MATHEMATICS ASSESSMENT: EVERYTHING OLD IS NEW AGAIN?

THE ANNUAL CLEMENTS / FOYSTER ORATION

ROSEMARY CALLINGHAM

University of Tasmania

Rosemary.Callingham@utas.edu.au

Over the past decade or so there has been much rhetoric about assessment. There are assessment websites replete with "rich tasks", work samples, standards, and definitions. The MySchool website reports data from large scale assessments. Teachers are exhorted to use assessment as a tool for learning. What has all of this activity achieved? Research evidence is scant and conflicting. It is time to assess mathematics assessment and to reconsider the purpose, nature and use of assessment information.

This is the first Clements-Foyster lecture to be delivered to a combined audience of practitioner and academic researchers. When MERGA was established in 1976 by John Foyster and Ken Clements, the AAMT already existed. At that time it had a research committee, which suggests that mathematics teachers recognised the importance of research. With the growth of MERGA, the AAMT research committee ceased to exist but over the years the two organisations have developed a strong mutual respect and have worked together productively to address a range of issues in mathematics education. With the introduction of the Australian curriculum, assessment of mathematics is a re-emerging focus and the topic of this address.

I aim first to briefly outline the history of assessment with a focus on mathematical knowledge. I will then examine a number of influential developments in more recent times, and consider current practices, before proposing some new ways of thinking about the purpose, nature and use of mathematics assessment information.

Assessing mathematics

Assessment of mathematical understanding goes back to ancient times. The traditional owners of the land we call Australia had a complex mathematics to describe kinship groups, arrangements for sharing food and other resources, and for navigation and describing position. This mathematical knowledge was passed to the youth of the group in a variety of ways: modelling, practice, direct instruction and story-telling. How was this assessed? Some of the knowledge would not have been formally assessed. Some may have been part of secret initiation ceremonies and some may well have been assessed in a public display of knowledge (Peterson, 2008). The key point is that the

"teachers" were also the assessors and getting the assessment right was fundamental to the survival of the group—very high stakes assessment.

Moving on to the ancient Greeks, Pythagoras is an important historical figure. In the Pythagorean brotherhood, whole numbers had religious status and formed the basis of secret rites. When one of their members, Hippasus, made the shocking discovery that the diagonal of a unit square could not be expressed as a whole number ratio, legend has it that he was drowned by the brotherhood members. This unhappy outcome was a consequence of challenging the teacher's assessment and knowledge base.

Imperial China used a complex system of examinations for admittance to the public service, the earliest system of standardised tests. Examinations took place at designated centres, and candidates were literally locked in for up to a week. Examinations were written, and all responses were copied by a scribe prior to assessment to prevent identification of the candidate. These examinations were very high stakes: success would guarantee a comfortable life not only for the examinee but for family and village as well. One of the "Six Arts" examined was mathematics, both applied, as in taxation, and pure problems being given. Successful candidates were "called to the bar" which separated them from the unsuccessful—similar language is used by lawyers to this day. Assessing mathematical knowledge has been an important element of education from the earliest days.

More recent developments

As schooling developed and became more formal, so did mathematics assessment processes. Teachers remained the principal assessors. Oral questioning of students, sometimes in public, was a recognised and respected approach to assessing students' knowledge for the purposes of determining attainment, and this tradition is continued in the oral defence of PhD theses. Such oral examinations "… allowed teachers to ask probing questions or even to help pupils by providing permissible hints" (Lewy, 1996, p. 225).

As educational opportunities expanded, formal examinations became more widespread in the west. Printed examinations were first used at Harrow School, one of the great public schools of England, in 1830. During the twentieth century the assessment emphasis moved to standardised tests and objective measurement that focussed on aspects such as identical conditions of testing, and statistical measures such as those relating to reliability. External tests at key points in schooling became widespread in some western countries, although not all. Bodin (1993), for example, described the French system, where students did not automatically move upwards from year to year, as one where assessment was carried out continuously by the teacher who did not have to account to anyone. The teacher awarded marks, calculated averages and these were assumed to be a measure of the achievement of the learner. Examinations were unknown.

The rise in external examinations, often presented at key points in schooling, in effect, separated testing and test development from the process of teaching (Grouws & Meier, 1992; Lewy, 1996). This separation was not unnoticed. Dennis (1926), for example, wrote

> In mathematics … the methods of testing have a strong effect upon the teaching. … For years we have been discussing and revising the teaching of mathematics, its aims, its

> curriculum, the materials to be used, and the methods to be employed. But we have not given equal attention to the ways of testing the results (p. 58).

Today, 85 years later, the same comments could be made about the new Australian Curriculum.

