving, and on Junior Cycle mathematics*. Dublin: Educational Research Centre.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-130.

Morsanyi, K., Mammarella, I. C., Szucs, D., Tomasetto, C., Primi, C., Maloney, E. A., eds. (2017). *Mathematical and Statistics Anxiety: Educational, Social, Developmental and Cognitive Perspectives*. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-076-3

Ní Shuilleabháin, A. (2013). Lesson study in a community of practice: A model of in-school professional development. *Trinity Education Papers, 2(1),* 22-40.

OECD. (2014). *PISA technical standards*. Paris: Author.

OECD. (2015). *Students, computers and learning: Making the connection*. Paris: OECD Publishing.

OECD. (2016). *PISA 2015 results. Excellence and equity in education (vol. 1)*. Paris: OECD Publishing.

OECD. (2017). *PISA 2015 Technical report*. Paris: Author.

OECD. Organisation for Economic Cooperation and Development. (2013). *PISA 2012 assessment and analytic framework: mathematics, reading, science and problem solving and financial literacy*. Paris: OECD Publishing.

Perkins, R., Shiel, G., Merriman, B., Cosgrove, J., & Moran, G. (2013). *Learning for life: The achievements of 15-year olds in Ireland on mathematics, reading literacy and science in PISA 2012*. Dublin: Educational Research Centre.

Pinto (2013). Variability in university mathematics teaching: A tale of two instructors. In B. Ubuz, Ç. Haser & M. A. Mariotti (Eds.), *Proceedings of the Eighth Congress of the European Society for Research in Mathematics Education* (pp. 2416-2425). Ankara, Turkey: Middle East Technical University.

Polya, G. (1945). *How to solve it.* Princeton, NJ: Princeton University Press.

Primi, C., Busdraghi, C., Tomasetto, C., Morsanyi, K., & Chiesi, F. (2014). Measuring math anxiety in Italian college and high school students: Validity, reliability and gender invariance of the Abbreviated Math Anxiety Scale (AMAS). *Learning and Individual Differences*, *34*, 51–56.

Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2013). Math anxiety, working memory, and math achievement in early elementary school. *Journal of Cognition and Development*, *14*(2), 187–202

Schoenfeld, A. H. (2007). Method. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 69-110). Greenwich, CT, USA: Information Age Publishing.

Shiel, G., Kelleher, C., McKeown, C., & Denner, S. (2015). *Future ready? The performance of 15-year-olds in Ireland on science, reading literacy and mathematics in PISA 2015*. Dublin: Author.

Sierpinska, A., & Kilpatrick, J. (Eds.) (1998). *Mathematics education as a research domain: a search for identity*. Dordrecht, The Netherlands: Kluwer.

Sowder, L., & Harel, G. (2003). Case studies of mathematics majors' proof understanding, production, and appreciation. *Canadian Journal of Science, Mathematics and Technology Education*, *3*, 251-267.

Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limit and continuity. *Educational Studies in Mathematics, 12*, 151-169.

Vahedi, S., & Farrokhi, F. (2011). A confirmatory factor analysis of the structure of abbreviated math anxiety scale. *Iranian Journal of Psychiatry*, 6(2), 47–53.

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*(4), 261–329.

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the Statistical Anxiety Scale. *Psicothema*, *20*, 174-180.

Vinner, S., & Hershkowitz, R. (1980). Concept images and common cognitive paths in the development of some simple geometrical concepts. In R. Karplus (Ed.), *Proceedings of the 4th Annual Meeting for the Psychology of Mathematics Education* (pp. 177-184). Berkeley, CA: PME.

Wu, S. S., Barth, M., Amin, H., Malcarne, V., & Menon, V. (2012). Math anxiety in second and third graders and its relation to mathematics achievement. *Frontiers in Psychology*, *3*, 162.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology, 61*, 319-328.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). Charlotte, NC: Information Age Publishing.