The past, present and future of Australian national assessment

Jane Greenlees, Charles Sturt University, Australia

Introduction

In the early 1990s there was a worldwide push towards centralized testing including the infamous 2001 No Child Left Behind Act in the United States. It was argued that this was closely aligned to an era of economic rationalism and subsequent public accountability. For Australia it included State based assessment programs which were later replaced with a national standardized test. The Australian National Assessment Program: Literacy and Numeracy (NAPLAN) for years 3, 5, 7 and 9 was officially implemented nationwide in May 2008.

Several writers have made the point previously, as noted by Klenowski and Wyatt-Smith (2012), that much of the merit of large-scale testing initiatives was in the nature of multiple choice tests that could “deliver objective measurements in which society could have confidence” (p. 68). It was believed that a test that was developed independently of schools, classrooms and teachers would be ‘uncontaminated’ and therefore capable of yielding objective measures of a student’s real achievement (Klenowski & Wyatt-Smith, 2012).

While there has been much debate about the legitimacy of standardised tests and one which I will not venture into, we cannot overlook their prevalence in schools today and over the past number of years. Essentially they are still viewed within society as the most efficient way of providing a rich analysis of what a child can and cannot do or as a means of assessing an educational system. In fact, it was estimated in the early 1990s about 30 to 35 million tests were administered annually in the United States (Pandey, 1991).

Yet what is important to note is that while the prevalence of standardised tests have remained, the entire nature of test design has changed dramatically in recent years with the incorporation of graphical and visual representations as well as the use of context and ‘real life’ scenarios (Lowrie & Diezmann, 2009). The purpose of this paper is to explore these changes in test item design and possible implications to student performance and mathematical reasoning. These issues will then be examined in light of future test item design, particularly moving towards a digital alternative.

Assessment now

In order to appreciate current test item design it is important to analyse what these constructs are attempting to measure. In the NAPLAN context, the testing bodies emphasise that it is not a test of content but rather skills in *numeracy*.

According to Connolly (2011), the most relevant Australian definition of the term numeracy is the one produced by the Australian Association of Mathematics Teachers (AAMT):

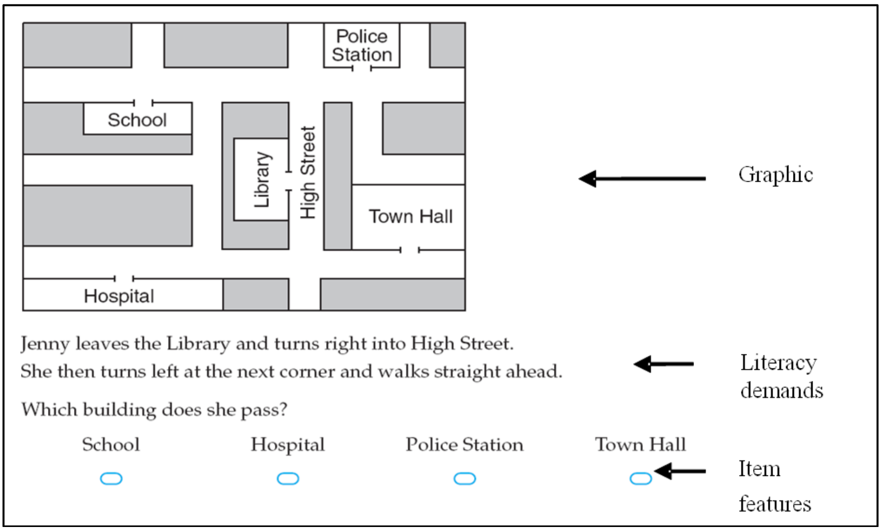
In school education, numeracy is a fundamental component of learning, discourse and critique across all areas of the curriculum. It involves the disposition to use, in context, a combination of:

- underpinning mathematical concepts and skills from across the discipline (numerical, spatial, graphical, statistical and algebraic);
- mathematical thinking and strategies;
- general thinking skills, and;
- grounded appreciation of context (AAMT, 1998, cited in Connolly, 2011, p. 910).

Connolly (2011) noted that this definition became the tool on which NAPLAN item design was based and resulted in a number of styles of items including:

- some items that address underpinning mathematical concepts;
- some items that address mathematical thinking and problem solving strategies;
- some items that include general thinking skills – including general reasoning;
- some items that are grounded in a directly meaningful context (Connolly, 2011, p. 910).