In Britain, by the early 1980s testing was widespread with over three-quarters of the responsible authorities using some form of testing, of which mathematics was usually a part. Much of this testing was driven by debate about standards of education (Gipps, 1988). The concerns about standards were not new. Early in the 20th century high failure rates in tests were accepted as a way of maintaining standards – only the brightest and best survived the process. As pressure grew for a better educated workforce, compulsory schooling was extended, and it became the norm for children to move through the years of schooling with their age peers. New arguments for testing developed, based on equity, but still rooted in standards (Resnick, 1980). Tests were used to set standards and test results were assumed to be a measure of the success of the system. Large-scale testing programs were used as part of a "carrot-and-stick" approach to improving the quality of education and teachers were expected to change their practice in response to this external pressure to raise standards of education (Darling-Hammond, 1990). The question of the use of tests not only to describe standards but also to raise them continues today.

In 1998, Black and Wiliam's seminal work changed the face of assessment. Their meta-analysis of a variety of research studies led to a series of influential publications about the use of feedback in classroom assessment (Black & Wiliam, 1998a, 1998b). Again the emphasis was on raising standards but this time through improving the classroom assessment process. Hattie (2009) reinforced the effectiveness of feedback, and reasserted the importance of teachers. When assessment and teaching are seamless, useful feedback is provided to students, and both teacher and students change what they do as a result, classroom assessment is a powerful tool.

Towards the end of the twentieth century, there were calls to build closer links between teaching, learning and assessment (e.g., Pellegrino, Chudowsky & Glaser, 2001; Shepard, 2000), and to involve teachers more closely in the assessment process. There was an expectation that teachers would not only test knowledge recall, but instead would use complex tasks intended to provide an intellectual challenge (Lewy, 1996). One approach to this matter was termed "authentic" assessment (Archibald & Newmann, 1988). Authentic assessment aimed to provide assessment tasks for students that were meaningful outside the school context, and which expected students to communicate their ideas through coherent writing, rather than through multiple choice responses. These ideas underpinned Queensland's Rich Tasks as part of the New Basics project (Education Queensland, 2004), although there were also other theoretical considerations around intellectual quality.

During the late 1980s and early 1990s, in various parts of the world attempts were made to return assessment to the classroom. In Britain, common assessment tasks were used at Key Stages in education. California had a large-scale program of teacher-judged assessment, as did Ontario in Canada. In Australia, the idea of a student "profile" took hold and this was seen as one approach to improved accountability in which teachers played a major role, culminating in the publication of *Mathematics: A curriculum profile for Australian schools* (Curriculum Corporation, 1994). These attempts to

develop large-scale teacher-judged assessment processes failed on two counts. The first was political. Authorities did not believe the evidence that teacher judgement was as reliable as a multiple choice test. The second was industrial. Teachers refused to accept the additional responsibility and workload. Perhaps this was an opportunity lost.

The situation today

Today, the situation in Australia is bewildering. NAPLAN provides an external measure of numeracy but teachers are urged to use formative assessment. External testing has become high stakes, with schools compared to other like schools using widely available, complex statistical information on the MySchool website. At the same time, systems advocate use of assessment *for* learning or assessment *as* learning and provide examples of open tasks, rubrics, descriptions of expected standards and many other resources aiming to lift the quality of teaching. Wiliam and Black (1996) used the ideas of meaning and consequences as one way of distinguishing formative and summative assessment. Formative assessment, they suggested, is characterised by some action as a result—the consequences of the assessment—whereas summative assessment has a focus on maintaining the same meanings across individuals and groups, as well as across time.

Despite the stress on assessment for learning, the emphasis on feedback and the plethora of advice to teachers, and external testing, there is little evidence that overall this activity has created improvement in students' learning outcomes (Stiggins, 2007). Over a twenty-year period mathematics performance on statewide tests in Tasmania remained stable, although the tests themselves became harder, leading to a perception of falling standards (Griffin & Callingham, 2006). Initial comparisons on NAPLAN numeracy from 2008 to 2010 do not indicate any significant change across time for any grade group (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2010). Burgess, Wilson and Worth (2010), writing from an economics perspective, claimed that "league tables" reporting assessment results for schools in England led to improved performance in contrast to Wales where league tables are not used. They quoted effect sizes of around 0.2 which is below Hattie's (2009) suggestion that an effect size of 0.3 represents what would happen through a process of natural development. Internationally, Australia has slipped somewhat in rankings, and in PISA 2009 its performance also declined significantly. In addition, a significant difference between male and female performance first seen in 2006 was confirmed, suggesting that gender equity issues are still of importance (Thomson, de Bortoli, Nicholas, Hillman, & Buckley, 2011). This decline happened despite the increased emphasis on statewide testing that grew throughout the 1990s and became NAPLAN in 2008. The evidence about improved performance from competitive assessment results is limited.

