Looking beyond the answer: The code-breaking world of mathematics assessment

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy at
Charles Sturt University

by
Jane Kathryn Greenlees
BEd (Honours)
Charles Sturt University
2013
Certificate of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma at Charles Sturt University or any other educational institution, except where due acknowledgment is made in the thesis. Any contribution made to the research by colleagues with whom I have worked at Charles Sturt University or elsewhere during my candidature is fully acknowledged.

I agree that this thesis be accessible for the purpose of study and research in accordance with the normal conditions established by the Executive Director, Library Services or nominee, for the care, loan and reproduction of theses.

Signature:  
Date:  
30th January 2013
Acknowledgements

This section became one of the last things that I needed to write for this thesis but indeed one of the most satisfying and enjoyable aspects of this PhD journey.

First and foremost, I would like to thank my amazing husband Jason, who has been my biggest fan and greatest asset. From the conception of the PhD possibility he has provided overwhelming support and encouragement far beyond my expectations. Thank you to my gorgeous children Harrison, Emily and Mikayla, who sacrificed a lot of ‘mummy time’ but were always there for kisses and cuddles when I needed them the most.

Sincere thanks to my talented supervisor, Professor Tom Lowrie, with whom I have a love/hate relationship. His knowledge knows no bounds, of which he can attest to, and for this I am so grateful. His insight into every aspect of the project allowed me to remain confident and motivated despite my moans and groans in between. He has not only been a wonderful mentor and supervisor but a caring, compassionate and faithful friend.

To my associate supervisor Dr Lindy Croft-Piggin who lovingly read every word of this thesis while sharing her special ‘writing room’ where ideas and sentences magically fell from the sky. Thank you for your calming influence, patience and kindness.

To all the wonderful RIPPLE staff for their encouragement and support. In particular Tracy Logan, whose door was always open to bounce ideas off and analyse results. Her genuine belief in my ability to complete this thesis helped me see the light at the end of the tunnel even when it was very small. Furthermore, I would like to thank Kim Woodland who took on a huge
amount of editing work in a short period of time but still managed to do it with a smile on her face.

Finally to my fabulous friends and family who I would love to name individually but the list may end up longer than the thesis, who remained my committed and loyal cheer squad from beginning to end. I am especially grateful to my sisters Sarah, Elizabeth, Kate and Amy for their love and concern not just for me but for Jason and the kids. However it goes without saying that the loudest cheers were always coming from my amazing parents. Words cannot express how grateful I am to you for your love and support for not only this aspect of my life but at every stage along the way.
Abstract

For many years high-stakes testing has been a contentious issue in the education field, and despite findings for and against the use of such tests, their prevalence remains. In an Australian context, the launch of the first National Assessment Program for Literacy and Numeracy (NAPLAN) in May 2008 intensified debate in this area. Since that time there has been a heightened awareness of using such tests to measure students’ mathematical ability and as a tool to evaluate teachers and schools’ performance. Even greater controversy was created when the Australian Government announced that schools’ results would be publicly reported upon and compared to others of similar demographics. With increased accountability comes great responsibility that the tests being used are achieving their intended rationale, in this case accurately measuring a child’s level of numeracy. It was for this reason that one of the purposes of this thesis was to investigate the impact of current test item design on students’ mathematical performance.

This study acknowledged the dramatic change in test item design over the past 10 years with greater attention afforded to graphical and pictorial representations and tasks which embedded ‘real-life’ situations. Specifically, the study outlined the changed behaviour (and performance) of students as they solved a selection of items taken from the NAPLAN that had been modified in relation to: (1) graphics; (2) language; (3) real-life situations; and (4) design features.

The study involved a sequential mixed method design. 143 Year 5 students participated in an experimental design that required them to solve items from the 2010 Mathematics NAPLAN and modified versions of these items.
In addition, 37 purposely-sampled students were interviewed to gain a deeper appreciation of the strategies they used when solving the respective tasks.

The study found that students experienced difficulties engaging with the decoding and encoding of graphics, in particular, the discernment between those graphics deemed useful in solving the task and those not necessary. The students were challenged by the unnecessary language often incorporated in creating real-life situations, and benefited from the decontextualising of items according to the graphic or language used.

A model of ‘the contexts of mathematics assessment’ also emerged from the quantitative and qualitative data. The model proposed that three contexts: (a) assessment; (b) conditional; and (c) item, need to be considered when attempting to measure a child’s numeracy. The creation of the model was examined in light of test item design and teaching practices.

There were a number of implications for theory, learning and teaching, and test and curriculum design arising from the study. From a theoretical perspective, the findings of the study suggest that the use of items developed within real-life situations can be problematic for some students. In terms of teaching and learning, the model provides teachers with a tool that scaffolds their understanding of task design which can better equip students to be empowered in high stakes testing situations. The thesis does not advocate ‘teaching to the test’, but rather the development of sound mathematical teaching strategies for the long-term benefit of the students. Finally, the implications for test designers are presented, which highlight the dramatic effect that slight change in test design can have on student’s sense making.
Table of Contents

Certificate of Authorship .......................................................................................... iii

Acknowledgements .................................................................................................. v

Abstract ..................................................................................................................... vii

List of Tables ............................................................................................................. xiii

List of Figures .......................................................................................................... xvi

Chapter One: Introduction ...................................................................................... 1
  The Context and Scope of the Investigation ....................................................... 1
  The Mechanics of the Code ................................................................................. 3
    The Tools Required .............................................................................................. 6

  Measuring Success ................................................................................................. 7

  Aims of the Study ................................................................................................... 8

  Significance of the Proposed Research ............................................................... 8

  Plan of the Thesis .................................................................................................. 9

Chapter Two: Research literature and theoretical underpinnings .................. 10
  Introduction ........................................................................................................... 10
  The assessment story so far ................................................................................. 10
    Making mathematics real? ................................................................................ 13

  Assessment in practice ......................................................................................... 19

  Graphics ................................................................................................................. 20

  Language ............................................................................................................... 32

  Situations ............................................................................................................. 41

  Item Features/Design ......................................................................................... 48

  Relationship/Connectivity of Components ...................................................... 54

  Summary: The Fantastic Four of Mathematics Assessment ........................ 65

  Implications for the Present Study ..................................................................... 66

Chapter 3: Methodology ......................................................................................... 67
Introduction ........................................................................................................................ 67
Research Questions ........................................................................................................... 67
Research Design ............................................................................................................... 68
  Internal and External Representations ................................................................. 68

Utilising the Strengths of Quantitative and Qualitative Approaches ........................................ 72

The Site and the Participants ......................................................................................... 75
  Implementation of Qualitative and Quantitative Instruments .................................. 76

Methods ....................................................................................................................... 79
  Quantitative Testing ................................................................................................. 86
  Quantitative Analysis ............................................................................................ 101
  Qualitative Interviews ........................................................................................... 103
  Qualitative Analysis ............................................................................................... 105

Quality and Rigour of the study ................................................................................ 107
  Credibility ................................................................................................................ 108
  Dependability and Transferability ....................................................................... 109
  Ethics ......................................................................................................................... 109

Chapter Summary .................................................................................................... 111

Chapter 4: Presentation and Analysis of Data ................................................................. 112
  Student Performance of Initial NAPLAN Item Instrument ................................ 112
  Quantitative Data .................................................................................................... 112

  Qualitative Data ...................................................................................................... 116

  Standard versus Modified Instruments—A Tale of Two Tests ............................ 118
    Analysis 1—Descriptive “Smiley Faces” Approach .......................................... 119
    Analysis 2—Cross Tab Contingency Tables ....................................................... 120
    Analysis 3—Analysis of Variances ..................................................................... 127

  Graphics Quantitative Analysis ............................................................................ 130
    The Impact of Modifying Graphics on Student Performance ........................... 131
List of Tables

Table 2.1 Guidelines for Writing Multiple Choice Items .................................. 51

Table 3.1 Means and standard deviations between test results .................. 85

Table 4.1 Year 5 2010 NAPLAN Test Results .................................................. 113

Table 4.2 Coding analysis of interviews by item component ...................... 117

Table 4.3 Combined smiley analysis ............................................................... 119

Table 4.4 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 10. ............................. 121

Table 4.5 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 4. ................................. 121

Table 4.6 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 5. ................................. 122

Table 4.7 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 3. ................................. 123

Table 4.8 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 8. ................................. 123

Table 4.9 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 2. ................................. 124

Table 4.10 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 9. ................................. 125

Table 4.11 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 6. ................................. 125
Table 4.12 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 1 ................................. 126

Table 4.13 Contingency Table for Cross Tab Analysis for performance on NAPLAN and Modified NAPLAN Question 7 ................................. 126

Table 4.14 Gender means and standard deviations ........................................ 128

Table 4.15 Means, standard deviations and univariate statistics for the items on the NAPLAN and Modified NAPLAN ........................................ 128

Table 4.16 Univariate analysis and modifications of graphics items ........ 130

Table 4.17 Univariate analysis of Liquorice allsort item in NAPLAN and Modified NAPLAN ................................................................. 132

Table 4.18 Univariate analysis of Coins item in NAPLAN and Modified NAPLAN ................................................................. 134

Table 4.19 Univariate analysis of Shoe item in NAPLAN and Modified NAPLAN ................................................................. 138

Table 4.20 Univariate analysis and modification of language item .......... 140

Table 4.21 Univariate analysis of pie-graph item in NAPLAN and Modified NAPLAN ................................................................. 142

Table 4.22 Univariate analysis and modifications of item features ........ 144

Table 4.23 Univariate analysis of Boxes item in NAPLAN and Modified NAPLAN ................................................................. 146

Table 4.24 Univariate analysis of Garden plan item in NAPLAN and Modified NAPLAN ................................................................. 149

Table 4.25 Univariate analysis of Street map item in NAPLAN and Modified NAPLAN ................................................................. 152
Table 4.26 Univariate analysis of Baby mass item in NAPLAN and Modified NAPLAN
........................................................................................................................................155

Table 4.27 Univariate analysis and modification of context items .......... 156

Table 4.28 Univariate analysis of Meeting schedule item in NAPLAN and Modified NAPLAN
........................................................................................................................................157

Table 4.29 Univariate analysis of scale item in NAPLAN and Modified NAPLAN
........................................................................................................................................161
List of Figures

Figure 2.1. Comparison of 1997 Year 3 NSW BST to an item from the 2010 Year 3 NAPLAN.......................................................... 17

Figure 2.2. The components of mathematics assessment tasks. ............... 18

Figure 2.3. Example of a contextual graphic. ........................................... 22

Figure 2.4. Example of an information graphic. ........................................... 22

Figure 2.5. Example of an Axis language item........................................... 24

Figure 2.6. Example of an Apposed-position language.............................. 25

Figure 2.7. Example of a Map language item............................................ 26

Figure 2.8. Example of a Retinal-list language item.................................... 26

Figure 2.9. Example of a Connection language item................................... 27

Figure 2.10. Example of a Miscellaneous language item ......................... 27

Figure 2.11. Newman’s model of reading skills for the language of mathematics. ........................................................................ 34

Figure 2.12. The angel task utilised in Clements (1980). ......................... 37

Figure 2.13. The comprehension task utilised in Clements (1980). ............ 38

Figure 2.15. Marbles problem................................................................. 45

Figure 2.16. Ham problem................................................................. 45

Figure 2.17. The banana problem.......................................................... 47

Figure 2.14. Interpret statistical diagrams.................................................. 62