Subsequently, a direct result of this new measure has been the inclusion of items within ‘meaningful’ contexts. These include attempting to provide realistic scenarios and applications of mathematics in a real world beyond the classroom. However the result of making the assessment item more ‘realistic’ has inadvertently added several other components to the task the student must now contend with and decode. This includes the added emphasis on language, graphics and the use of specific situations. Therefore no longer are assessment items only measuring a child’s mathematical understanding but also how well they can read, utilise graphics and familiarise themselves within a particular situation. Figure 1 illustrates the role each of these plays in regards to an assessment item attempting to measure a child’s numeracy.



Jenny leaves the Library and turns right into High Street.
She then turns left at the next corner and walks straight ahead.

Which building does she pass?

School ○ Hospital ○ Police Station ○ Town Hall ○

Figure 1. The components of numeracy assessment

In deconstructing numeracy test items, Lowrie, Diezmann and Logan (2012) found that, typically, many assessment items consisted of three elements that organised mathematical information: graphics, literacy demands and item features. The graphic is defined as a diagram or picture used to convey information or contextualise a scenario. In this case the graphic is a street map. The literacy demands include text associated with the purpose of the task, the mathematical symbols and language needed to operate the tasks and the posing of the question(s). The item features contain the answer format and in this instance provide text associated with the multiple-choice format. In addition, item features involve the placement of text and graphics within the item. Research findings indicate that each of these components can impact student performance and understanding.

Graphics

Graphics have been defined in many unique and different ways. Within the context of this paper, graphics refers to any diagram, pictorial representation or graph used within an item.

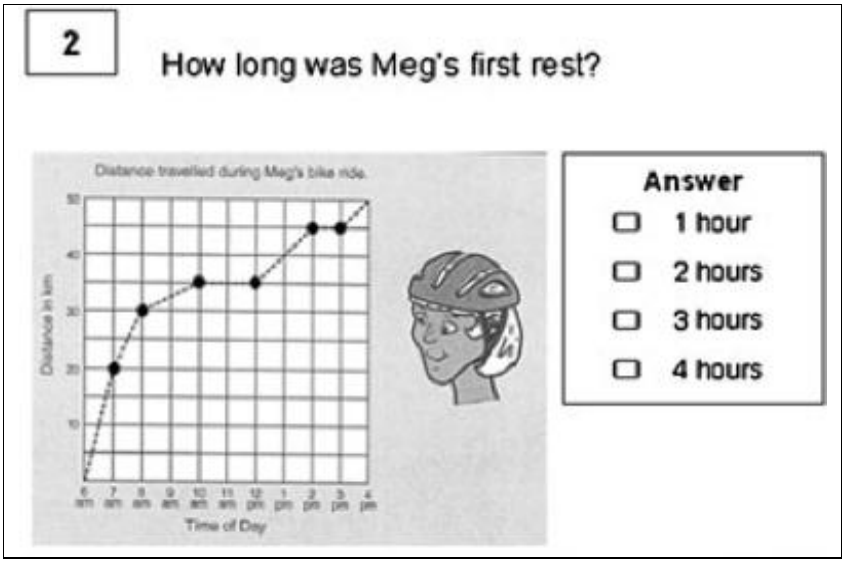
Although graphics are often considered “one of the simplest symbolic systems for interpreting information on the relationship between two or more sources” (Parmar & Signer, 2005, p. 250) primary students often find such representations overloaded with information and therefore difficult to decode (Lowrie & Diezmann, 2007). It is for this reason that the impact of using graphics in numeracy assessment tasks needs to be analysed particularly in light of student performance. Yet although there is current research surrounding the use of graphics in information texts, few studies have been conducted on the use of graphics in assessment. This is despite the fact that of the 75 tasks in the 2008 NAPLAN primary grades numeracy tests (Grades 3 & 5), 64 of these tasks (85%) contained a graphic (Lowrie & Diezmann, 2009).

Of particular interest has been the research conducted by Logan & Greenlees (2008) and further developed by Lowrie, Diezmann and Logan (2012) utilising the modification paradigm to explore the effect of graphics on students’ processing of content and their ability to decode and solve the task. Using semi-formal interviews, 40 Year 6 students (11-12 years) were given the opportunity to verbalise and justify their thinking processes when solving graphical tasks from various numeracy assessments. This included items from State based numeracy assessments and National mathematics and science competitions. These tasks could be viewed as representative of those included in the NAPLAN for numeracy or other international high-stakes testing such as Trends in International Mathematics and Science Study (TIMSS) or Programme for International Student Assessment (PISA). These items were included in Test A.

