The Terminology of Mathematics Assessment

Jane Greenlees
Charles Sturt University
<jgreenlees@csu.edu.au>

Standardised testing has received a lot of political and public attention recently in Australia. This paper describes the sense-making of Year 3 students as they interpret items from the 2008 NAPLAN. Results show that student performance changed dramatically when the terminology of an item was modified and subsequently were not a true indication of student mathematical knowledge and understanding. Implications include the need for test designers to carefully consider the terminology included within assessment items and the need for comprehensive analysis of student results.

The introduction of the 2008 National Assessment Program – Literacy and Numeracy (NAPLAN) - across schools in all states and territories heralds a new era in Australian education. Just like reforms and policy changes before, the NAPLAN was deemed necessary to "better the nation's competitive edge" (Webb, 1992, p.661) and hold teachers and schools accountable for student results. Therefore with such high stakes involved it is important that test results are a reliable and credible representation of student's knowledge and understanding. But "how well do current standardised mathematics tests reflect the extent and nature of mathematical knowledge and ability that students have?" (Kulm, 1991, p. 72).

Although national standardised testing is new to the Australian education system, the concept of mandatory numeracy tests is not. Nevertheless, what has incrementally changed in the past 10 years is what numeracy is being assessed and particularly the nature and composition of the assessment tasks (Lowrie & Diezmann, 2009). Additionally the idea of making mathematics 'real' and relevant by incorporating 'everyday' contexts has been a growing trend in schools in the past 30 years (Boaler, 1994). The nationally agreed Statements of Learning (NSL) in mathematics outline in its Year 3 Working Mathematically that students will "actively investigate everyday situations as they identify and explore mathematics" (MCEETYA, 2006, p. 5). It is believed that such an approach would help students realise the relevance of maths as it is applied to their world outside the classroom.

As a result, test designers are attempting to make questions more realistic and possibly authentic but whether this is problematic is yet to be seen. Test items therefore have moved beyond simple word problems and algorithms.

The Four Components of Assessment Items

There are four components of assessment items that need to be implicitly taught within the classroom for student success. These include mathematical content, literacy demand/terminology, contextual understanding and graphics (see Figure 1).

Mathematical content can be defined as the core elements children are taught throughout their school career as outlined in state and territory curriculums. The role of assessment therefore is to examine these mathematical understandings and concepts. However research has found that often other components of a test item, resulting from an attempt to make them more realistic, confound these understandings, thus affecting a child's performance (see, for example, Abedi & Lord, 2001; Boaler, 1994; Logan &

L. Sparrow, B. Kissane, & C. Hurst (Eds.), *Shaping the future of mathematics education: Proceedings of the 33rd annual conference of the Mathematics Education Research Group of Australasia.* Fremantle: MERGA.

Greenlees, 2008). In such situations, students tend to use prior knowledge and understanding of general contexts and previous experience to shape their decision making rather than specifically focusing on the task at hand. Consequently, too much misleading information can affect performance (Logan & Greenlees, 2008).