Figure 3.1. The relationship between internal and external representations. ........................................................................ 71
Figure 3.2. Research design.................................................................80
Figure 3.3. Pilot test item.................................................................82
Figure 3.4. Modified pilot test item..................................................84
Figure 3.5. Graphics modification.....................................................93
Figure 3.6. Language modification...................................................95
Figure 3.7. Item Features modification............................................97
Figure 3.8. Situation modification....................................................100
Figure 3.9. Definitions of Codes......................................................106
Figure 4.1. The Liquorice allsort item.............................................131
Figure 4.2. The Coins item..............................................................134
Figure 4.3. The Shoe item...............................................................137
Figure 4.4. The Pie graph item.........................................................141
Figure 4.5. The Boxes item.............................................................146
Figure 4.6. The Garden plan item....................................................148
Figure 4.7. The Street map item.......................................................151
Figure 4.8. The Baby mass item......................................................154
Figure 4.9. The Meeting schedule item............................................157
Figure 4.10. The Scales item...........................................................160
Figure 4.11. Max—Coins.................................................................165
Figure 4.12. Max—Modified coins..................................................166
Figure 4.13. Max—Shoe.................................................................167
Figure 4.14. Max—Modified shoe....................................................169
Figure 4.15. Max—Garden plan. ................................................................. 170
Figure 4.16. Max—Modified garden plan. .................................................. 171
Figure 4.17. Emily—Pie graph. .................................................................. 173
Figure 4.18. Emily—Modified pie graph....................................................... 174
Figure 4.19. Emily—Scales. ....................................................................... 175
Figure 4.20 Emily—Modified scales. ........................................................... 176
Figure 4.21. Sophie—Boxes. ....................................................................... 179
Figure 4.22. Sophie—Modified boxes. ......................................................... 179
Figure 4.23. Sophie—Baby mass................................................................. 181
Figure 4.24. Sophie—Modified baby mass. ................................................. 182
Figure 5.1. A model of the contexts of mathematics assessment. ............... 187
Figure 5.2. The first circle of the model. ....................................................... 188
Figure 5.3. The middle circle of the model.................................................. 188
Figure 5.4. The use of one-way arrows within the model......................... 189
Figure 5.5. The final circle of the model...................................................... 191
Figure 5.6. A model of Emily’s numeracy............................................... 192
Figure 5.7. Emily’s assessment context...................................................... 192
Figure 5.8. Emily’s conditional context...................................................... 193
Figure 5.9. Emily’s item context............................................................... 193
Figure 5.10. A train timetable. ................................................................. 211
Chapter One: Introduction

The Context and Scope of the Investigation

In May of 2008 the first National Assessment and Planning: Literacy and Numeracy (NAPLAN) test took place in Years 3, 5, 7 and 9 in schools across Australia. As anticipated, this is the first initiative toward a national curriculum. This initiative has sparked debate among education researchers, with some raising concern in regard to “assumptions about knowledge, learning, teaching and assessment” (Atweh & Singh, 2011, p. 189). The NAPLAN test has forced assessment further into the spotlight with increased accountability placed on teachers due to the public reporting of results. The Australian Government’s push for high stakes testing, it is claimed, is drive by “a desire to meet public accountability, demonstrate transparency and maintain public confidence in the standards of schooling (Klenowksi & Wyatt-Smith, 2012, p.65). They have used assessment as a means of identifying and targeting low-performing schools and guaranteeing resources to ensure no child is disadvantaged. The current study does not consider the political and social implications of high-stakes testing but rather the foundational structure of assessment items. The title of the thesis was framed around the phrase the code-breaking world of mathematics assessment since it is imperative that we understand the manner in which students engage with mathematics assessment. Furthermore, it is necessary that we understand how students make sense of these items when it is widely recognised that the tasks presented in the NAPLAN are fundamentally different in nature and composition to
mathematics tasks previously administered to students of this age (Lowrie & Diezmann, 2009).

In terms of the analogy of the code-breaker entitled in this thesis, the term ‘code’ is used to describe an unknown, something that needs to be cracked. The dictionary defines it as “a system used for brevity or secrecy of communication, in which arbitrarily chosen words, letters, or symbols are assigned definite meanings” (Dictionary.com, n.d.). Similarly, assessment (and especially high-stakes assessment) holds a degree of secrecy as to what the results actually reveal about students’ numeracy as well as the nature and impact of its composition (see Haladyna, 2006; Hargreaves, Shorrock-Taylor, Swinnerton, Tait, & Threlfall, 2004; Haladyna, Haas, & Allison, 1998; Clements & Ellerton, 1996; Gipps, 1994). Its use of particular words and graphics are assigned meanings and purpose set out by the coder (or test designer) that need to be interpreted or cracked by the decoder (or students).

Both students and teachers need to have access to the keys necessary to unlocking the code of mathematics assessment. In particular: (a) the structure of mathematics assessment; (b) the way students interact with these elements; and (c) the extent to which specific components of a task influence sense making. It is for this reason that this study will explore each of these issues.

In order to unlock these assessment codes, there was a need to consider the structure and modification of mathematics assessment items. The literature detailing the development of assessment over time to its prevalence today was examined in light of the paradigm exploring the modification of items within assessment. This included modifying aspects within an item to recognise any possible changes in students’ responses utilising both
quantitative and qualitative methods. Another key to unlocking the codes of assessment lay in the decoding and encoding preferences of those attempting to crack the assessment code. Interviews have been used in this study to uncover these strategies. It was also necessary to examine the establishment of meaningful understandings of students’ interpretation of tasks and to determine the relationship between quantitative measures (task correctness) and in depth qualitative analyses (students’ thinking). This was achieved through a mixed-method approach where results utilising ANOVAS were interpreted in light of the interview data. To relate this back to the code-breaking scenario, these three aspects deal with the mechanics of the code, the tools and insight required, and the way success can be transferred and measured, all of which will now be explored further.

**The Mechanics of the Code**

When a computer hacker is preparing to infiltrate a system they must first become familiar with the mechanics of the programming code. This may require the hacker to research and practice to gain confidence in their ability as well as develop appropriate strategies and plan a course of action. Similarly, when a child encounters a question within a high-stakes test, they need to understand the complexity of the task, be able to identify the key elements, practice and develop code-breaking strategies and approach the task with assurance. For just as a computer hacker may get caught and serve gaol time, poor performance in high-stakes tests condemn children to a sense of failure and inadequacy, often a precursor to future mathematical challenges. Although this is an extreme comparison, such analogies highlight the importance of analysing the mechanics and structure of high-stakes mathematics assessment to better inform all stakeholders. Indeed, as
Boaler (2012) has recently claimed, one poor performance of a publicly ‘timed’ test can dramatically impact on a students’ performance for the rest of their schooling and beyond. It was for this reason that the first key focused on in this study was one to unlock the code of mathematics assessment by investigating the mechanics of the task.

When examining the structure of mathematics assessment tasks in the past, research has often focused on just one aspect of the item. These have included studies analysing the use of: (1) language (De Corte, Verschaffel, & De Win, 1985; Kintsch & Greeno, 1985; Adetula, 1990; Gerofsky, 1996; Zevenbergen, 2000b); (2) the influence of graphics (Mayer & Gallini, 1990; Lowe, 1993; Lowrie & Diezmann, 2005; Lowe & Promono, 2006); (3) the use of real-life contexts (Boaler, 1993; Verschaffel, De Corte, & Lasure, 1994; Cooper, 1998b; Cooper & Harries, 2002; Sullivan, Zevenbergen, & Mousley, 2003); or (4) the types of questions used e.g., multiple choice versus short-answer response (Haladyna & Downing, 1989a; Haladyna, Downing, & Rodriguez, 1999). Although these studies have provided invaluable findings into the impact of such aspects on a child’s ability to solve tasks, they do not analyse or acknowledge the relationship that exists between them all.

In the case of our computer hacker, this is similar to just hacking into one aspect of a computer’s code, allowing the hacker restricted access to one area only. However if the hacker wishes to experience the full benefit of total access to all the architecture of the site they would need to be familiar with more than just the language it is written in. They would also need to be able to understand the database’s schema, security settings, and how to access the file structures and directories to gain control over all areas of the
site. It was for this reason that all the mechanics of a mathematics assessment item were examined in the present study. This included the relationship and impact of the language, graphics, situations and item features contained within a task.

Another aspect of the study apart from what was being decoded was the way in which this was being achieved. To break the code of the four components of test item design, a relatively under-researched modification paradigm was utilised. This included modifying aspects within an item to analyse its impact on a child’s mathematical reasoning and to record changes in performance. The process of modification can be defined as “a process by which the test developer starts with a pool of existing test items with known psychometric properties, and makes changes to the items, creating a new test with enhanced accessibility for the target population” (Kettler, Elliott & Beddow, 2009, p. 531).

While the systematic process of item modification had previously been utilised in research to improve test accessibility for students with disabilities (Kettler, Elliot, & Beddow, 2009), its purpose within the present study was to improve accessibility for all students. The term “access” has been defined as providing the opportunity for students to adequately and fairly demonstrate their content knowledge on which they are being tested. (Kettler, Elliot & Beddow, 2009). The intended results of increasing and improving access of items was threefold. Initially it was anticipated that by modifying items according to one aspect of test item design it would highlight its impact on student performance. Consequently, this led to the second expectation of recognising possible gaps in children’s code-breaking abilities. The final expectation was emphasis on high-quality test-item
design and the need for explicit instruction of code-breaking strategies in the classroom. This included the need to identify and acknowledge the role the four components play in measuring a child’s mathematical understanding.

However the planning and understanding of the mechanics of the code that one is attempting to crack is only the first step in the code-breaking process. The next step is identifying the relevant tools needed to complete the job.

**The Tools Required**

In regards to the computer hacker there are a number of tools available to them when attempting to crack the code surrounding a site. These are often specific in nature to the aspect they are trying to infiltrate. For example, if attempting to crack the code on a site one may use a tool such as Adobe Dreamweaver, however to access the database a Structured Query Language (SQL) chart tool would be required. Another alternative is to customise and build your own automated program or ‘bot’ that allows the computer hacker access to the back end of the site and negates the need to individually address every aspect. The other benefit of the development of the ‘bot’ is that it can then be used to hack into other sites with similar programming characteristics. In the same manner the present study attempted to build a ‘bot’ or code-breaking model that could be utilised by all stakeholders involved in the assessment process. This includes the many people impacted by the implications of high stakes assessment. For teachers and schools, it provides insight into how elements of the NAPLAN need to be explicitly incorporated into their classroom teaching; for test designers the need to carefully consider the design of mathematics assessment; and for governing bodies to be alert to the impacts of the contexts surrounding high-stakes tests.
To create this model or tool it was necessary to examine the encoding and decoding techniques of students within each individual component or ‘mechanics’ of the test item. These components include the language, graphics, item features and situations referred to in this study as the ‘fantastic four’ of mathematics assessment. It was anticipated that these techniques and findings could be used to create a model that would allow ‘back end’ access to any assessment with a similar framework. By recognising the relationship between the way students decode the four aspects of an assessment item and subsequent encoding of the item, a model of appropriate and inappropriate problem-solving strategies could be derived. This ‘bot’ will provide an invaluable insight into how the mechanics of assessment items impact on students’ mathematical reasoning. It also accesses the holy grail of mathematics assessment, that is, the relationship between a student’s correct answer and their mathematical thinking and ability.

**Measuring Success**

The journey of the computer hacker leads to this pinnacle point, access to the very heart of the site. So far they have carefully analysed the mechanics, implemented appropriate tools, and possibly even developed their own automated program all in an effort for total insight and control of the site. Similarly, by recognising the mechanics of mathematics assessment through the modification paradigm and developing a tool or model to assist in cracking the complicated assessment code, the research study aims to expose the reality of mathematics assessment results. Fundamentally, how well do results reflect mathematics ability and can success be a measure of competence?
**Aims of the Study**

In order to investigate the code-breaking world of mathematics assessment the following four research questions were posed. The manner in which these questions evolved is detailed in Chapter 3.