It was evident from the responses that particular elements of the task composition impacted on the way the participants interpreted and solved the items. These items were then modified according to the specific element that most influenced performance and became known as Test B. Such modifications included the addition of shaded backgrounds to graphics or rewording tasks to explicitly indicate that only one option was plausible. It was important that the modification did not change the intent of the task nor the complexity. An example of this was Task 2, also known as the Line Graph Task. Logan & Greenlees (2008) reported a substantial change in student performance and reasoning when the graphic in the item was modified.

The Line Graph Item

When students first attempted this item (Figure 2), the performance was very poor (22% correct).



2

How long was Meg's first rest?

Answer

□ 1 hour

□ 2 hours

□ 3 hours

□ 4 hours

Figure 2. The Line Graph Item
Queensland Studies Authority (2002)

It became apparent through the student's interview responses that the appearance of dots on the line graph at various intervals along the line was being interpreted as a stopping point.

> (ANSWER: 1 hour) It has 6am and 7am, so that took 1 hour until she had a rest. Because it's a line graph the circle/dot is like a rest and it tells you how long she rode [Alex].

Lowrie, Diezmann & Logan (2012) reported similar responses to the item:

> I chose 1 h because she started at 6 am and she stopped at 7 am because here it has a dot where it was a new hour [Rebecca].

It was this reason that the dots were removed in the modified task to allow the students the opportunity to read the graph without the distraction of the dots (Figure 3).

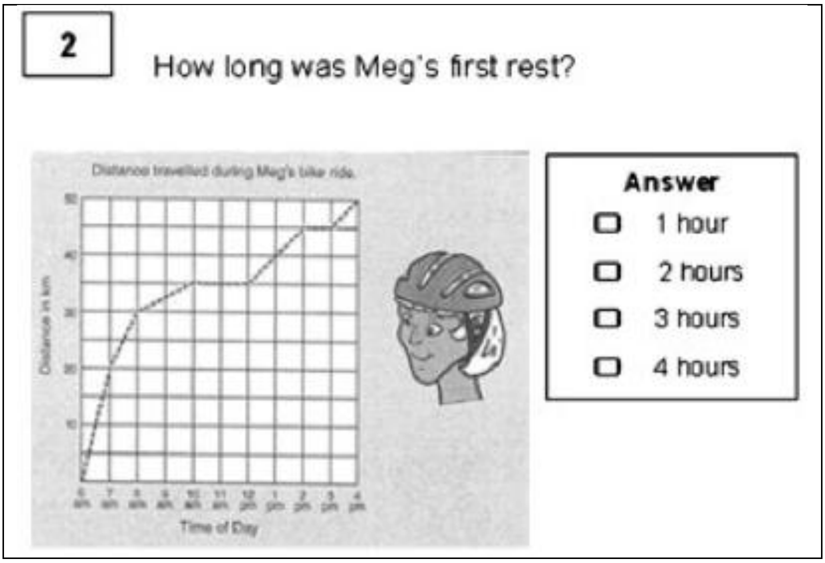


Figure 3. The modified Line Graph item

Subsequently the students' performance on the Line Graph item in Test A and the item included in Test B had a large effect size with a dramatic improvement in performance on the modified test. Table 1 highlights these results.

Table 1. Students' performance on Line Graph item across Test A and Test B

	Line graph item	
	Test A	Test B
% Correct	22	60
Effect size (Cohen's *d*)	.83	

The interview responses also supported these quantitative results with students displaying a new clarity in being able to interpret the graph. For example Rebecca now chose 2 hours as her answer with the following explanation:

> I chose two hours because on the graph it keeps on going up until she gets from 10am to 12pm and then it just goes straight so she's not moving any distance which means she must have stopped [Rebecca].

As argued by Logan & Greenlees (2008), this change in Rebecca's thinking demonstrates how visual features included in graphics, in this case a line graph, can affect students' interpretation and understanding (Gattis, 2002). It also highlights the impact the design of the graph plays in students' comprehension and reasoning processes (Carpenter & Shah, 1998). Consequently, test designers need to analyse whether the graphics included in assessment tasks accurately represent the intended information and whether they are appropriate and within the capability of the student.