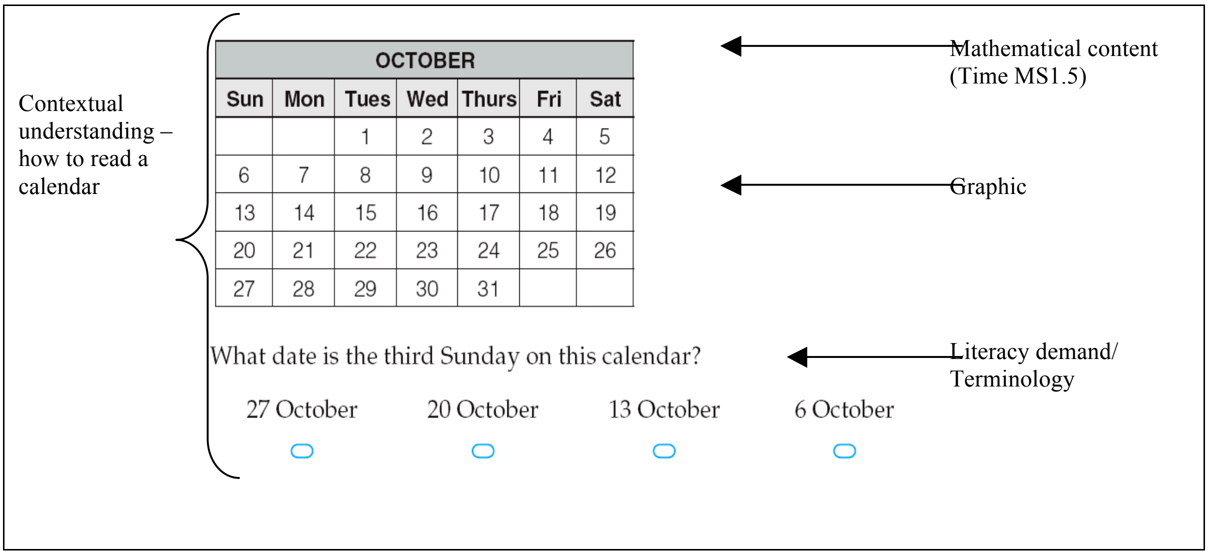

Figure 1 The four components of mathematics assessment items
(MYCEETYA 2008a: Year 3 Numeracy Test, Item 12)

In Figure 1 the mathematical content being assessed according to the NSW Board of Studies K-6 Mathematics Syllabus (2002) is outcome MS1.5 – compares the duration of events using informal methods and reads clocks on the half hour. As such students should be able to identify the day and date on a calendar. However in order to do this a child must first decode the graphic according to calendar conventions, understand the context of the use of a calendar as well as comprehend specific terminology associated with the question. As a result, terminology, which is intended to be related to the task gets interpreted within a broader context. In this investigation it is argued that it is difficult to separate the context that surrounds the question from the terminology. Subsequently for an assessment item to be accessible to all students these four elements need to be valid and relevant to the mathematical construct being measured, that is, how to read a calendar. However research has found that often assessment outcomes are "confounded with nuisance variables that are unrelated to the construct" (Abedi, 2006, p. 377) thus threatening the validity of the assessment, in particular the use of unnecessary and unfamiliar terminology.

Background

Mathematics is often associated with numbers and symbols. In fact many people's mathematical experiences and memories would include the stereotypical times tables in primary school and later on, algebraic expressions and formulae. However the shift towards making mathematics relevant has seen an increase in the literary demand placed on assessment tasks. As Thomas (1988) points out these demands involve both technical terminology and ordinary language. Teachers now have an obligation to "provide opportunities for students to strengthen their understanding of mathematics terminology and concepts" (Adams, 2003, p. 789). In fact specific mathematical terminology was explicitly defined in the NSW Department of School Education K-6 Mathematics Syllabus

(1989) so that teachers could intentionally refer to them as part of the mathematical content.

While there has been an extensive body of literature which address language in mathematics (Adams, 2003; Fuentes, 1988; Perso, 2009; Pugalee, 1999; Wakefield, 2000), Matteson (2006) notes few studies have focused on connections between mathematical literacy and achievement on mathematical assessments. Yet while teachers have some control over the mathematical terminology used in their classroom they have no influence on the unnecessary and unfamiliar linguistic structures used in an assessment task. According to Abedi (2006) it is these language barriers that can "threaten the validity and reliability of content-based assessments" (p. 380). Abedi & Lord (2001) found that minor changes in the wording of test items resulted in significant differences in mathematics performance. For example, "rewording a verbal problem can make semantic relations more explicit, without affecting the underlying semantic and content structure; thus, the reader is more likely to construct a proper problem representation and solve the problem correctly" (Abedi, 2006, p. 380). Abedi & Lord (2001) found that scores on linguistically modified mathematics tests were slightly higher than the original version. This highlights the serious impact unnecessary terminology may have on student performance. The purpose of this paper is to explore and scrutinize the terminology used in test items from the 2008 Year 3 Mathematics NAPLAN, in order to provide informed comment on the interpretation of student results.

Research Design and Methods

This investigation is the beginning of a three-year study that aims to explore the way in which test items are constructed and how this impacts on a student's capacity to make sense of mathematics. This study will include exploring the relationship between mathematical content, mathematical terminology (mathematical literacy), graphical representations and contextual understanding. The aims of this initial component of the study were to:

1. Analyse student responses and sense-making on standardised test items through a mixed method research design.
2. Examine the effect of modified items on student performance.
3. Identify important components of a test item that positively or negatively influence the validity of student results.

The Participants

170 Year 3 students (aged 8-9 years) from 4 Catholic NSW schools participated in the quantitative phase of the study. The qualitative dimension included 40 students (10 from each school) randomly selected from the original cohort. All participants were from varying socioeconomic and academic backgrounds and participation was strictly voluntary with the available option to discontinue at any time.