The situation in mathematics assessment in Australia today is thus somewhat confused. All states and territories undertake NAPLAN and these results are used for accountability at the local level. Australia participates in various international studies which are used as measures of the success of government policies. At the same time, teachers are bombarded with advice and resources about formative classroom assessment. There is an emphasis on giving feedback, improving teaching and providing detailed information to parents. Media and systems decry falling standards in numeracy, and parents are advised to consider assessment outcomes reported through MySchool

when they choose a school for their child. In summary, teachers and schools are getting mixed messages about assessment. On the one hand they are urged to bring assessment closer to teaching, on the other the assessment that counts is externally imposed testing. Confusion reigns.

It seems that the education community has not clearly communicated to those outside it what assessment is about, and what inferences can be validly drawn from the information presented. In part this is an issue of numeracy—it is, after all, a data interpretation exercise. There are also, however, issues around the use of assessment information that have remained unquestioned. Messick (1989) coined the term "consequential validity" to describe the use of assessment information. He stated

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. (p. 13, emphases in original).

The question that always needs to be asked is "does this assessment provide suitable information on which to base future actions" about whatever claim is being made, whether that claim is about an individual student, a school or a system.

Productive assessment

Assessment is arguably the most powerful element in teaching and learning. Quality assessment can provide information to students, teachers, parents and systems in effective and useful ways. To be helpful, however, it must be broad ranging, collecting a variety of information using a range of tasks before, during and after a teaching sequence. At present there is a lack of consistency—in terminology, in approach and in use of assessment information.

One resource that may provide some direction is the AAMT position paper on the "Practice of Assessing Mathematics" (Australian Association of Mathematics Teachers [AAMT], 2008). Taking account of both classroom and external assessment, this document clearly makes the call that

> Students' learning of mathematics should be assessed in ways that:
> - are appropriate;
> - are fair and inclusive; and
> - inform learning and action (p.1).

This statement is consistent with Messick's (1989) view of validity, and also recognises the reality of modern education. Large-scale external testing is here to stay, but does not have to have a negative impact on learning if it is used appropriately.

Assessment that provides useful, timely and appropriate information in fair and equitable ways is productive assessment. It may address the mathematical understanding of a child, the achievement of a class or the performance of a system, and can take place at any point in the learning and teaching cycle. Productive assessment includes productive tasks, productive dialogue, productive teaching practices and productive reporting. To illustrate these points, some examples of productive assessment are described here.

There are numerous wonderful tasks that promote and develop mathematical understanding in children. The key to making these tasks productive is to trust the students and allow them some freedom and control over what they choose to do. For

example, a Year 7 class started exploring the Task Centre activity called "Sphinx" (Martin, 2000). They became very engaged with the problem and asked the teacher whether they could make a video about their investigation. Ultimately, the class produced a video that showed their learning about geometry, algebra, problem solving and many other incidental aspects of mathematics. This was an unintended assessment activity but one that produced very rich results for all concerned.

Productive dialogue can be any discussion that improves understanding. Take this example from a Year 8 classroom in a disadvantaged school during a learning sequence addressing 2D and 3D shapes:

> Student: We live on a circle, don't we?
> Teacher: Are you sure? If we cut the earth in half we'd see a circle… Do we live on a circle?
> Student: Hang on, no, it's a [long pause] It's a cubic circle.

The student successfully demonstrated his understanding of the difference between a circle and a sphere without having the technical language to describe this. The teacher was able to build on this understanding and to develop the appropriate mathematical language—a productive episode for both parties.

Quality mathematics teachers can turn almost any activity into a productive teaching event. A Year 1 teacher decided to use her students' birthdays as a starting point for what she intended to be a unit on time addressing the months of the year, and so on. When trying to sequence the birthdays in the class by hanging cards on a line, the children were very insistent that the sequence should begin in the current month, rather than January which the teacher had anticipated. The teacher decided to throw the challenge to the class to represent the birthdays in ways that could be understood by other people. The representations produced gave some deep, and surprising, insights into the children's understanding of data representation.

Productive reporting can be at any level. This scenario was observed in a Tasmanian primary school (Callingham, 2010).

> The teachers are meeting in grade teams. They are sharing the "big books" about mathematics that the children in their class have produced. The discussion centres on what the books demonstrate about the children's understanding, and what the teachers need to do to move that forward. In the discussion, teachers compare the work samples and make judgements about their own and other teachers' students. They refer frequently to the state curriculum documents, NAPLAN results, the school policies and "throughlines" that have been developed collaboratively to ensure a common language and focus across the school. By the end of the meeting, all teachers have a commitment to some action for their class, and to increase the school focus on specific aspects of mathematics at which the students appeared to do less well on the NAPLAN.