Literacy Demands

According to Kiplinger, Haug and Abedi (2000), there has been significant literature and research available highlighting the impact of language on student performance in mathematics assessment. In fact, their study supported findings of Carpenter, Corbit, Kepner, Lindquist and Reys (1980) nearly 20 years earlier of a significant improvement in children's performance based on the complexity of the language used and the use of alternate numeric formats. They concluded that "unnecessarily complex linguistic structures or difficult vocabulary in a mathematics assessment introduces non-construct-related variance that can be removed by careful attention to construction of the assessment to measure the construct of math knowledge—not reading ability" (p. 15).

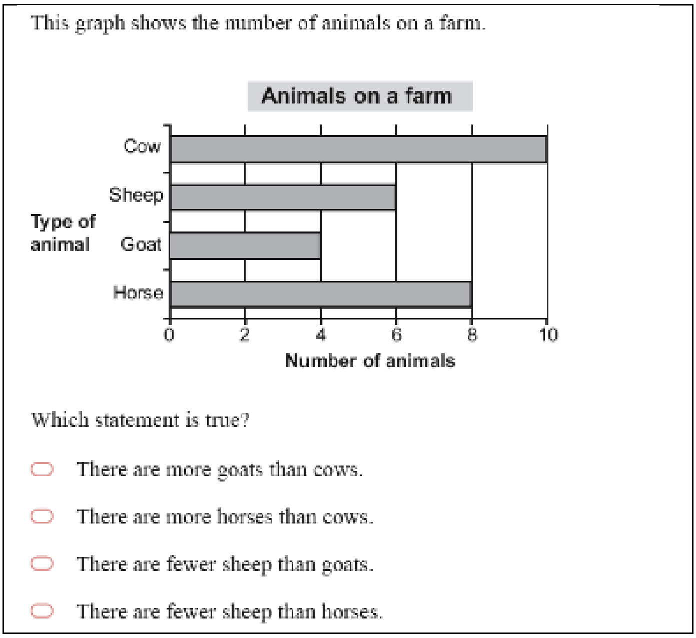
Subsequently, when Abedi and Lord (2001) linguistically modified test items to simpler versions while keeping the mathematics task and terminology the same, they achieved statistically significant improvement in students' results. This was particularly obvious for low-performing students. These findings highlighted the growing relevance and relationship between reading ability and mathematics problem-solving ability.

When investigated within the NAPLAN context, Greenlees (2010) reported comparable trends and outcomes. Using a methodology similar to Logan & Greenlees (2008) and Lowrie, Diezmann & Logan (2012), 170 Year 3 students (aged 8-9 years) were originally tested on a selected number of items taken from the 2008 Year 3 NAPLAN (Test A). Forty of these students were interviewed further to explore mathematical reasoning and possible misunderstandings. Based on these responses, items were modified according to the graphic, literacy demands or item features (Test B). An item of particular interest regarding literacy demands was the Farm Item (Figure 4).

The Farm Item

The quantitative results revealed that only 44% of the 170 students answered this item correctly. The interview results informed this data revealing that students were struggling with the definition of the word 'fewer'. 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:

> (ANSWER: C) 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].



This graph shows the number of animals on a farm.

Which statement is true?

○ There are more goats than cows.

○ There are more horses than cows.

○ There are fewer sheep than goats.

○ There are fewer sheep than horses.