1. From a selected number of often used mathematics tasks, which do students find most difficult to solve?
2. How does the use of graphics, language, situations and item construction impact on a child’s capacity to make sense of mathematics?
3. To what extent does assessment reflect sense making and mathematical understanding?
4. What effects will these findings have on classroom teaching?

**Significance of the Proposed Research**

Research into these questions comes at a very poignant time when assessment is held in high regard as an indicator of a child’s mathematical knowledge. Answers to these questions will assist in analysing what assessment results actually reveal about a child’s understanding and possibly highlighting the ambiguity of it all. The research findings will also throw light on whether, in future, teachers might be able to prepare students for mathematical assessment not by ‘teaching to the test’ but incorporating some of the research findings within their classroom practice. This means explicitly teaching skills necessary in achieving within the NAPLAN, but making sure they benefit the child overall, and are not just a rehearsal of test items. It is also anticipated that it will alert test designers and pyschometricians to the impact of test-item design on students’ performance.
and the need to ensure validity within assessment, that is, that the assessment is examining its intended purpose.

**Plan of the Thesis**

This thesis is divided into six chapters. This first chapter provides an introduction to the problem studied, to the background of the study, and its purpose. The four research questions proposed for the investigation are identified.

Chapter 2 reviews the literature on the changing nature of mathematics assessment, and explores the theoretical frameworks surrounding the four components of test item design.

Chapter 3 describes the setting in which the inquiry took place, the design and methodology used, results of an initial pilot study and the subsequent development of the research instruments.

Chapter 4 presents the results of univariate statistical analyses in light of qualitative data obtained from interviews conducted with students who were outlining their problem-solving techniques. A more detailed analysis of certain aspects of test-item design is explored through three case studies.

Chapter 5 includes a model developed from the research findings and highlights its relevance to teachers and test designers.

Chapter 6 offers summary responses to the research questions posed in Chapter 1. It acknowledges limitations of the study and the implications of the findings for future research are discussed.
Chapter Two: Research literature and theoretical underpinnings

Introduction

The central concern of this thesis is to identify the various components of mathematics assessment items and investigate the impact they have had on a student’s performance. Consequently the theoretical basis for the investigation evolved from the literature on the use of graphics, language, situations and design in mathematics assessment items. Various theoretical frameworks are described to assist in creating an overall understanding of the issues students contend with in regards to assessment items today.

These frameworks assisted in not only describing the nature of assessment items but also in gaining an understanding of how and why these came about. The literature suggested that the changing nature of mathematics assessment must be analysed and understood according to the changing expectations of society and the subsequent teaching of mathematics in the classroom.

The assessment story so far...

A National Curriculum for schools in all states and territories of Australia, from Kindergarten to Year 12, is currently being implemented. The first stages are scheduled to commence in 2013, with some States and Territories already beginning to transition and implement a draft version within schools. Presently, each state and territory has its own ‘leaving’ certificate for senior students and it is indicated that this will not become a national initiative. In Queensland it is called the Queensland Certificate of Education
(QCE), for Western Australia it is the Western Australian Certificate of Education (WACE) and for New South Wales the Higher School Certificate (HSC). For the purpose of this study the assessment context will be examined and contained within New South Wales however the analogies used are similar across all Australian States and Territories.

Many people can recall some form of mathematics assessment during their years of schooling. Whether it was their High School Certificate (HSC), their School Certificate, a classroom exam, State Mathematics Test or more recently the National Assessment Program: Literacy and Numeracy (NAPLAN). I can vividly remember my Mathematics HSC even though it was over 15 years ago. I can recall waiting nervously outside our school auditorium beforehand with my sharp 2B pencils, erasers and calculator in hand. Some of my fellow classmates were agonising over well-worn notes in a last ditch effort to retain some vital theorem and nervously predicting what may or may not be in the exam to come. The truth is that assessment has been a reality for all of us and a memorable part (both good and bad) of our education experience.

However, as Lowrie and Diezmann (2009) noted, it was not until the early 1990s that a worldwide push towards centralised testing began. This included the infamous 2001 No Child Left Behind Act in the United States. They argued that this was closely aligned to an era of economic rationalism and subsequent public accountability. In regards to large scale assessment programs in New South Wales, the Basic Skills Testing Program (BSTP) was introduced into government schools in 1989 (Doig & Masters, 1992). According to Mitchelmore and White (2003) the test was
developed by the DET (the New South Wales Department of Education and Training) over a 12-year period using Rasch modelling and test-equating techniques [to ensure] that scores [were] reported on a fixed scale, and—within the limits of computer scored testing—[were] regarded as a most valid and reliable numeracy assessment instrument (p. 515).

This test was replaced in 2008 with a national standardised test, the NAPLAN, as part of the Australian Federal Government’s commitment to implementing a national curriculum.

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 use of standardised tests (Kulm, 1991; Linn, 2000; Webb, 1992) 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 (Howson, 1993). In fact, Pandey (1991) estimated that about 30 to 35 million tests are administered annually in the United States.

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). Although classroom practices have not always mirrored changes in assessment rationale, it is more likely the case that changes will take place when assessment is framed around such high stakes measures. Thus, either explicit or gradually emerging changes in classroom practices will occur and these changes will merge toward NAPLAN-like content. Furthermore, there now seems to be a blending of assessment contexts that are realistic (e.g., bus timetable) and representations that are realistic (e.g., real bus timetables being inserted into assessment items).

**Making mathematics real?**

According to Boaler (1993a) “the abstractness of mathematics is synonymous for many with a cold, detached remote body of knowledge” (p. 13). It was for this reason that in the late 1970s there was a major shift toward the ‘everyday’ use of mathematics in a hope that mathematics taught in the classroom would be seen as relevant and easily transferred into ‘real’ situations (Boaler, 1994). These advocates of everyday mathematics believed that this kind of focus not only prepared students in understanding the content studied, but also the use of ‘real world’ questions bridged the gap between the abstract role of mathematics and their roles as members of society (Boaler, 1993b). Freudenthal (1991) noted this was the creation of context-rich mathematics instruction later to become known as rich contexts. As a result textbooks started to exhibit questions utilising common scenarios that an adult or child may face outside the walls of the classroom (Gerofsky, 1996). Items such as shopping lists, timetables, graphs and maps replaced traditional algorithms in an attempt to improve students’ abilities to interpret events around them. Boaler (1993a) suggested that using “real
world, local community and even individualized examples which students may analyse and interpret is thought to present mathematics as a means with which to understand reality” (p. 13).

Furthermore, Bonotto (2001) stressed that

bringing real world situations into school mathematics is a necessary, although not sufficient, condition to foster a positive attitude towards mathematics, intended both as an effective device to know and critically interpret reality, and as a fascinating thinking activity (p. 76).

However, what needs to be acknowledged is that this educational objective can only be achieved if the students can bring mathematics into reality, not simply reality into mathematics. Indeed, it is not the case that providing a realistic scenario will result in students being able to apply the mathematics to other situations.

It is the child’s inability to recognise the need to include mathematics into their reality that has raised questions over the effectiveness of utilising real life situations in the mathematics classroom. Can students actually treat an item as a description of some real-world context and not view it as just a mathematics question? De Corte, Verschaffel and Greer (2000) argued that authors have found

by the end of the elementary school many pupils have constructed a set of beliefs and assumptions about doing mathematical application problems, whereby this activity is reduced to the selection and execution of one or a combination of the four arithmetic operations with the numbers given in the problem, without any serious consideration of possible constraints of the realities of the problem context that may jeopardize the appropriateness of their standard models and solutions (p. 66).

Therefore, in order to provide students with the appropriate mathematical tools and understandings (from a conceptual perspective), notions of what
could be real or not become blurred and non-existent. Once again there is a
tension between providing realistic scenarios and applications of
mathematics in a real world beyond school.

It is therefore no coincidence that the new Australian mathematics
curriculum that is currently being developed echoes a similar ideology as it
highlights its aim to create “confident, creative users and communicators of
mathematics, able to investigate, represent and interpret situations in their
personal and work lives and as active citizens” (Australian Curriculum,
Assessment and Reporting Authority [ACARA], 2012). Similarly this
philosophy of teaching ‘real’ mathematics is also reflected in the Dutch
domain-specific educational theory for mathematics education that was
known as *Realistic Mathematics Education* (RME) (Van Den Heuvel-
Panhuizen, 2005). In RME, “this connection to reality is not only
recognizable at the end of the learning process in the area of applying skills,
but also reality is conceived of as a source for learning mathematics” (Van
Den Heuvel-Panhuizen, 2005, p. 2). It is therefore a natural progression that
this evolution of mathematics instruction has been reflected in the changing
nature of mathematics assessment.

This becomes apparent when we compare an item from the 1997 Year 3
NSW Basic Skills Test to an item from the 2010 Year 3 NAPLAN (see
Figure 2.1). The top item from this figure is contextualized within a scenario
that may well represent aspects of an authentic child’s representation of a
map. This type of map would be described as a mud map (or treasure map)
that would be devoid of scale and perspective. By contrast, the item below
could be described as a more realistic rather than authentic item since it
would be common to represent a room in such a 2D form. This item does have elements of perspective and proportion (if not scale) within its context.

Both questions reflect the same Board of Studies New South Wales syllabus outcome i.e., “SGS2.3 Uses simple maps and grids to represent position and follow routes” (2002, p. 136). However in 1997 the term ‘map’ is given to a diagram that does not contain any map conventions such as a grid or coordinates but rather records Carol’s path using a segmented line and arrows. Comparatively in the 2010 NAPLAN a more sophisticated map or plan of a bedroom is used. It is assumed that this is a graphic that students would encounter regularly and be familiar with its conventions and use.

The first representation in the figure has been constructed in order to tell a picture story rather than it being a map that would depict an authentic situation. It is also the case that this diagram is not likely applicable or used for any other situation. By contrast, the second figure is built on map conventions with the diagram certainly applicable to those conventions in architecture and design.
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. Subsequently relevant research and literature needs to be reviewed in relation to these emerging components of current mathematics assessment tasks and how each one individually contributes to a child’s ability to solve the task as well as their holistic impact. The literature explored so far has focused on the following four components:

1. Graphics;
2. Language;
3. Situations; and
4. Item features.

Figure 2.2 illustrates the role each of these play in regards to an assessment item.

Figure 2.2. The components of mathematics assessment tasks.¹

¹ Item example sourced from (ACARA, 2009)

² To aid in the recognition of answers throughout the thesis a system of A, B, C and D will be utilised to describe student responses.

³ Real life scenarios are used in other national and international high-stakes testing e.g.
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. The final component of the item, situational understanding, involves the interaction of the preceding three components, however, this also involves specific reference to student understanding of particular conventions (e.g., knowing how to read a street map).

To the best of my knowledge, no other study has considered these components both individually and holistically. In this study a combination of relevant theoretical frameworks and findings are utilised to create a detailed model of the nature of mathematics assessment items. It is anticipated that this model will assist teachers and test designers in understanding and subsequently addressing some of the aspects students must contend with in assessments. It is also hoped to highlight the ambiguous nature of assessment results as a true reflection of sound mathematical understanding.

**Assessment in practice**

Literature regarding current assessment practices will be analysed in two ways. Firstly within the context of the four individual components of test item design and secondly in light of the connectivity and relationship that exists between graphics, language, situations and item features. This will
draw upon specialised research as well as theoretical frameworks to assist in understanding the interaction that exists between the four components.