Data Collection and Analysis

The following section describes the three phases of the project.

Phase 1. The initial interview. The 40 randomly selected students were interviewed on their thinking processes and strategies used when solving the 2008 NAPLAN (Test A).

Phase 2. The modification. From the interview data, students' responses were analysed and collated to ascertain the problem-solving processes used to solve respective items. The analysis revealed that often a correct answer was given yet an incorrect strategy was used. This highlighted an obvious misconception of what the child actually knew and what could be considered an educated guess. It was therefore assumed that by modifying the question slightly it would verify and reveal a true understanding or not. Further similarities between student's interpretations became obvious and the impact certain elements of the item including the graphic and the mathematical terminology had on student success. As a result these items were slightly redesigned without changing the complexity of the question and Test B was created. Test A and Test B were then given to the larger cohort of 170 students in random order. For example in one school Test B was administered first and then followed by Test A, while in another school Test A was before Test B. These two tests were carried out on the same day with the exception of one school where there was a two-day reprieve.

Phase 3. The re-interview. Following the large scale testing, the original 40 students were re-interviewed based on their responses to Test B. Once again these structured, in-depth interviews allowed students the opportunity to verbalise and justify the mathematical processes they used.

Results

This paper focuses on the two items that had the largest effect sizes from Test A (NAPLAN) to Test B (modified test) when the terminology was modified. Thus, this study was concerned with items where student performance (in relation to correctness) changed the most. Table 1 highlights these results.

Table 1
Students' Performances across Test A and Test B

	Item 2		Item 15	
	A	B	A	B
% Correct	95	87	44	95
Effect size (Cohen's *d*)	.31		-1.34	

Largest versus smallest

Item 2 of Test A (see Figure 2) could be considered quite easy for many of the students with no children getting it wrong in the interview and 95% correct in the mass testing.

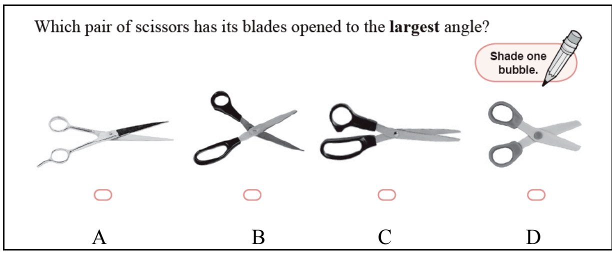

Figure 2 Test A Largest versus smallest
(MYCEETYA 2008a: Year 3 Numeracy Item 9)

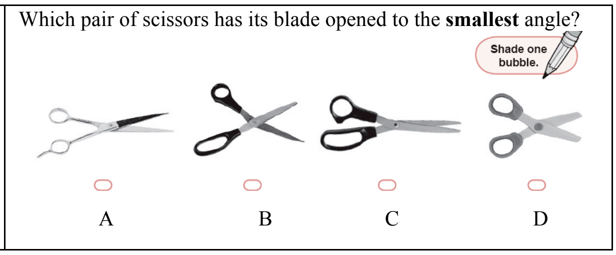

Figure 3 Test B Largest versus smallest
(MYCEETYA 2008a: Year 3 Numeracy Item 9)

> I chose B because that's pretty small (points to D) and that one's smaller (points to A) and that one's the smallest (points to C) [HT4].

It could therefore be assumed that most of the children had a competent understanding of the mathematical content, that is, SGS2.2.b Identifies, compares and describes angles in practical situations (Board of Studies NSW). However when students were asked to find the smallest angle instead of the largest, only 90% got it correct in the interview and 87% correct in the larger cohort.

> I chose B because it is bigger than all the other acute angles. C is the smallest acute and A is the second, D is the third and B is the largest [HT4].

Now when asked to find the smallest angle (C) students still chose the largest (B). While many students were able to justify their mathematical reasoning for choosing B it was often confusing and complicated. Test B results now indicated this was an area of concern and raises questions of teacher competency. The reality is that changing the terminology, not the mathematical content, impacted negatively on student results. It could be argued that performance differed due to familiarity of the item and an automated response from the students as they failed to notice the change in wording from Test A to Test B. However, students involved in the interviews had the opportunity to read the question out loud, drawing attention to the terminology change, and still answered incorrectly to a question in which they originally had shown a competency. Thus, the likelihood of an automatic response was reduced.

