Special issue on summative assessment

What can summative assessment in mathematics education tell us?

Paola Iannone* and Ian Jones*
Mathematics' Education Centre
Loughborough University

This special issue presents research into the summative assessment of mathematical learning. Summative assessment is perhaps an unfashionable topic for research, having been eclipsed over recent decades by formative assessment. Yet few would deny the influence, for good or ill, of summative assessment on the practice of teaching and learning mathematics. In fact, there is much high-quality research into the summative assessment of mathematical learning, but it is rarely collated into one place. Consequently researchers working on similar problems can feel isolated and lack a forum for sharing ideas and findings. The aim of this special issue, which comprises six original research articles and two book reviews, grouped to reflect emerging themes of validity and social consequences of assessment, is to provide such a forum.

Scholarly discussion of summative assessment commonly uses the language of validity. Accordingly, five of the six research papers included in this issue use the word "valid" or its derivatives. However, there is little consensus on the nature of test validity across researchers. Colin Foster explores this issue in his review of Paul Newton and Stuart Shaw's (2016) recent book *Validity in Educational and Psychological Assessment*. Foster notes that the definitions and value of even ostensibly straightforward notions such as construct, convergent and predictive validity are hotly debated. In the case of mathematics assessment, Foster draws our attention to a state of affairs that is perhaps less controversial: high-stakes mathematics tests tend to contain predominantly procedural items that assess the recall of facts and the application of algorithms. This promotes an impoverished view of mathematical understanding that has a negative backwash on teaching and learning in mathematics classrooms. It is appropriate, then, that some of the papers in this special issue focus on the development of tests intended to capture a richer view of mathematical understanding.

In this spirit Holmes, He and Meadows evaluate the outcome of recent drives to increase the quantity and quality of problem solving items in General Certificate of Secondary Education (GCSE) Mathematics examinations, which are taken by the majority of students aged 16 in England and Wales. Their research is situated within the tension of improving the richness of national assessments while maintaining a political and moral requirement for consistent marking by a large pool of examiners. Rather than defining problem solving and then applying this definition to the analysis of GCSE examinations, Holmes, He and Meadows instead collate expert perceptions of the problem-solving skills required to score highly on selected test questions. They employ two related methods for evaluating examinations: comparative judgement (Bramley, 2007), which enables items to be quantified in terms of the 'amount' of problem solving they assess, and Kelly's Repertory Grids (Suto & Nádas, 2009), which enable broad constructs, in this case mathematical problem solving, to be broken down into sub-constructs. Holmes, He and Meadows report that the teachers and examiners who participated in their study considered problem solving items to be those that offer students no standard method and allow multiple possible

approaches to deriving solutions. Conversely, participants did not strongly associate problems to be solved within 'real world' contexts with the broad construct of mathematical problem solving. The authors note that this finding contrasts with well-known approaches to problem solving as exemplified by assessment materials associated with the Bowland Maths Initiative (Swan & Pead, 2008) and Realistic Mathematics Education (van den Heuvel-Panhuizen, 1996).

Continuing the focus on validity, Mejía-Ramos, Lew, de la Torre and Weber describe the processes of design and evaluation of a short test aimed at assessing undergraduate students' comprehension of proofs. Undergraduate mathematics students notoriously find it difficult to read and produce proofs, and mathematics educators often cannot assess whether students have understood the proof presented in their teaching (Weber, 2012). In order to facilitate such assessment, the authors describe the design and validation of short tests to assess the comprehension of three well-known proofs (common to many university curricula), with the intent of generalising the process to the construction of short tests to assess other proofs. The authors list some practical and some theoretical benefits for the availability of such short tests. There are benefits for teachers, who have a time-effective tool to assess their students' progress on understanding proofs, and there are benefits for students, who can identify important aspects of proofs that they may have not considered. These tests can also have some interesting applications for research. They could be used, for example, to evaluate alternative ways of presenting proofs, such as those described by Leron (1983) and Rowland (2001), or to help categorise those aspects of proof that undergraduate students tend to find most difficult.