**Graphics**

It is important to note that within mathematics education research, the term ‘graphics’ has been utilised in many unique and different ways. For example, Postigo and Pozo (2004) defined graphic formats as “geographical maps, diagrams, illustrations and numerical graphs” (p. 624), while Lowrie, Diezmann and Logan (2012) introduced ‘graphics tasks’ as containing text stimulus, a text question, symbolic answer and a graphic component. Wainer (1980) and Kosslyn (1985) measured a child’s graphicacy as their ability to read graphs while Bertin (1967/1983) defined graphics as visual representations for “storing, understanding and communicating essential information” (p. 2). Within the context of this study, graphics refers to any diagram, pictorial representation or graph used within an item. This definition implies that the graphic contains its own conventions (Mayer, 1989) and consequently needs to be identified as a separate component particularly since it also has obvious relationships to other components of the task.

According to Lowe and Promono (2006), “text and graphics have long been combined in various ways to provide complementary sources of information on a wide variety of topics” (p. 22). Although their research has focused on the use of graphics to support comprehension of dynamic information in texts, the issues and concerns they raised are applicable to the use of graphics within mathematics assessment items. They argued that although supplementing text with a suitable graphic can assist in comprehension, this understanding depends on the characteristics of the graphic and its
interaction with the reader. Therefore the use of a graphic within an assessment item has the potential to have both a positive and negative effect.

It is this interaction with the reader that Lowe (1993) noted was the key to understanding the effectiveness of using graphics. He maintained that a graphic’s usefulness not only relied on quality design but the experience of the reader who was attempting to utilise it. Therefore he concluded that “a research agenda for improving the effectiveness of diagrams as a means of communication needs to include a study of potential users, as well as the quality of the diagrams” (p. 4). Thus, it is imperative that students’ reasoning and sense making is explored as they engage with graphics and not simply to report on success and task correctness in relation to understanding. Furthermore, the graphic conventions may at times have more influence on the demands of the task than the mathematics concepts embedded in the task.

Although other researchers have claimed that there are “numerous and varied ways in which people illustrate ideas or concepts” (Kosslyn, 1989, p. 185) the classification of graphics can be framed within two distinct forms which have fundamentally different purposes. The two graphics forms are contextual graphics and information graphics.

**Contextual graphics versus information graphics**

According to Diezmann and Lowrie (2008), contextual graphics can be defined as those that ‘complement’ a text or symbolic representation while information graphics ‘supplement’ and represent mathematical information. A similar distinction was made by Lowe and Promono (2006), who identified one group of graphics responsible for echoing the information
already contained in the text and the other group providing additional information to the text. For example, in Figure 2.3, an item taken from the 2010 NAPLAN, the picture of the shoe *complements* the text but contributes no mathematical information. By contrast, in Figure 2.4, also containing an item from the 2010 NAPLAN, the picture of the lollies *supplements* the text and is essential to calculating the answer.

![Figure 2.3. Example of a contextual graphic.](image1)

![Figure 2.4. Example of an information graphic.](image2)

In the Year 5 2010 NAPLAN, within the 31 items that contained a graphic, 29 were information graphics while only 2 were contextual. With such a
disproportionate variance in the use of information graphics over contextual, it raises the issue of the necessity of a contextual graphic within mathematics assessment. This is particularly relevant when considering the extraneous load it may put on a child’s mental ability to solve the problem.

**Information Graphics**

According to Lowrie, Diezmann and Logan (2012), information graphics can be defined as those that “contain information essential to the task which is not presented elsewhere (i.e., in text or symbols)” (p. 170). These included items such as graphs, diagrams, charts, tables and maps. They need to be decoded in tandem with any text or symbols for successful completion of the task.

Yet contained within the graphics themselves Bertin (1967/1983) noted a number of visual-spatial elements that make each one unique in the way it conveys data and is represented. These included size, value, texture, colour, orientation and shape. Kosslyn (1989) also highlighted these differences in the communication of information and identified four common types of displays, that is, graphs, charts, maps and diagrams. Mackinlay (1999) elaborated further and developed a classification of information graphics according to six types of ‘graphical languages’. Within each language are a group of graphics that use similar encoding techniques and perceptual elements. Like text-based languages, Lowrie and Diezmann (2009) asserted that “graphical languages have unique signs, symbols and characteristics” (p. 150). The six graphical languages are: Axis, Apposed-position, Retinal-list, Map, Connection and Miscellaneous. As Mackinlay (1999) clarified:

**Single position languages** encode information by the position of a mark set on one axis. **Apposed-position languages** encode information by a mark set that is
positioned between two axes. **Retinal-list languages** use one of the six retinal properties of the marks in a mark set to encode the information. Since the positions of the marks do not encode anything, the marks can be moved when retinal list designs are composed with other designs. **Map languages**, which have fixed positions, encode information with graphical techniques that are specific to maps. **Connection languages** encode information by connecting a set of node objects with a set of link objects. **Miscellaneous languages** encode information with a variety of additional graphical techniques (p. 75).

Using these definitions and encoding techniques, Lowrie and Diezmann (2005) identified examples of each of the graphical languages. They found that there was often a mismatch between the types of graphics utilised in curriculum documents and those presented in high-stakes test situations. Lowrie and Diezmann (2005) indicated that the Mackinlay (1999) definitions had much merit for mathematics classrooms since the explicit teaching of conventions could only help and support students’ decoding skill development.

From a mathematics assessment perspective, **Axis languages** are typically represented as number lines (usually on a horizontal axis) or temperature graphics (usually on a vertical axis) or measurement tools (usually represented on either axis) (see Figure 2.5).

![Figure 2.5. Example of an Axis language item.](image-url)
Apposed-position graphics combine features of both vertical and horizontal axes and are commonly represented as \( x \) and \( y \) axis. Mathematics examples include line graphs, bar graphs and column graphs (see Figure 2.6).

Figure 2.6. Example of an Apposed-position language.

Map graphics contain most of the Apposed-position conventions but are often represented from a different perspective (e.g., a street map as a bird’s-eye view or from a coordinate perspective). These graphics also contain other features and conventions that are often associated with the graphic but not necessarily embedded into the actual graphic (e.g., the legend on a map) (see Figure 2.7)
Retinal-list languages are often found in intelligence tests or instruments that measure spatial reasoning ability. These graphics typically require the decoder to translate, rotate, or reflect information on a point or a line (see Figure 2.8). The decoder generally conducts these transformations in the mind’s eye (Kosslyn, 1989).

Connection languages are typically decoded from a top down or bottom up convention (e.g., a flow chart or a tennis draw). These representations
sometimes include family trees and also contain directional arrows as part of the processing features (see Figure 2.9).

![Diagram](image)

*Figure 2.9. Example of a Connection language item.*

*Miscellaneous* languages do not represent conventions on points or lines—which is the case for the other five graphic categories. Consequently these graphics are often circular in nature and include pie charts and Venn diagrams (see Figure 2.10).

![Pie Chart](image)

*Figure 2.10. Example of a Miscellaneous language item.*
In an analysis of the inaugural Year 3 and Year 5 NAPLAN’s, Lowrie and Diezmann (2009) revealed that all graphical languages were represented across both tests with the following breakdown: Miscellaneous (38%), Retinal-list (29%), Map (8%), Axis (8%), Apposed-position (7%) and Connection (4%). The remaining 6% of graphical items contained contextual graphics. However what was surprising in their findings was that Apposed-position items (which included bar and column graphs) “were very much under represented despite the fact that they feature so predominately in school curricula” (p. 152).

With such a high prevalence of information graphics being utilised in national standardised testing, it is important to analyse the impact this may have on a child’s ability to solve mathematical tasks.

**Contextual Graphics**

The effects and use of contextual graphics within mathematics assessment is not as well documented as issues surrounding information graphics. As previously discussed, contextual graphics are those included in a task to complement the information contained within the text of the question and contain no information necessary to solve the task. In a study on the use of contextual graphics within American and Japanese textbooks, Mayer, Sims and Tajika (1995) refer to these graphics as irrelevant illustrations. Although little research has investigated the use of these ‘irrelevant’ kinds of graphics in an assessment context, the research surrounding contextual graphics within instructional settings and as a promotion of learning can be considered applicable to the context of this study.
According to Levie and Lentz (1982) there could be a number of reasons why contextual graphics or illustrations could be utilised in instructional text. These included, aiding in the interaction and interest of the text with the child by adding an element of enjoyment to their reading. They also may be particularly effective in providing “spatial information that is difficult to express in words” (p. 196). Both of these reasons are relevant when considering the use of contextual graphics in assessment. The use of a picture to illustrate an assessment task may add interest and aid in maintaining students’ focus and engagement within the test conditions. It may also assist poor readers in comprehending the text contained within the question through the provision of a reading prompt.

Shimada and Kitajima (2006) referred to these two processes as the motivation effect and the elaboration effect of incorporating illustrations within text. They argued that if an illustration is included, students are more likely to want to engage with the text rather than skim over it. Similarly they also noted that a picture can elaborate or assist in the comprehension of the text. Furthermore, research by Glenberg and Langston (1992) confirmed that not only do pictures aid and facilitate comprehension but also memory of the text, “even when the pictures add no new information” (p. 140). These authors concluded, contrary to other findings, that the effects of pictures on working memory were beneficial by reducing the need to identify referents within the text. They proposed that pictures help to build mental models of what the text is about.

Although these reasons for the inclusion of contextual graphics or illustrations seem to be relevant and valuable, some research has also indicated a positive effect on students’ results. In their review of the
research surrounding the effects of text illustrations on students’ learning. Levie and Lentz (1982) provided a number of guidelines for practices and possible avenues of further research. Firstly they noted that “illustrations can help learners understand what they read, can help learners remember what they read, and can perform a variety of other instructional functions” (p. 226). As a result, illustrations can provide a context and aid comprehension of associated text. This may be particularly beneficial in an assessment setting where illustrations can assist in creating a situation in which the task can be embedded. Subsequently this may also aid a child’s understanding of the situation through the use of visual cues.

However with such diversity in the classroom these positive findings need to be examined in light of student’s spatial ability. Although the use of pictures can assist in supplementing a text, unless children are able to effectively interact between the components and build appropriate mental models the pictures could be considered more of a hindrance than an asset. In Mayer and Sims’ (1994) study on the impact of pictures on high and low-spatial ability students, they found that the inclusion of illustrations benefited high-spatial ability students who were able to competently build mental connections between the verbal and visual representations. As a result, low-spatial ability students struggled to build connections between the two. Mayer and Sims (1994) concluded that “researchers need to examine more fully the role individual differences might play in multimedia learning” (p.400).

In a similar vein, Levie and Lentz’s (1982) final guideline of the inclusion of illustrations that was relevant to assessment was that “illustrations may be somewhat more helpful to poor readers than to good readers” (p. 226). This
is particularly relevant when considering students who have English as the second language (Ercetin, 2003). It is understandable why the inclusion of supporting illustrations would aid poor readers. However the benefit of their inclusion, particularly in mathematics assessment items, also needs to be considered for good readers. With an ever growing presence of graphics in assessment, both information and contextual, there not only exists the possibility of student’s not being able to comprehend the text but also the inability to discriminate when and when not to utilise the graphics incorporated in the question. For this reason the inclusion of illustrations may be detrimental for good readers who do not necessarily require the illustration for comprehension but may attempt to inappropriately incorporate it within their mathematical thinking.