Less versus Fewer

It was evident that students found it difficult interpreting some of the terminology in Test A, in particular the word "fewer" in Item 15 (see Figure 4).

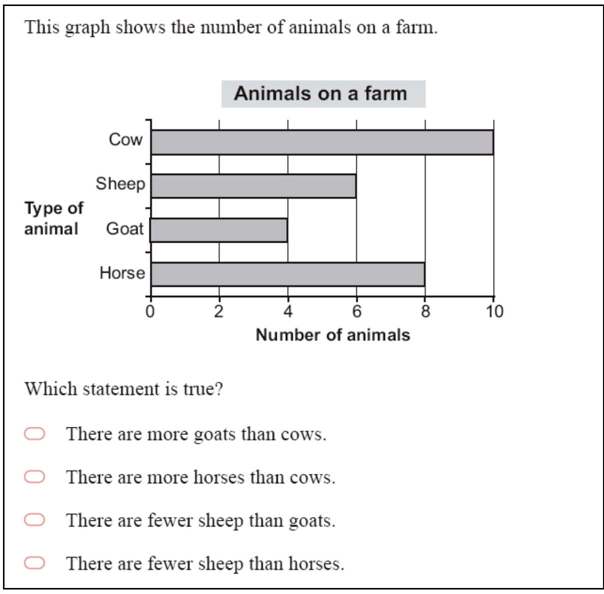

Figure 4 Test A Fewer versus less (MYCEETYA 2008a: Year 3 Numeracy Item 29)

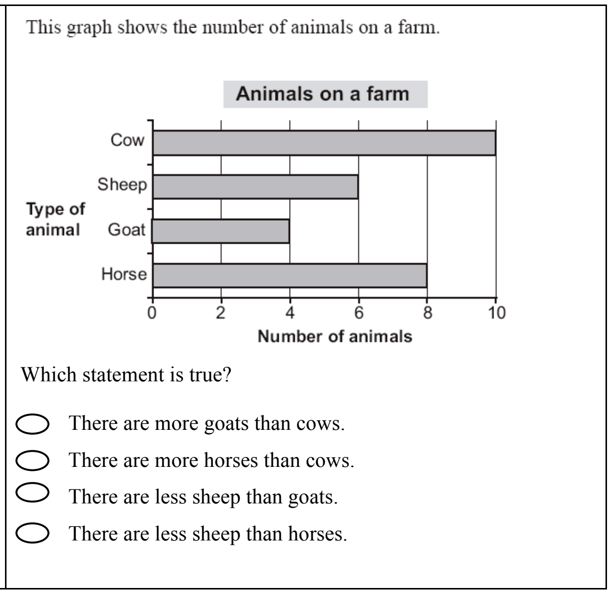

Figure 5 Test B Fewer versus less (MYCEETYA 2008a: Year 3 Numeracy Item 29)

In fact almost half the interview cohort (48%) and over 56% of all students answered this question incorrectly, choosing answer C. When questioned on how they drew their conclusions nearly all students could successfully read the graph but simply did not understand the terminology. For example:

> Because it shows on the graph that there's more sheep than goats and this one says that there's fewer sheep than goats and that's what it shows on the graph [SJ1].

> I looked A and it wasn't right. Looked at B didn't look right. I looked at C and it looked right and then I looked at D and it didn't look right so I picked C and coloured that in. (So there are fewer sheep than goats. How many sheep are there?) There are 6. (And how many goats are there?) 4. (What's another way of saying that?) There are more sheep than goats [HT5].

It was for this reason that the word 'fewer' was replaced with 'less' in Test B (see Figure 5). According to Quirk & Greenbaum (1993) when making a comparison between quantities there is a choice between these two words, however 'less' is definitely used when referring to statistical or numerical expressions. Subsequently only 5% of all students answered this question incorrectly in Test B.

> I looked are there more goats than cows and no because there are only 4 goats and there are a maximum of cows. And I looked at there are more horses than cows and that is not true. There are less sheep than goats and that's not true. And then I looked at there are less sheep than horses and I could see that answer had to be because the horses had 8 and the sheep had 6 and then I coloured answer D [HT5].

With the growing emphasis of standardised testing within the education system it is important that these assessments provide a valid picture of what students know and can do. If we are to read the results of Test A, with no insight into children's mathematical thinking, it could be assumed that over half the students were unable to read a graph correctly. The reality is that 95% could successfully complete the mathematical component of the question but were unable to access the item due to literary restraints. As Abedi (2006) argues, "to provide fair and valid assessment for all students … the impact of terminology unrelated to content-based assessments must be controlled" (p. 377).