Mac an Bhaird, Nolan, O'Shea and Pfeiffer start from the widely reported concern in the literature that students can be successful in their university mathematics education by mostly employing repetitive reasoning and not engaging in conceptual understanding and problem solving (Fukawa-Connelly, 2005). To this end, the authors analyse tasks in terms of opportunities for (global) creative reasoning (GCR, in the sense of Lithner, 2008) in calculus modules offered to first year students in one specialist module (pure mathematics) and two non-specialist modules (science mathematics and business and mathematics). The authors classified 632 tasks collected from assignments, tutorial work and examinations using the Lithner (2008) task analysis framework. They found that although the lecturers on the three modules all made efforts to include GCR tasks in their teaching and formative assessment, the tasks in the pure mathematics module required a much higher percentage of GCR than the other two modules. Interestingly, when restricting the analysis to examination tasks, all three modules presented similar (low or near zero) percentage of GCR tasks but with the specialist module having a higher percentage of GCR tasks in the examinations. The authors offer a variety of explanations for their findings but conclude that, given the influence that summative assessment has on student engagement with learning (Scouller, 1998), it is worrying that so few creative reasoning tasks are found in non-specialist modules.

Bramley investigates the validity of summative mathematics examinations, but takes a very different approach to other authors in this special issue. Like Holmes, He and Meadows, his interest is in the validity of GCSE examinations, but his focus is on the effects of differentiated examination papers. In England, national mathematics examinations are tiered: there is a Higher tier for students expected to obtain higher grades, and a

Foundation tier for students expected to obtain lower grades. Students are entered for just one tier (Higher or Foundation) and are awarded a grade based on an aggregation of their total marks from the papers they have taken. A different model for tiering, used until recently in Scotland, is the adjacent levels model: students are usually entered for an adjacent pair out of three tiered papers, each targeted at a different and non-overlapping range of grades. Students are awarded the highest grade achieved with no aggregation of total marks across papers. Bramley explores the implications for validity of such approaches using simulation techniques based on scores from a GCSE mathematics examination, and finds that different approaches have both strengths and dangers. The current GCSE tiering model uses all a student's marks to award a grade, which maximises reliability, but offers more than one route to a given grade, thereby threatening validity because a given grade cannot be associated with particular knowledge, skills and understanding. The adjacent levels model discards some marks when awarding grades, which reduces reliability, but could offer more valid outcomes because there is only one route to each grade.

Bramley's contribution is concerned with validity but leads into issues of social consequences of assessment. The differentiation of examinations, and the method adopted for differentiation, has an impact on teachers and students in mathematics classrooms. Teachers must decide which students will sit which papers early on in their GCSE study. His contribution thereby provides a link between the articles in this special issue that we have discussed so far and those we discuss below. McCusker's review of White's (2014) book *Who Needs Examinations? A Story of Climbing Ladders and Dodging Snakes* provides an introduction to the focus of the remaining articles, namely the social consequences of high stakes examinations. McCusker reminds us of the strength of these consequences: examinations have been used to maintain the social order, being introduced in England in the 19th Century as a meritocratic route that enabled less advantaged children to compete academically, and becoming later a tool for the middle classes to maintain their status. McCusker considers White's alternative proposal to the current assessment status quo, namely assessment of pupils' profiles based in teacher assessment, but remains unconvinced. While recognising the value of challenging current assessment practices, the dangers of such profile-based assessment, especially when extended to personality traits, are considerable and may be hard to overcome. The risk of just perpetuating the social order as it is rather than promoting social change remains. The last two papers in this special issue explore exactly some of the social consequences of high stakes assessment, which are linked to the analysis of the status quo which McCusker highlights in his review.

The article by Marinho, Leite and Fernandes explores the depth of the social consequences of national assessment regimes based on examinations. They examine assessment practices in two high schools in Portugal: one at the very top of the league tables (in terms of academic results) and the other at the bottom. Through observations of classroom practice and thematic analysis of semi-structured interviews with teachers, Marhino, Leite and Fernandes present a comparative case study that illuminates the social consequences of summative assessment. They find that in both schools the educational practices observed are akin to an examining pedagogy rather than about teaching and learning, and this is due to the extreme importance of state-wide mathematics examinations. Both sets of participant teachers believe that through summative tests they can ultimately motivate students and so enhance learning. The authors also find, however, that this is true only for

the higher-achieving students. High stakes summative tests remain, for lower achieving pupils, only a mechanism of certification and exclusion rather than a motivation for learning.