It is this process of knowing when and how to use visuals included in a task that is considered to be visual literacy. Although this term has a number of definitions attached to it as noted by Avgerinou and Ericson (1997), these authors observed an accepted statement regarding the term visual literacy that is particularly relevant to mathematics assessment. This included the ability to use “visuals for the purpose of communication: thinking; learning; constructing meaning; creative expression; aesthetic enjoyment” (p. 284). The inclusion of contextual graphics within assessment items requires a student to be able to recognise and appreciate the graphics purpose without impacting on their mathematical thinking. A visually literate child will be able to effectively utilise all the aspects of the task including the interaction of the text and the illustration to construct meaning and knowledge.

The next component of mathematics assessments tasks that will be explored in light of the current literature is the use of language.
Language

Language versus literacy

Before an analysis of the role language plays in mathematics assessment can be addressed it is necessary to define the term language and its use within this study. Yet to understand the term language it is also important to recognise the term literacy, since they are often used synonymously (Anstey and Bull, 2000). However, they have noted a crucial difference between the two referring to language as a “system of signs and symbols used by a group or a society to communicate meaning” (p. 201) whereas literacy is defined as the “particular social practices that the group uses when employing the language” (p. 202). “Language is about signs and symbols whereas literacy is about events in which social practices occur” (Anstey & Bull, 2000, p. 202). These definitions are important especially when terms like ‘mathematical literacy’, ‘critical numeracy’ and ‘mathematics language’ have been used throughout research in the past with varying meanings and definitions attached (see Kaiser & Willander, 2005; Matteson, 2006; Pugalee, 1999). In regards to this aspect of the study, the focus is on the language used within mathematics assessment. Inevitably it is the way pupils come to terms with this symbol system and how they use this alongside the natural language system they may still be learning and are yet to master (Shorrocks-Taylor & Hargreaves, 1999). From the perspective of this study both language and literacy need to be considered. For example, language relates to the composition and structure of text within a given assessment task while literacy would be how the students engage with the text. The notion of mathematical literacy however will be addressed further on in the study.
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 (see, for example, Abedi, Lord & Plummer, 1995; Abedi, Lord, & Hofstetter, 1998; DeCorte, Verschaffel, & DeWin, 1985). In fact, their study supported findings of Carpenter, Corbit, Kepner, Linquist 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 student’s 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.

In Australia, it was Newman (1977; 1983) who developed a model highlighting specific reading skills crucial to performance on mathematical word problems. The model as outlined by Newman (1983) is divided into three major areas: (1) mechanical reading skills, which are used mainly to decode the material; (2) comprehension skills, which are associated with the purpose of reading, that is, reading for meaning; and (3) application skills which are involved in the ability to interact with, and use written material
successfully. Figure 2.11 gives a detailed summary of Newman’s (1983, p. 22) model.

A  Mechanical reading skills

1. Word and symbol identification
   - Functional words
   - Instruction words
   - Specific word terms
   - Symbols

2. Rate and style of reading

3. Directionality

4. Word analysis

5. Context clues

B  Comprehension skills

Level one: Literal comprehension
   - Reading to follow directions
   - Common words and numerals in mathematics
   - Gathering stated information
   - Noting a stated sequence

Level two: Interpretative comprehension
   - Interpreting purpose
   - Interpreting terminology
   - Ability to predict outcomes
   - Interpreting general meaning
   - Disregarding irrelevant information
   - Interpreting graphic material

Level three: Evaluative comprehension

C  Application skills

- Using a textbook
- Note taking

Figure 2.11. Newman’s model of reading skills for the language of mathematics.

Newman’s model is both compelling and useful for this study. Importantly, the model considers the mechanical function of reading skills which relate directly to construction of text within an assessment item (i.e., the language of composing a task). In addition the model describes comprehension skill which provides insights into the literacy aspects of the present study; that is,
the way in which students interpret text. The application phase of the model considers a utilisation aspect of reading which is highly appropriate to this study since the study is embedded within a regimented assessment framework i.e. the NAPLAN.

**Mechanical Reading Skills**

According to Chapman (1993) “particular kinds of activities require particular kinds of language” (p. 37). Halliday (1978) used the linguistic term ‘register’ to describe “a set of meanings that is appropriate to a particular function of language, together with the words and structures which express these meaning” (p. 195). Zevenbergen (2000a) argued that in order to successfully crack the code of the mathematics classroom and similarly with assessment, children must come to understand the register of mathematics.

Chapman (1993) described features of the register of school mathematics to include highly specialised vocabulary and the reinterpretation of existing words. Zevenbergen (2000a) also commented on the use of a very particularised and highly technical vocabulary, trigger words such as ‘more’, ‘less’, ‘got’ or ‘took away’, the semantic structure used within problems and the lexical density of the item. Similarly, Perso (2009) noted that this would include such code breaking as:

- English language words and phrases;
- words and phrases particular to mathematics;
- words and phrases from the English language that have a particular meaning in the mathematics context such as the word ‘flip’; and
• symbolic representations such as ‘4’ representing ‘four’ or ‘+’ representing ‘addition’, ‘plus’, ‘sum’, ‘increase’, ‘more’, ‘group’ and ‘combine’.

All these features, including the lexical density, “results in a high level of complexity in the translation of the problem” (Zevenbergen, 2000a, p. 207).

Of particular interest to research in the past has been children’s acquisition and use of the terms “more” and “less” in problem solving. Lean, Clements and Del Campo (1990) argued that although children start school with a general understanding of these terms, “for many school beginners their receptive and expressive understanding and use of the important “more-less” pair is shaky and still subject to modification as a results of feedback from others and of listening to how others (e.g. parents, teachers, and peers) use the words” (p. 168). This study’s findings were similar to those of Donaldson and Balfour (1968) where the positive member of the pair, in this case “more” was understood before negative members such as “less”. Other linguistic research found that when children encounter a word to which they do not fully understand, they will select words from the same semantic fields and substitute them for the unknown term (Clark, 1972).

It is therefore the initial ‘reading’ of the question or ‘reading’ of the data in which the child’s ability to solve the task lay. As noted by Fuentes (1998), “to improve their mathematics, we must improve their reading” (p. 81). As previously noted it is necessary for the student solving the task to understand the text and mathematics signs and symbols (the language).

Clements (1980) offered an example to highlight this relationship when a student from a sixth-grade class pointed to the question shown in Figure
2.12 and asked: “Sister, when it says here ‘Which angel is a right angel, does it mean that the wings should go this way or that way?’”

![Figure 2.12. The angel task utilised in Clements (1980).](image)

This is an instance where important words in a question have been misread and have prevented the child from proceeding any further and once again drawing the conclusion that reading ability plays a critical role in mathematics assessment. It is not surprising that research has also shown a correlation between poor performances on mathematics word problems with low-ability readers or ESL students (see Abedi & Lord, 2001; Abedi, 2002). It is for this reason that Matteson (2006) argued that “mathematics educators need to be aware of the role of reading as applied to standardized assessment, especially those of the multiple choice format” (p. 207).

However, as noted by Newman (1983), it is not just the case of reading ability but being able to subsequently comprehend it correctly and attach the appropriate meaning that is important.

**Comprehension Skills**

Newman’s three levels of text comprehension move from a basic literal understanding to a more complex and inferential application. It includes comprehension, transformation, process skills and encoding (White, 2009). This sequence is very similar to the one outlined by Curcio (1987) in
regards to graph comprehension. He moves from reading the data, to reading between the data and finally reading beyond the data. In both instances it is about moving beyond the reading of the words and symbols to being able to apply them appropriately.

To highlight the importance of reading comprehension, Clements (1980) provided the following example (see Figure 2.13).

![Figure 2.13. The comprehension task utilised in Clements (1980).](image)

After recording an answer of ‘15’, the child was asked to provide an explanation of how he came to that conclusion. He responded by saying “it says John goes to bed fifteen minutes later, so the answer must be ‘15’” (Clements, 1980, p. 2). Clements noted that the child could read the words but had been unable to grasp the meaning of all the information given in the question.

However the child also utilised an unsuccessful comprehension strategy categorised by Hegarty, Mayer and Monk (1995) as “direct-translation strategy” (p. 18). They argued that often unsuccessful problem solvers begin by selecting numbers and keywords from the problem and base their solution plan on these. Also referred to as “compute first and think later” (Stigler, Lee, & Stevenson, 1990, p. 15), the keyword method (Briars & Larkin, 1984), and number grabbing (Littlefield & Reiser, 1993), this shortcut approach results in the problem solver producing an answer based on the key propositions contained within the question rather than a
qualitative representation of the situation described in the problem (Hegarty, Mayer, & Monk, 1995).

Another problem experienced by students when attempting to comprehend word problems is the inability to identify appropriate semantic relations. Cummins (1991) argued that successful comprehension of the text not only included understanding the words but also the actual mathematical process required. Utilising the semantic model developed by Riley and Greeno (1988), Cummins (1991) identified three semantic types of standard problems: Combine, Change, and Compare. She described Combine and Compare problems as those that required static relation between quantities. That is, involving either the combination of two quantities to produce the answer or the comparison of two quantities and their difference quantified. In a Change problem, an exchange occurs that changes the size of a given set of objects. Of the three forms of questions, Lean, Clements and Campo (1990) found that Compare problems proved to most difficult for primary schoolchildren with children often misinterpreting the meaning behind ‘less’ and ‘more’. In many cases children replaced questions that they heard or read with semantically less complex and more familiar questions.

According to Riley, Greeno and Heller (1983) a proficient problem solver is able to acknowledge these semantic relations and correctly link them to appropriate solution sequences. They hypothesise:

The acquisition of <problem solving> skill is primarily an improvement in children’s ability to understand problems—that is, in their ability to represent the relationships among quantities described in the problem situation in a way that relates to available solution procedures. (p. 173)
Clements (1980) referred to this process as the ability to “transform” from the written problem to an acceptable set of mathematical procedures. It is for this reason that the analysis of semantic structures has proved to be invaluable to researchers in understanding the processes necessary for successful and unsuccessful problem solving (see Cummins, Kintsch, Reusser & Wener, 1988; Kintsch & Greeno, 1985).

According to Fuentes (1998), another cause of comprehension failure is when the reader interprets the message consistently but the interpretation is different from the one intended by the author. Therefore it is as much a product of the characteristics of the reader as of the text (Shorrocks-Taylor & Hargreaves, 1999). It is this notion of the meaning being constructed by the child which leads to the notion of ‘mathematical literacy’.

**Mathematical Literacy**

“Mathematical literacy is defined in PISA (Programme for International Student Assessment) as the capacity to identify, understand and engage in mathematics, and to make well-founded judgments about the role that mathematics plays in an individual’s current and future private life, occupational life, social life with peers and relatives, and life as a constructive, concerned and reflective citizen” (Kaiser & Willander, 2005, p. 49). This concept of mathematical literacy has been gaining momentum and status since the late 1980’s when Cooper (1998b) noted a change in pedagogy in the English school system. According to findings by Her Majesty’s Inspectorate, it was recommended that mathematics become concerned with ‘real-life’ applications in an attempt to break down the barriers between mathematics in the classroom and the outside world. Even the highly influential have identified mathematical literacy as:
…an individual’s capacity to identify and understand the role mathematics plays in the world, to make well-founded judgements, and to use and engage with mathematics in ways that meet the needs of that individual’s life as a constructive, concerned and reflective citizen (OECD, 2004, p. 5, cited in Council of Australian Governments, 2008, p. 4).

However two major issues have resulted from the inclusion of ‘real life’ scenarios in the mathematics classroom. Firstly how does this affect the validity of assessment especially considering the diversity of each child and subsequently their construction of meaning? And secondly, how realistic are these ‘real life’ mathematics problems?