Figure 4. The Farm Item
MYCEETYA 2008a: Year 3 Numeracy Item 29

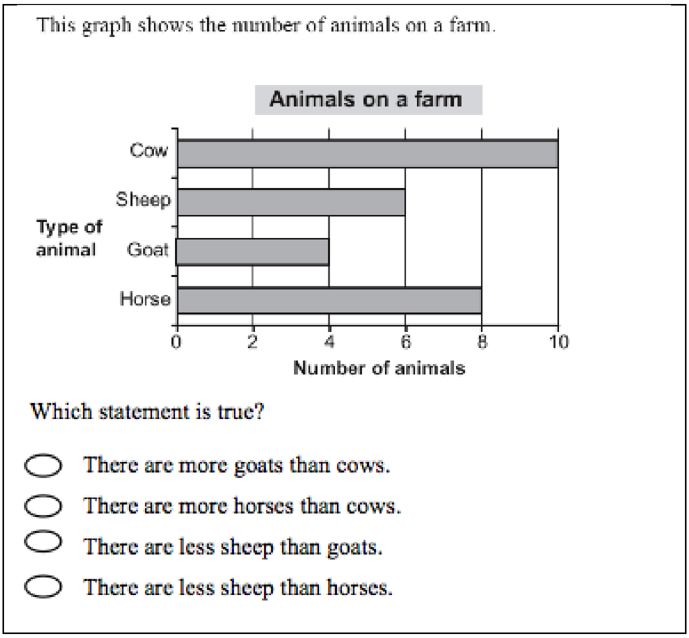
It was for this reason that the word 'fewer' was replaced with 'less' in Test B. According to Quirk & Greenbaum (1993) when making a comparison between quantities there is a choice between these two words, however 'less' is most often used when referring to statistical or numerical expressions. The replacement of this word saw a substantial improvement in students' performance with only 5% answering incorrectly. Table 2 highlights these results.

Table 2. Students' performance on Farm Item across Test A and Test B

	Line graph item	
	Test A	Test B
% Correct	44	95
Effect size (Cohen's *d*)	-1.34	

The modified Farm Item (see Figure 5) also resulted in more logical and articulate interview responses.

> I looked at there are 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].



This graph shows the number of animals on a farm.

Which statement is true?

○ There are more goats than cows.

○ There are more horses than cows.

○ There are less sheep than goats.

○ There are less sheep than horses.

Figure 5. The modified Farm Item

It could be assumed by analysing the results of this item from Test A that less than half the cohort were unable to effectively read a bar graph correctly. However the modification of some of the terminology used in the task revealed 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) argued "to provide fair and valid assessment for all students ... the impact of terminology unrelated to content-based assessments must be controlled" (p. 377).

Item features

Item writing has been described as a particular art form (Rodriguez, 2005). Within the Australian context, considerable time and effort goes into writing items to be included in the NAPLAN each year. This includes pre-testing with a sample group prior to the administration of the test, inclusion of common items across the years to improve validity and reliability and continuing research into effective test item design (Connolly, 2011). However, as noted by Rodriguez (2005), the science of item writing is still under development since the construction of multiple choice items in the early 1900s.

The effectiveness of pencil-and-paper items has been a long time debate and issues surrounding these forms of items are particularly relevant due to the current design of the NAPLAN. In Clements and Ellerton's (1996) study on the effectiveness of items similar to the NAPLAN format—multiple choice and short-answer pencil-and-paper items—they revealed a serious ineffectiveness in the items of measuring student understanding. They found that over one-third of correct answers "were given by students who did not have a sound understanding of the correct mathematical knowledge, skills, concepts and relationships which the questions were intended to cover" (p. 159). They also identified a misalignment between incorrect responses and partial understanding of the mathematics the questions were designed to assess. They concluded that if pencil-and-paper mathematics tests are being used “then it is inevitable that invalid results will be obtained” (p. 160).

For this reason there has been extensive research in the design of multiple-choice questions with a particular focus on content and formatting, however little has been reported on the placement of the components within the task. This includes spacing between the question stem and the graphic or the location of the graphic within the task. While exploring the role of graphics within numeracy assessment items, Greenlees & Logan (2014) found that item features such as the use of ‘white space’ and the placement of the graphic could impact on a child’s performance and ability to access the required information.

Following on from the initial work of Greenlees (2010), Greenlees & Logan (2014) utilized a similar research design. The study included an initial test (Test A) of 15 items sourced from the 2010 Year 5 NAPLAN based on their relevance to the particular design elements: graphics, literacy demands and item features. These items were administered to 143 Year 5 students (aged 10-11 years) with interviews conducted with 37 of these children. Once again the interview process provided invaluable data into understanding the children’s responses and subsequent modifications were made to the items based on this insight (Test B). One of the items that reported a statistically significant difference between Test A and Test B based on changes to the item features was the Garden Plan Item.