Logan and Lowrie adopt a longitudinal approach to investigate the consequences of aspects of national examinations in Australia. Their focus is on gender differences in performance on examination questions that require spatial reasoning. Previous research provides mixed evidence that, on average, males outperform females on tasks requiring spatial visualisation and spatial orientation. To investigate gender differences in performance on examination questions requiring spatial reasoning, Logan and Lowrie draw on national assessment data spanning five years. They find no overall systematic differences in performance between males and females, but report that males outperform females on specific types of examination questions at particular ages. Their analysis enabled nuanced conclusions to be drawn. For example, there was no gender difference on items requiring a single aspect of spatial orientation movement, but males outperformed females when two aspects of orientation movement were required. Through fine-grained analysis of a large dataset the authors offer interpretations that go beyond mere performance comparisons. For the case of spatial orientation movement, they conclude males have a tendency to make use of environmental reference points, and this enables greater success on questions requiring two-stage orientation movements. Such nuanced findings can help teachers and educational designers to understand and so better develop specific aspects of mathematical thinking in classrooms.

In this editors' introduction we have loosely categorised the articles around what might be termed construct validity and consequential validity. This categorisation is based on the submissions received and was not an *a priori* decision. As discussed at the start of this guest editorial, a key motivation for producing this special issue was our awareness of two types of published contributions to research into summative mathematics assessment. Perhaps most familiar to readers of RME is the study and design of high-stakes assessments by mathematics education researchers. Well known recent examples include Andrew Noyes' work on mathematical pathways, which included a substantive assessment component (e.g. Drake, Wake & Noyes, 2012), and Candia Morgan's work on the evolution of GCSE test questions (e.g. Morgan & Sfard, 2016). Perhaps less familiar to some readers is work by psychometricians into the validity and reliability of mathematics tests (e.g. Newton, 1996; Crisp & Johnson, 2007). We are pleased to present papers from both the mathematics education and the psychometric research communities in this special issue, and hope it may lead to productive collaborations in the future.

References

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 264–294). London: QCA.

Crisp, V., & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal, 33*, 943-961.

Drake, P., Wake, G. & Noyes, A. (2012). Assessing 'functionality' in school mathematics examinations: what does being human have to do with it? *Research in Mathematics Education, 14*, 237-252.

Fukawa-Connelly, T. (2005). Thoughts on learning advanced mathematics. For the Learning of Mathematics, 25, 33-35.

Leron, U. (1983). Structuring mathematical proofs. American Mathematical Monthly, 90(3), 174-184.

Lithner, J. (2008). A research framework for creative and imitative reasoning. Educational Studies in Mathematics, 67, 255-276.

Morgan, C., & Sfard, A. (2016). Investigating changes in high-stakes mathematics examinations: a discursive approach. *Research in Mathematics Education, 18*, 92-119.

Newton, P. E. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal, 22*, 405-420.

Newton, P., & Shaw, S. (2014). *Validity in Educational and Psychological Assessment*. SAGE.

Rowland, T. (2001). Generic proofs in number theory. In S. Campbell & R. Zazkis (Eds.), Learning and teaching number theory: Research in cognition and instruction (pp. 157–184). Westport: Ablex.

Scouller, K. 1998. 'The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. Higher Education 35: 453–72.

Suto, W. I., & Nádas, R. (2009). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education, 24*, 335-377.

Swan, M., & Pead, D. (2008). *Bowland Maths Professional Development Resources*. Bowland Trust/Department for Children, Schools and Families.

van den Heuvel-Panhuizen, M. (1996). *Assessment and Realistic Mathematics Education* (Vol. 19). Utrecht: Utrecht University.

Weber, K. (2012). Mathematicians' perspectives on their pedagogical practice with respect to proof. International Journal of Mathematics Education in Science and Technology, 43(4), 463-482.

White, J. (2014). Who needs examinations? A story of climbing ladders and dodging snakes. Institute of Education Press.