**Situations**

**How real is real?**

A situation can be defined as a ‘real life’ scenario used within a mathematics question. This concept has been defined in previous research as the use of ‘contexts’ and is described as a situation in which a problem is embedded (Sullivan, Zevenbergen, & Mousley, 2001). The use of contexts have been utilised in classrooms in an attempt “to motivate, to illustrate potential applications, as a source of opportunities for mathematical reasoning and thinking, and to anchor student understanding” (Sullivan et al., 2003, p. 109). Although this study does not consider task design in relation to assessment, Sullivan et al. maintained that the same teaching task can end up being fundamentally different depending on the manner in which it is presented and the way in which the cohort (students) interact and engage with the task (students). However in this investigation the context or situation is only one aspect of how the task is constructed—with the task
being the actual assessment artefact rather than the activity or lesson plan described above.

It has been illustrated that the use of graphics and language also contribute to what will be referred to as the ‘item context’ (as described in Figure 2.2 in the previous section). In this case the item context will examine the relationship between the graphics, language, design and situations and the impact they have on students’ results, thus, what students bring to the artefact.

According to Boaler (1993a), the effects on student performance of the use of situations has been widely underestimated for many years. Not only has it failed to be recognised for its impact on performance but even its ability to influence a child’s choice of mathematical procedure. Boaler (1993a) illustrated this effect, citing Taylor’s (1989) research study comparing students’ responses to two questions on fractions: one asking the fraction of a cake that each child would get if it is shared equally between six; and one asking the fraction of a loaf to be shared between five. She noted:

One of the four students in Taylor’s case study varied their methods in response to the simple variation of the word ‘cake’ and ‘loaf’. The cake was regarded by the student as a single entity to be divided up in sixths, the loaf of bread was regarded as something that had to be divided into a certain number of slices – the student therefore had to think of the bread as cut into a minimum of ten slices with each person getting two tenths of the loaf (p. 12).

Boaler (1993a) concluded by suggesting that “students interact with the context of a task in many different and unexpected ways and this interaction is, by its nature, individual” (p. 22). This revisits the notion of mathematical literacy and the role it plays in a child’s understanding and interpretation of the test item. The issue therefore is “not with the demonstration that some
underlying competencies may not express themselves in certain assessment situations but as another demonstration of the ways in which context constrains/facilitates individuals’ mathematical productions and the implications of this for fairness and validity in testing” (Cooper, 1998, p. 528).

Another difficulty in creating perceptions of reality according to Boaler (1993a) was “when students are required to engage partly as though a task were real whilst simultaneously ignoring factors pertinent to the ‘real life version’ of the task” (p. 15). One highly published example of this was that of French and German researchers (Institut de Recherche sur l’Enseignement des Mathematiques de Grenoble, 1980, cited in Greer, Verschaffel, & DeCorte, 2003) who posed problems such as “There are 26 sheep and 10 goats on a ship. How old is the captain?” and found that many children supplied answers by using arithmetical operations on the numbers in the text.

Greer et al. (2003) noted that these findings and many others suggested “that many children in elementary school develop the very specific beliefs that a word problem should be solved by adding, subtracting, multiplying, or dividing the two numbers supplied, and that the choice may be guided by superficial cues, such as the presence of ‘key words’ in the text” (p. 275). So while we endeavour to coat mathematics assessment items with a “thin-veneer of ‘real-world’ associations” (Boaler, 1993b, p. 343) children have difficulty perceiving mathematics as any more than arithmetic operations. Similarly in their theoretical discussion on the impact of English word problems on Filipino students, Verzosa and Mulligan (2012) identified a tendency for students to be guided by the “minimal effort principle” (p. 4).
Within this principle student answers are a reflection of “pseudo-analytical” thinking based on the classroom or teachers expectations rather than the word problem’s association to real life.

Verschaffel, De Corte and Lasure (1994) identified two reasons why students had a tendency to ignore real-world knowledge and they lay in the way school word problems were addressed in the classroom:

1) the impoverished and stereotyped diet of standard word problems, which can always be modelled and solved through the straightforward use of one or more arithmetic operations with the given numbers;

2) several features of the current instructional practice and culture, such as the premature imposing of the formal arithmetic approach towards arithmetic word problem solving by requiring that pupils must identify the correct arithmetic operation to solve a word problem, and—more generally—the lack of systematic attention to the modelling perspective as one of the building blocks of a genuine mathematical disposition (p. 274).

So while teachers are being encouraged to present mathematics in realistic settings, they have trouble presenting concepts as anything but an arithmetic problem with a ‘real-world’ frill. It can therefore not be assumed that accepted content based schemes that offer numerous ‘real world’ examples will automatically mean a successful transfer to ‘real world’ situations (Boaler, 1993b). However, Van Den Heuvel-Penhuizen (2005) argued it is the inappropriate use of the word ‘realistic’ in these word problems that needed to be addressed.

According to Van Den Heuvel-Penhuizen (2005) the term ‘realistic’ has normatively been employed as a requirement that the teaching and the problems used are authentic and reflect real-life situations however this is an incorrect use of the term. She believed that being ‘realistic’ is about making
something real in your mind either from the student’s real world or fantasy world. “The fantasy world of fairy tales and even the formal world of mathematics can provide suitable contexts for a problem, as long as they are real in the student’s minds and they can experience them as real for themselves” (p. 2).

She argued that many word problems are often merely number problems that are ‘dressed up’ with a cosmetic reality. In these circumstances the situation is not even necessary and can be exchanged for another without substantially altering the problem. For instance, Van den Heuvel-Panhuizen (2005) used the marble and ham examples (see Figures 2.15 and 2.16). She highlighted problems involving marbles, in which someone has 16 marbles and gains 10 more, might just be changed to a problem involving pounds of ham.

**Figure 2.15.** Marbles problem.

```
Jim has 16 marbles and wins 10 more.
How many does he have now?
```

**Figure 2.16.** Ham problem.

```
The butcher has 16 pounds of ham in his shop and orders 10 pounds more.
How much does he have now?
```

Moreover, Van den Heuvel-Panhuizen (2005) noted that the reality used in these word problems is not in tune with the real situation being presented. For example, in the situation of the butcher’s shop, some of the hams in stock might have been sold by the time the new ham arrives. Boaler (1994)
also found that one of the most significant problems of the use of word problems occurs “when students are required to engage partly as though a context in a task were real whilst simultaneously ignoring factors pertinent to the ‘real life version’ of the task” (p. 555). This means that children may offer a common sense answer, which would be marked as ‘incorrect’, based on their real world understanding compared to one that required them to suspend reality and answer according to ‘mathsland’. As William (1992, cited in Boaler, 1994, p. 555) observed:

> Over the last eight years, I have visited a lot of mathematics classrooms, and it seemed to me that in most of them, it was as if there were a kind of check-in desk just outside the classroom labelled ‘common sense’, and as the pupils filed into the classroom, they left their common sense at the check in desk saying “Well we won’t be needing this in here”.

Boaler (1994) concluded that ‘Mathsland’, the fantasy world created by many mathematics questions, is likely to be most harmful to those students who are socially aware and concerned about the relevance of subjects in their future lives” (p. 557). Therefore it is important to reflect on the question, if current assessment word problems are just numeric equations ‘dressed up’, would it be more beneficial for all students to strip these questions and reveal their ‘naked truth’? According to Van den Heuvel-Panhuizen (2005), no.

While acknowledging the unresolved issues of the use of ‘real-life’ assessment items including the possibility of hindering a child’s answer, students’ unwillingness to take into account the ‘real-life’ scenario and the need to often suspend reality, she has found that real-life problems when compared with bare problems produce better results. Within her research, problems were presented both as a real-life problem and as a bare number
problem. It was found that large discrepancies in the achievement scores emerged, for example the *Banana* problem (see Figure 2.17). Within this question fifth-grade students (aged 10-11 years) were asked to give an estimate of $1.49 \times 0.740$. The question was asked in two ways: as a bare number problem and as a ‘real-life’ problem. According to Van den Heuvel-Panhuizen the results “were remarkable. The bare version was answered correctly by only 4% out of the 26 students, whereas 46% of the students came up with the correct answer to the (real-life) version of the problem” (p. 8).

These findings mirrored those of Clements (1980) who also paralleled bare arithmetic problems with ‘real-life’ and found that the use of imagery helped support children in achieving better results (see Figure 2.17).

*Figure 2.17. The banana problem.*

This research highlighted the need for careful analysis of the use of situations to create ‘real-life’ situations in large-scale assessment. This included the need to consider children’s different levels of mathematical
literacy, the reality behind ‘real-life’ test items and further study into the comparison of situation and situation-free mathematical test items. However the situation, along with the graphics and language, have so far only been discussed in isolation within this current study. The final aspect that needs to be considered is the way all of these parts are presented as an item and the impact this may have on a child’s ability to answer the question.

**Item Features/Design**

Item writing is an art. It requires an uncommon combination of special abilities. It is mastered only through extensive and critically supervised practice. It demands, and tends to develop, high standards of quality and a sense of pride in craftsmanship (Ebel, 1951, p.185, cited in Rodriguez, 2005, p. 3).

Item writing is indeed an art. 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 (Hancock, 1994; Rodriguez, 2003; Threlfall, Pool, Homer, & Swinnerton, 2007). 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).

As the mathematics NAPLAN is largely made up of multiple choice questions compared to single answer responses (in the 2010 Year 5 booklet 73% were multiple choice), a more detailed analysis of this use of question will now be explored.

**Multiple choice**

According to Haladyna and Downing (1989a) there are a number of reasons why testing organisations prefer multiple choice formats:

1. Sampling of content is generally superior when compared to other formats; the use of MC formats generally leads to more content-valid test-score interpretations.

2. Reliability of test scores can be very high with sufficient numbers if high-quality MC items.

3. MC items can be easily pretested, stored, used, and reused, particularly with the advent of low-cost, computerised, item-banking systems.

4. Objective, high-speed test scoring is possible.

5. Diagnostic subscores are easily obtainable.

6. Test theories (i.e., item response, generalisability, and classical) easily accommodate binary responses.
7. Most kinds of content can be tested using this format, including many types of higher level thinking (p. 37).

For these reasons, multiple choice is the preferred format for most questions contained within the Australian mathematics NAPLAN including those conducted in high school with 78% of the Year 9 2010 NAPLAN being multiple choice. However it is interesting to note that by Year 12, the 2010 Higher School Certificate (HSC) for General Mathematics allocated 22 marks to multiple choice and 78 marks to the working out of constructed-response questions. In the 2010 Mathematics HSC, there were no multiple choice questions at all with the whole test presented in constructed-response item format with the opportunity to gain marks based on working out. Of particular interest was the absence of open-ended tasks in all mathematics standardised assessment with questions allowing only one correct response. In both instances it is important to note that the item features (see Figure 2.2) of the assessment tasks are being altered through levels and years of schooling. The higher levels of mathematics resist multiple choice task features and no high-stakes testing provided opportunities for open-ended engagement. There is a question then of the extent to which each type of mathematics is being valued.

According to Bormuth (1970, cited in Haladyna, 1999) item writing is not a science but rather a collection of guidelines captured in textbooks based on experience and the wisdom of mentors. It was for this reason that Haladyna and Downing (1989) developed a taxonomy of 43 multiple choice item-writing rules from within 46 textbooks and other sources, presenting an array of advice from as early as 1935 (see Table 2.1). Haladyna (1999) noted that there was author consensus for many of these guidelines however
they were not mutually agreed upon. It was for this reason that a second study by Haladyna and Downing (1989b) involved an analysis of more than 90 research studies testing the validity of these item-writing guidelines. Yet only a few of these guidelines received extensive study, resulting in a reprised study (Haladyna, Downing, & Rodriguez, 2002) of 23 new textbooks and a revision of the original 43 guidelines.