The Garden Plan Item

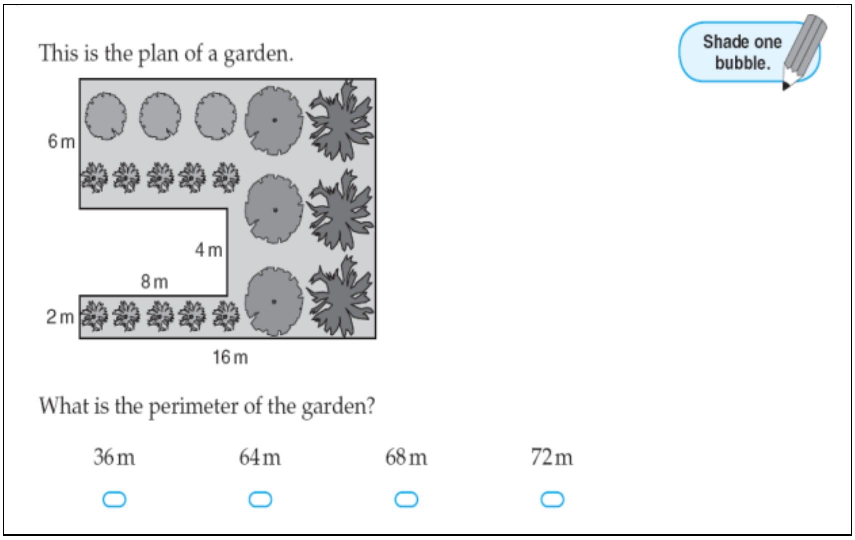
In Test A the Garden Plan Item (Figure 6) recorded a 33% success rate. The analysis of the interview data revealed that students failed to incorporate all the sides into their calculation of the perimeter of the garden, in particular the side closest to the question stem. This was in spite of students’ accurately defining the term perimeter:

> (ANSWER: 36m) I added them all together and then I did partners to 10. So I did 16 and 4 and 8 and 2 and that equalled 30 and I added the 6. (What does perimeter mean?) The outside of the object. [Elise]

It was for this reason that the presentation and layout of the question was modified by centering and rotating the graphic to a more prominent position. Such a change to the item features was statistically significant and improved the mean score of Test B by 14 (see Table 3).

Table 3. Students’ performance on Garden Plan Item across Test A and Test B

	Line graph item	
	Test A	Test B
% Correct	33	47
F (df 1,285) * p ≤ *0.05*	5.91 * p = .016	



Shade one bubble.

This is the plan of a garden.

What is the perimeter of the garden?

36 m	64 m	68 m	72 m
○	○	○	○

Figure 6. The Garden Plan Item

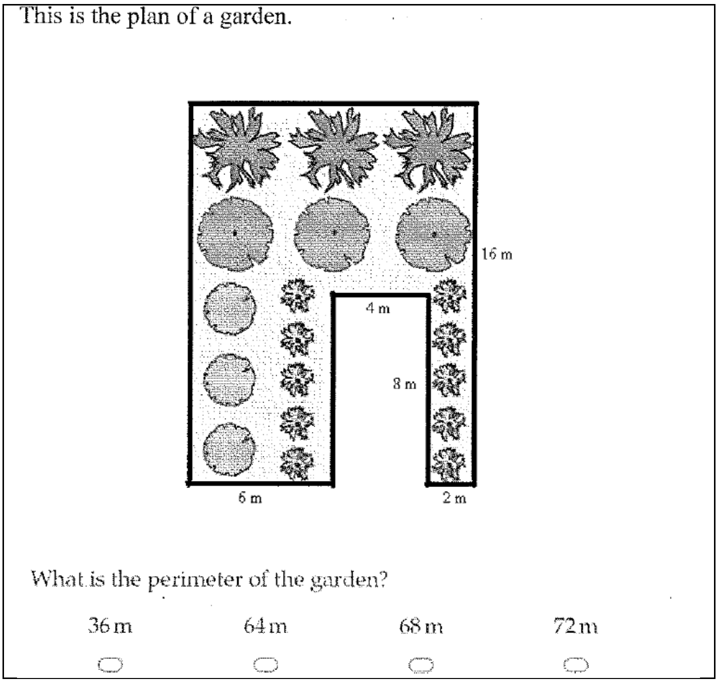
By moving the picture of the garden away from the clutter in Test B (Figure 7), the qualitative data uncovered a heightened awareness to incorporate all the sides into the perimeter equation.