Conclusions and Implications

Given the increased accountability being placed on teachers (and education systems) in relation to national testing, it is imperative that specific items within tests are scrutinised (Diezmann, 2008). If teachers and schools are going to be targeted and held accountable for student's results, we need to guarantee that the assessment is a valid and accurate representation of what students know.

This paper does not suggest that children are ill prepared to engage in mathematics tasks and thinking but rather the phrasing of some of the items in NAPLAN are inappropriate for contemporary teaching and learning. Any mathematical test should be a reflection of children's mathematics performance, not a student's capacity to interpret tasks that are foreign. We therefore need to evaluate what we are actually assessing - mathematical knowledge or a child's individual terminology ability. Teachers should not be expected to teach to the test but in light of the results from this paper if we are not careful this will happen in order to guarantee positive results for their students. Furthermore we need to be assured that slight modifications in terminology do not result in dramatic student performance differences. The reliability of items within the NAPLAN need to be well scrutinised and indeed the items need to assess what is being reported, particularly in our climate of intense accountability.

References

Abedi, J. (2006). Terminology issues in item development. In S. M. Downing, & T. M. Haladyna (Eds.). *Handbook of test development*. (pp. 377-398). New Jersey: Lawrence Erlbaum Associates, Inc.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Adams, T. L. (2003). Reading mathematics: More than words can say. *The Reading Teacher, 56*(8), 786-795.

Boaler, J. (1994). When do girls prefer football to fashion? An analysis of female under achievement in relation to realistic mathematics contexts. *British Educational Research Journal, 20*(5), 551-564.

Board of Studies NSW. (2002). *Mathematics K-6 Syllabus*. Sydney, Australia: Board of Studies NSW.

Diezmann, C. (2008). Graphics and the national numeracy tests. In M. Goos, R. Brown & K. Makar (Eds.). *Navigating currents and charting directions. Proceedings of the 31st annual conference of the Mathematics Education Reseacrh Group of Australasia* (Vol. 2, pp. 659-662). Brisbane, Australia: Mathematics Education Research Group of Australasia.

Fuentes, P. (1988). Reading comprehension in mathematics. *The Clearing House, 72*(2), 81-88.

Kulm, G. (1991). New directions for mathematics assessment. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics.* (pp. 71-78). Washington, DC: American Association for the Advancement of Science.

Logan, T., & Greenlees, J. (2008). Standardised assessment in mathematics: The tale of two items. In M. Goos, R. Brown & K. Makar (Eds.). *Navigating currents and charting directions. Proceedings of the 31st annual conference of the Mathematics Education Reseacrh Group of Australasia* (Vol. 2, pp. 655-658). Brisbane, Australia: Mathematics Education Research Group of Australasia.

Lowrie, T. & Diezmann, C. M. (2009). National numeracy tests: A graphic tells a thousand words. *Australian Journal of Education, 53*(2)*,* 141-158.

Matteson, S. M. (2006). Mathematical literacy and standardised mathematical assessments. *Reading Psychology, 27*, 205-233.

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) (2006). *Statements of Learning for Mathematics.* Retrieved 2 November 2009 from: http://www.mceecdya.edu.au/verve/_resources/SOL_Maths_Copyright_update2008.pdf

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) (2008a). *National assessment program: Literacy and numeracy. Year 3 numeracy test.* Melbourne, Australia: Curriculum Corporation.

NSW Department of School Education. (1989). *Mathematics K-6.* Sydney: NSW Department of School Education.

Perso, T. (2009). Cracking the NAPLAN code: Numeracy and literacy demands. *Australian Primary Mathematics Classroom, 14*(3), 14-18.

Pugalee, D. K. (1999). Constructing a model of mathematical literacy. *The Clearing House, 73*(1), 19-22.

Quirk, R., & Greenbaum, S. (1993). A university grammar of English. Retrieved February 1, 2010, from: http://grammar.ccc.commnet.edu/grammar/adjectives.htm

Thomas, D. A. (1988). Reading and reasoning skills for mathematics problem solvers. *Journal of Reading, 32,* 244-249.

Wakefield, D. V. (2000). Maths as a second language. *The Educational Forum, 64*, 272-279.

Webb, N. L. (1992). Assessment of students' knowledge of mathematics: Steps toward a theory. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning.* (pp. 661-683). New York: Macmillan.