It is Haladyna’s (1999) 30-point checklist of item writing guidelines that are, according to Connolly (2011), used for the training of NAPLAN item writers and also as a set of consistent standards to judge items by (see Table 2.1). He also noted the use of the Educational Testing Service (ETS) International Principles for Fairness Review: Guidelines (2007). This is the major testing agency in the United States that publishes a set of principles for fairness designed to avoid bias at a content level and at item review.

Table 2.1
Guidelines for Writing Multiple Choice Items (Haladyna, 1999, p. 77)

<table>
<thead>
<tr>
<th>CONTENT CONCERNS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Base each item on specific content and a type of mental behavior.</td>
<td></td>
</tr>
<tr>
<td>2. Keep the specific content of items independent from one another.</td>
<td></td>
</tr>
<tr>
<td>3. Avoid overly specific and overly general knowledge.</td>
<td></td>
</tr>
<tr>
<td>4. Focus each item on a single behaviour instead of a chain of behaviours.</td>
<td></td>
</tr>
<tr>
<td>5. Avoid opinion-based items.</td>
<td></td>
</tr>
<tr>
<td>6. Avoid trick items.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FORMATTING CONCERNS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Avoid true-false and complex MC formats.</td>
<td></td>
</tr>
<tr>
<td>8. Format the item vertically instead of horizontally.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STYLE CONCERNS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Edit and proof items.</td>
<td></td>
</tr>
<tr>
<td>10. Keep vocabulary simple for the group of students being tested.</td>
<td></td>
</tr>
</tbody>
</table>
11. Use correct grammar, correct punctuation, capitalization, and spelling.

12. Minimise the amount of reading in each item.

WRITING THE STEM

13. Use either a question stem or a partial sentence.

14. Ensure that the directions in the stem are very clear.

15. Include the central idea in the stem instead of the choices.

16. Avoid window dressing (excessive verbiage).

17. Word the stem positively, avoid negatives such as NOT or EXCEPT.

WRITING THE CHOICES

18. Use as many good choices as possible, but three seems to be a natural limit.

19. Make sure that only one of these choices is the right answer.

20. Vary the location of the right answer according to the number of choices.

21. Place choices in logical or numerical order.

22. Keep choices independent; choices should not be overlapping.

23. Keep choices homogenous in content.

24. Keep the length of choices about equal.

25. Avoid using none-of-the-above, all-of-the-above, or I don’t know.

26. Phrase choices positively; avoid negatives such as NOT.

27. Avoid giving clues to the right answer, such as
   
   a. Specific determiners including always, never, completely, and absolutely.
   
   b. Clang associations, choices identical to or resembling words in the stem.
   
   c. Grammatical inconsistencies that cue the test-taker to the correct choice.
   
   d. Conspicuous correct choice
   
   e. Pairs or triplets of options that clue the test-taker to the correct choice.
   
   f. Blatantly absurd, ridiculous options.

28. Make all distractors plausible.

29. Use typical errors of students to write your distractors.

30. Avoid humorous choices.
Of particular relevance to the present study are five features of the guidelines, namely: 10, 18, 23, 28 and 29. Although each of these 5 features, and indeed the entire 30 presented in the guidelines, relate to the item features, they have connected links to other aspects of task construction. For example the notion of keeping choices homogeneous in content fails to recognise the challenge of doing this given the impact situations have on students’ reasoning. Furthermore, keeping the vocabulary simple can be more problematic than first appears, especially since reduced vocabulary in essence results in each word having more importance and significance than they would in extended abstracts. It is also the case that having all distractors as plausible results in students ‘reading’ too much into any given context.

It is interesting to note that option development received the most emphasis in terms of rules, with 13 of the 30 relating to this aspect of multiple choice item writing. In fact a growing amount of research has been conducted into the number of options that should be presented within multiple choice items.

**Three, Four or Five Options?**

It is recommended within Haladyna and Downing’s (1989) taxonomy outlined in rule 24 to use as many options as feasible, however there has been growing evidence that fewer options are actually more desirable. In Rodriguez’s (2005) meta-analysis of over 80 years of research on item development he concluded that three answer choices are optimal for multiple choice items. According to Beddow, Elliott and Kettler (2009), Rodriguez’s results indicated that reducing items from 4 or 5 answer choices to 3 tended to result in:
small, insignificant changes in item difficulty, discrimination, and reliability, and such modification are sensible in that the reduced reading load results in a test that is more manageable for a group of students who typically take much longer to finish. Likewise, the decreased reading load may permit test developers to add more items and increase the reliability of the test (p. 3).

They concluded that reducing the number of response options from 4 to 3 eliminated an implausible distract and resulted in a balanced set of answer choices that were free of cueing. While Haladyna (1999) argued that one should write as many good distractors as one can, it should be expected that only one or two will really work as intended.

However, as noted by Haladyna (1999), test developers are often faced with the dilemma of creating a test according to predetermined standardised testing guidelines of 4 or 5 options. This may result in the existence of nonperforming distractors as nothing more than window dressing.

The four components of mathematics assessment tasks have been individually examined in light of the relevant literature to gain an appreciation of their use within current assessment practices. Nevertheless, these four components need to also be considered holistically since the interplay between some or all of these components has an effect on the manner in which task are decoded (Kosslyn, 1985; Mayer, 1985).

**Relationship/Connectivity of Components**

In the following section relevant psychological and sociological literature are considered in relation to one’s capacity to interpret graphics and text simultaneously. Theoretical foundations associated with dual coding, cognitive load theory, empowerment and social capital are addressed.
Dual Coding

According to Schnotz and Bannert (2003) there has been much research surrounding multiple external representations that has focused primarily on combinations of texts and pictures (see Levie & Lentz, 1982; Levin, Anglin, & Carney, 1987). They subsequently reported “that text information is remembered better when it is illustrated by pictures than when there is no illustration” (p. 142). Mayer and Anderson (1992) also found that pictures and words are most effective when they occur “contiguously in time and space” (p. 444). These findings were strongly aligned with Paivio’s dual coding theory (Paivio, 1986; Clark & Paivio, 1991). According to this theory,

...verbal information and pictorial information are processed in different cognitive subsystems: a verbal system and an imagery system. Words and sentences are usually processed and encoded only in the verbal system, whereas pictures are processed and encoded both in the imagery system and in the verbal system. Thus, the memory-enhancing effect of pictures in texts is ascribed to the advantage of a dual coding as compared to single coding in memory (Schnotz & Bannert, 2003, p. 142).

It could therefore be assumed that the addition of a contextual graphic or illustration within an assessment item could be beneficial for the students by making it easier to decode and comprehend provided that the graphic is appropriately represented (Lowe & Promono, 2006). Yet examples can be found of tasks in which two different processes compete to the detriment of each other (e.g., Kirby, Jurisich, & Moore, 1984). These ‘competitive effects’ Kirby (1993) noted, comprise limits on the benefits of dual processing. Subsequently he also recognised several other boundaries and possible limitations that need to be recognised in regard to dual coding theory. These are:
1. Task Interference—especially within complex and time restricted tasks if the dual forms of mental processing (verbal and spatial) are not automated responses and are thus competing for the limited executive resources.

2. Focus of Attention—additional processing, either verbal or spatial, may affect a learner’s inferences about what is important and misdirect their attention and focus.

3. Individual Differences—such things as natural ability, prior knowledge, strategies and interests that will impact on a child’s preference and effective use of dual coding. Consequently, as noted by Schnotz and Bannert (2003), “adding pictures to a text is not always beneficial, but can also have detrimental effects on the construction of task appropriate mental representations” (p. 153).

Similarly Mayer and Moreno (2003) noted three assumptions on how the mind works within multimedia learning that are relevant to the inclusion of graphics and language within mathematics assessment task. Multimedia learning is defined as learning from words and pictures but includes printed and spoken words as well as the inclusion of static or dynamic illustrations. Mayer and Moreno’s (2003) assumptions are closely related to those outlined by Kirby (1993) and concentrate on the dual or separate information processing channels required for verbal and visual information. Mayer and Moreno (2003) also commented on the disproportionate amount of processing capacity available in the verbal and visual channels compared to the cognitive load required in constructing meaning. They argued that often this limited capacity of the mind is overloaded by the active processing necessary when attempting to comprehend multimedia learning.
Although previous research suggests that processing of texts with text illustrations has profited poor readers in regard to comprehension and learning (Levi & Lentz, 1982; Rusted & Coltheart, 1979), the reality, according to Scaife and Rogers (1996), is that little is known about the cognitive value of any graphical representations. It appears in the light of the research, that when considering the addition of a picture within a text it is important to ensure that it is relevant and whether it supports the construction of a task-appropriate mental model.

**Cognitive Load Theory**

Cognitive load theory (CLT) is becoming increasingly relevant in mathematics considering the highly graphical nature of assessment tasks and the way mathematics is taught in the classroom. This is particularly in light of much of the research surrounding multimedia learning (Glenberg & Langston, 1992; Mayer & Moreno, 2003; Mayer & Sims, 1994). As topics are introduced through the use of more than one representation (e.g., text and pictures), greater cognitive demand is placed on the student in order to understand and decode the relevant information. In the initial learning phase in particular, Sweller (1999) noted that the high level of element interactivity in an item may cause difficulty in manipulating the elements of the task due to the limited working memory available. He argued that “each element and its relations to all the other elements must be considered and working memory may be unable to accommodate the various processes required” (p. 36). So how does this impact on the design of mathematics assessment tasks? What implications can this have on student achievement and results? While research on cognitive load theory has primarily focused on its effects on instructional design, its implications can also shed light on
These questions of assessment design (Threlfall, Pool, Homer, & Swinnerton, 2007).

Cognitive load theory, as outlined by Sweller (1999), is:

...heavily concerned with the consequences of limited working memory on instructional design. The theory assumes that when dealing with material that has high element interactivity, working memory limitations should be a primary consideration of instructional designers and that a basic difference between good and poor instructional design lies in the requirements imposed on working memory (pp. 36-37).

Subsequently, according to Mousavi, Low and Sweller (1995) “researchers have used CLT to suggest that many commonly used instructional procedures are inadequate because they require learners to engage in unnecessary cognitive activities that impose a heavy working memory load” (p. 319). This notion was examined in light of multimedia instruction and the impact of inappropriate and unnecessary language and graphics. Within this research Mayer and Moreno (2002) identified four design principles to aid students in learning more deeply and preventing the overloading of their visual and/or verbal working memories. These included the notion of contiguity, coherence, modality, and redundancy. Of particular relevance to mathematics assessment was the coherence principle where it was found that students “learn more deeply when they do not have to process extraneous words or sounds in verbal working memory or extra pictures in visual working memory” (Mayer & Moreno, 2002).

Similarly with regard to assessment, Beddow, Elliott and Kettler (2009) argued:

application to CLT focuses on certain aspects of test items that may require the test-taker to use cognitive resources in excess of those needed to demonstrate knowledge of the tested content. Excessive cognitive
load may be reduced by eliminating extraneous material from items, reducing reading load, and using visuals efficiently (p. 539).

In order to alleviate the demand placed on working memory, Sweller (1999) highlighted the need for instruction to be designed to minimise any unnecessary burdens on working memory and “maximise the opportunity for the acquisition and development of automated schemas” (p. 37). It is the development of these schemas that are particularly relevant to assessment items.