> (ANSWER: 72m) I added all the ones and the ones where there wasn't any numbers like I knew that would be 8 because that's the same as that one and that would be 16 and for that one there was a gap down here so I put 6 and 4 together and that's 10 and then 2 for 12. [Elise]

It appeared that the format of the Garden Plan Item in Test A clearly impacted on students' ability to solve the task. By placing the graphic too close to the question, students failed to notice the need to include those sides that were blank. However, when the sides of the graphic that needed to be calculated were clearly visible, students could effectively determine their value and include them within the perimeter. As noted by Greenlees & Logan (2014), "it could be the case that the inclusion of more "white space" and orientation of the graphic is influential in student performance and reasoning" (p. 216).

What is to come

There are a number of research issues that have emerged from these studies that warrant further exploration. This is particularly timely given both the increased use of graphics in our society and the concentrated focus on high stakes tests and in particular the results of the NAPLAN. It raises the question of the validity of current assessment practices of attempting to measure a child's numeracy. It may be the case that attempting to assess a child's ability to think mathematically and appropriately apply mathematics to real life scenarios is not something that can be measured using current assessment practices. Of course this also raises questions of the validity of future assessments as the Australian government pushes for the NAPLAN in a digital and computerised form.



This is the plan of a garden.

What is the perimeter of the garden?

36 m ○ 64 m ○ 68 m ○ 72 m ○

Figure 7. The modified Garden Plan Item

References

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

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4*, 75-100.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S. Jr., Lindquist, M. M., & Reys, R. E. (1980). Solving verbal problems: Results and implications from national assessment. *Arithmetic Teacher, 28*, 8-12.

Clements, M. A., & Ellerton, N. F. (1996). *Mathematics education research: Past, present and future.* Bangkok, Thailand: UNESCO Principal Regional Office for Asia and the Pacific.

Connolly, N. (2011). Refining the NAPLAN numeracy construct. In J. Clark, B. Kissane, J. Mousley, T. Spencer & S. Thorton (Eds.), *Mathematics: Traditions and (new) practices* (Proceedings of the AAMT-MERGA Conference, pp. 777–785). Alice Springs, NT: AAMT–MERGA.

Gattis, M. (2002). Structure mapping in spatial reasoning. *Cognitive Development*, *17*, 1157-1183.

Greenlees, J. (2010). The terminology of mathematics assessment. In 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, Vol. 1, pp. 218-224). Fremantle, WA: MERGA.

Greenlees, J. & Logan, T. (2014). The influence of graphics in mathematics test item design. Liljedahl, P., Nicol, C., Oesterle, S., & Allan, D. (Eds.). (2014).

Proceedings of the Joint Meeting of PME 38 and PME-NA 36 (Vol. 1). Vancouver, Canada: PME.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000*). Measuring math—not reading—on a math assessment: A language accommodations study of English language learners and other special populations.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Klenowski, V., & Wyatt-Smith, C. (2012). The impact of high stakes testing: The Australian story. *Assessment in Education: Principles, Policy & Practice, 19*(1), 65-70.

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 Research Group of Australasia (Vol 2)*, pp. 655–658. Brisbane: MERGA.

Lowrie, T., & Diezmann, C. M. (2007). Solving graphics problems: Student performance in the junior grades. *Journal of Educational Research, 100,* 369-377.

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

Lowrie, T., Diezmann, C. M., & Logan, T. (2012). A framework for mathematics graphical tasks: The influence of the graphic element on student sense making. *Mathematics Education Research Journal, 24*(2), 169-187.

Pandey, T. (1991). Power items and the alignment of curriculum and assessment. In G. Klum (Ed.), *Assessing higher order thinking in mathematics*, pp. 39-52. Washington, DC: American Association for the Advancement of Science.

Parmar, R. S., & Signer, B. R. (2005). Sources of error in constructing and interpreting graphs: A study of fourth-and fifth-grade students with LD. *Journal of Learning Disabilities, 38*, 250-261.

Queenland Studies Authority. (2002). Aspects of numeracy test: Year 5 (p. 5, 1 – of insert). Spring Hill, QLD:Author

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

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13.

Acknowledgement

I would like to thank Centenary Professor Tom Lowrie, Assistant Professor Tracy Logan and Professor Carmel Diezmann for their contribution to the research.

Jane Greenlees
Charles Sturt University
Boorooma Street
Wagga Wagga NSW 2678
Australia
jgreenlees@csu.edu.au