The teachers were reporting to each other, using data from various sources and committing to action as a result.

Teachers make a difference. They assess continuously in a variety of ways. It is time for a return to the traditions of assessment practice by recognising teachers' authority in the [new] practice of mathematics assessment.

References

Archibald, D. & Newmann, F. (1988). *Beyond standardized testing: Authentic academic achievement in the secondary school*. Reston, VA: NASSP Publications.

Australian Association of Mathematics Teachers (2008). *Position paper on the practice of assessing mathematics.* Adelaide: Author.

Australian Curriculum, Assessment and Reporting Authority [ACARA] (2010). *NAPLAN 2010 summary report*. Accessed 23 May 2011 from http://www.naplan.edu.au/verve/_resources/NAPLAN_2010_Summary_Report.pdf

Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, March, 7-74.

Black, P. & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan* [Online article]. Retrieved December 17 2002 from http://www.pdkintl.org/kappan/kbla9810.htm

Bodin, A. (1993). What does to assess mean? The case of assessing mathematical knowledge. In M. Niss (Ed.), *Investigations into assessment in mathematics education* (pp. 113-141). Dordrecht, "The Netherlands": Kluwer Academic Publishers

Burgess, S., Wilson, D. & Worth, J. (2010). *A natural experiment in school accountability: the impact of school performance information on pupil progress and sorting. (Centre for Market and Public Organisation, working paper 10/246.)* Bristol, UK: CMPO.

Callingham, R. (2010). Mathematics assessment in primary classrooms: Making it count. In C. Glascodine & K-A. Hoad (Eds.) *Teaching mathematics? Make it count. What research tells us about effective mathematics teaching and learning.* (Proceedings of the annual research conference of the Australian Council for Educational Research, pp. 39-42). Melbourne: ACER.

Curriculum Corporation (1994). *Mathematics: A curriculum profile for Australian schools*. Melbourne: Author.

Darling-Hammond, L. (1990). Achieving our goals: Superficial or structural reforms? *Phi Delta Kappan, 72*, 286-295.

Dennis, J. (1926). Chapter iv. Mathematics. In Institute of Inspectors, N.S.W. *Teaching and testing*. Sydney: Geo. B. Philip & Son.

Education Queensland (2004). *The New Basics research report*. Brisbane: Author.

Gipps, C. (1988). The debate over standards and the uses of testing. *British Journal of Educational Studies, 26*(1),104-118.

Griffin, P. & Callingham, R. (2006). A twenty-year study of mathematics achievement. *Journal for Research in Mathematics Education, 37*(3), 167-186.

Grouws, D.A., & Meier, S.L. (1992). Teaching and assessment relationships in mathematics instruction. In G. Leder (Ed.) *Assessment and learning of mathematics*. (pp. 83-107). Melbourne: Australian Council for Educational Research.

Hattie, J.A.C. (2009). *Visible learning: a synthesis of meta-analyses relating to achievement*. Abingdon: Routledge.

Lewy, A. (1996). Postmodernism in the field of achievement testing. *Studies in Educational Evaluation 22*(3), 223-44.

Martin, A. (2000). The sphinx task centre problem. *Mathematics in School, 29*(3), 6-9.

Messick, S. (1989). Validity. In R. Linn (Ed.). *Educational measurement*. (3rd ed., pp. 13 – 103). New York: American Council on Education and Macmillan Publishing Company.

Pellegrino, J.W., Chudowsky, N. & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Peterson, N. (2008). Just humming: the consequence of the decline of learning contexts among the Walpiri. In J. Kommers & E. Venbrux (eds.), *Cultural Styles of Knowledge Transmission: Essays in Honour of Ad Borsboom* (pp. 114-118). Amsterdam: Aksant Academic Publishers.

Resnick, D. P. (1980). Minimum competency testing historically considered. *Review of Research in Education, 8*, 3-29.

Shepard, L.A. (2000). *The role of classroom assessment in teaching and learning. CSE Technical Report 517*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Stiggins, R. (2007). Assessment through the student's eyes. *Educational Leadership, 64*(8), 22-26.

Thomson, S., de Bortoli, L., Nicholas, M., Hillman, K., & Buckley, S. (2011). *Challenges for Australian education. Results from PISA 2009.* Camberwell, VIC: Australian Council for Educational Research.

Wiliam, D. & Black, P. (1996). Meanings and Consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, *22*(5), 537-548.