**Schemas**

A schema is a “cognitive construct that organises the elements of information according to the manner with which they will be dealt” (Sweller, 1994, p. 296). The relevance of such schemas in regard to mathematics assessment lies in its ability to “both store knowledge in long-term memory and, when brought into working memory effectively bypass the limited capacity of working memory” (Tindall-Ford, Chandler, & Sweller, 1997, p. 258). By combining complex and generalised knowledge structures into single elements the student is enabled to focus on many aspects of an assessment task with little impact on working memory. This of course will only be made possible if the schema is automatic in nature and has been learnt and practiced repeatedly over a period of time compared to one that still requires conscious thought and control. Automated schemas allow the processing of large amounts of information in working memory that would otherwise be considered too overwhelming.

Therefore unless a child has developed relevant schema on the interaction between a particular graphic and statement, searching for relations between the two requires working memory resources. As noted by Kirschner (2002),
skilled performance consists of “building increasing numbers of increasingly complex schemas by combining elements consisting of lower level schemas into higher level schemas” (p. 3). Consequently children that have practised and used graphics and text in the classroom would find a graphical item easier to solve than those with limited experience. Once children have mastered the use of both graphic and text within an item, the combination of the two within a mathematical assessment item will decrease cognitive load as appropriate schemas are already established. This highlights the need not only for instruction on the use of graphics but the important role text and language plays in children’s understanding and performance.

_Social and Individual Capital: Viewpoints from Bernstein and Bourdieu_

According to White (2011), Australia is regarded as one of the most multicultural countries in the world, second only to Luxembourg, and equal with Switzerland. And yet while we are quick to celebrate this diversity as a country, it fails to be acknowledged in national assessment particularly with the use of ‘real life’ mathematics items. And these differences not only lay in a child’s culture but also their socioeconomic background. It is therefore beneficial according to Cooper (1998b) and Zevenbergen and Lerman (2001) to employ Bernstein (1996) and Bourdieu’s (1984) work on class and culture in the exploration of a possible set of unintended consequences that might arise from the use of this particular type of test item in large-scale testing contexts.

Bernstein’s work is particularly relevant when analysing children’s responses and the rules that they employ according to their social
background. Cooper (1998b) noted from Bernstein’s (1996) recently published discussion that “middle-class children were much more likely to offer reasons which had an indirect relation to a specific material base and that the working-class children were much more likely to offer reasons which has a direct relation to a specific material base” (Bernstein, 1996, p. 34, cited in Cooper, 1998b, p. 518). This means that middle-class children recognise situations that require esoteric mathematical knowledge while working-class students select more generic understandings. Subsequently ‘real life’ test items may “discriminate against working class testees as a consequence of these children’s tendency to interpret formal educational contexts as legitimate arenas for the employment of everyday knowledge when such use is actually “inappropriate”” (Cooper, 1998b, p. 511).

To illustrate this concept Cooper (1998b) analysed the responses of a child classified as working class (named Mike) and another from a middle class family (named Diane) using an item from the 1994 English National Curriculum assessment (SCAA, 1994, cited in Cooper 1998b, p. 516) (see Figure 2.14).
Mike’s answer was based on his understanding of what kinds of socks boys wore rather than utilising the information contained within the item. This was compared to Diane who did note a higher presence of boys at her school in her real-life situation but was able to focus on the facts contained within the question and concluded “although it’s the same proportion, there are more girls” (Cooper, 1998b, p. 516). It was found that the student from the professional middle-class family ‘Diane’ was able to avoid the use of inappropriate real-world experience, although she did refer to it, compared to ‘Mike’ from a working class family who failed to avoid the trap. Although it is conceivable that England has a more structured class society than that of Australia, the distinction between children who have high
access to resources and capital and those students who typically come from backgrounds of less opportunity remains relevant.

Cooper (1998b) acknowledged that such dichotomous observations are purely illustrative and cannot be claimed to prove any case, nevertheless the differences need to be noted with regard to assessment item construction especially since other research studies have reported similar findings.

Specifically, Cooper and Dunne (1998) found a greater difference between service and working class students’ means due to the inappropriate use of everyday knowledge. Subsequently this led to particular groups of students being disadvantaged in terms of access to language that resulted in an underestimation of their mathematics competence. Similarly, Iversen and Larson (2006) noted “the influence of students’ task-specific attitudes, beliefs and motivation on performance” (p. 290). In addition, Boaler (1993b) found that the individual nature of a student’s construction of meaning was an important factor. She concluded that “contexts should not be randomly assigned to assessment tasks without consideration of the effect that these have upon students’ choice of methods and the extent to which a student’s mathematical understanding, with relation to context, is being assessed” (p. 369).

In a study exploring the use of different types of situations in mathematics including low fantasy, high fantasy, adult’s real world and children’s real world, Weist (2001) found that they impacted significantly on a number of variables. These included “the children’s interest in, attentiveness to, and willingness to engage with problems, the strategies they used, their effort, their perception of and actual success, and the extent to which they learned the intended mathematics” (Sullivan, Zevenbergen, & Mousley, 2003, p.
Therefore student dispositions play an important part in how students engage with tasks. From a high stakes testing perspective, some students will be less engaged than others due to a range of factors including self-esteem, confidence and readiness (especially for Grade 3 students). Consequently it is difficult to establish the extent to which particular words and phrases create situations and the context in which students engage with the tasks.

It is for this reason that Cooper (1998b) included the work of Bourdieu (1984) in regard to the use of ‘real-life’ contexts in mathematics assessment.

Fundamentally, Bourdieu employed a:

concept of habitus to describe the tendencies people have to respond to particular contexts in habitual, if not predictable, ways. Habitus is rooted in individuals’s socioeconomic and cultural experience (Cooper, 1998b, p. 526).

A child’s habitus is their habitual way of being and doing and is the result of the social, cultural, financial and physical influences on their lives. According to Zevenbergen (2000a) it is “the embodiment of culture, and (it) provides the lens through which the world is interpreted” (p. 202). Therefore children are not blank slates when entering the school environment and already have many well developed ideas and concepts of the world including the use of a particular language. As noted by Zevenbergen (2000a)

As students come to hear and use particular forms of language, this language becomes embodied to constitute a linguistic habitus. When students enter mathematical classrooms, they have accepted the language of their home environment, the consistency of which will vary with respect to formal school language (p. 202).

Although it is important to recognise such differences, it should not be treated as a deterministic factor of student achievement but rather to identify
the existence of differences between home and school languages and the impact these may have on students’ performance.

Bernstein and Bourdieu’s findings illustrate the strong influence of social class background and culture on mathematical literacy levels of a child. Therefore as noted by Cooper (1998b) this is one of many factors that need to be considered as part of a fuller account of children’s problems with ‘realistic’ test items. Boaler (1993b) raises the very important notion of how ‘real’ these items are by pointing out that one person’s meaning of a phrase or a word can be different to another’s based on their previous experiences as well as their access to language conventions. These ideas are elaborated on in more detail in the following section where Boaler (1993b) raises issues about the use of realistic situations and the extent to which these situations can ever be authentic.

**Summary: The Fantastic Four of Mathematics Assessment**

In today’s technological age, information is represented in visually rich modes, with quite different language, context and graphics demands. Multiple representations in graphs are regularly found in computer games and in the media. It is not surprising that recently introduced national instruments have followed a similar theme. This study will examine the four components of assessment items (i.e., graphics, language, situations and item features) and the impact they have on student performance. Just like the ‘Fantastic Four’ of Marvel comics fame, each component has its own unique ‘power’ that makes it strong enough to stand alone but as a group can be a force to be reckoned with.
Implications for the Present Study

The literature review has drawn attention to theoretical and practical issues which affected the design, instrument selection, methodological considerations and data analysis for the present study. The following points were specially noted and taken into consideration in this study:

1. Assessment items have attempted to become more ‘realistic’ in nature and this has influenced their subsequent design. This has resulted in four distinct components of test items including language, graphics, situations and design.

2. The need to consider the individual characteristics of students when attempting to measure the effectiveness of the language or graphics included within an item.

3. The problematic nature of attempting to include ‘real-life’ situations within mathematics assessment.

4. Current mathematics assessment practices emphasise a correct or incorrect response with no opportunity for children to explain their mathematical reasoning.

5. Present research has focused specifically on the impact of certain aspects of mathematics assessment on students’ results but has not embraced its accumulative affect.

The following chapter will outline the methodology, methods and instruments used in light of the literature and theoretical perspectives.
Chapter 3: Methodology

Introduction

The purpose of this chapter is to describe and explicate the design and methods used within this investigation. The chapter begins with the investigation’s four research questions and the research design by introducing external and internal representations. Following this the site and participants of this study are introduced. The methods utilised in developing and executing this study are recounted and examined in light of the design process. The process of the pilot study is elaborated upon leading into the methods used in the development of the instruments. These procedures are detailed throughout the remaining part of the chapter.

Research Questions

The following research questions emerged from the literature reviewed within the preceding chapter.

1. From a selected number of often used mathematics tasks, which do students find most difficult to solve?
2. How does the use of graphics, language, situations and item construction impact on a child’s capacity to make sense of mathematics?
3. To what extent does assessment measure sense making and mathematical understanding?
4. What effects will these findings have on classroom teaching?
Research Design

The methods used in the study were selected to inform and answer the research questions seeking depth and substance in an exploration of student performance in mathematics assessment tasks. The NAPLAN “instrumentation” was used as the baseline for investigation since the construction of items incorporated graphics, language, situations and item features. Moreover, its items were representative of current principles of assessment tasks. After an analysis of the pilot data, an experimental design was conducted which incorporated modified NAPLAN items as a way of measuring the influence of item structure on children’s mathematical understanding. This was further elaborated upon with a mixed method design through the incorporation of both qualitative and quantitative data collection techniques. The sequential nature of the design provided opportunities for further in-depth analysis of children’s sense making through three instrumental case studies. The implications of the findings on classroom teachers are summarised within a model. At the heart of the research design was the need to explore children’s internal and external representations as outlined in the following literature.

Internal and External Representations

This inquiry involved ‘looking beyond the answer’ and therefore required a different lens to focus upon assessment. Currently results from the NAPLAN are reported to parents as individual student reports and to schools as summative reports containing extensive information including results for each item and for each student. According to White and Anderson (2011) these school reports:
enable teachers to analyse the results for each year group to determine which items appear to be understood and which are problematic. In addition, school data can be compared to the Australian student data (p. 777).

Yet while these reports are important, especially in addressing common errors as well as to aid planning and programming of future learning (Perso, 2009), they fall short of providing any real insight into a child’s mathematical thinking. They also fail to acknowledge that a correct answer is not necessarily indicative of sound mathematical understanding.

The need to understand a child’s mathematical thinking has led to a growing interest in the role of representations in mathematics. It is often easy to categorise representation as the concrete materials used to represent a sign, or a configuration of signs, characters or objects (Goldin & Shteingold, 2001). In fact, as Goldin and Shteingold (2001) point out, much of the history of mathematics has been about the creating and the refining of representational systems, and much of the teaching of mathematics has been about students learning to work with them and solve problems with them. However while these external representational systems are useful, they cannot be used as a measure of a child’s understanding. As Goldin and Shteingold (2001) acknowledge, “Rules in mathematics can be learned and followed mechanically and definitions can be measured—without very much conceptual development having taken place” (p. 5). This is particularly relevant to the assessment process. External representations as a measure of a child’s mathematical understanding are therefore clearly limited and only provide one perspective of what a child knows.

The term ‘internal representation’ is now being used to refer to the “mental configurations of individuals, such as learners or problem solvers” (Goldin
& Kaput, 1996, p. 399). It is this form of representation that goes beyond the obvious pen and paper response, to understanding the mental processes and strategies used.

Therefore “external representations permit us to talk about mathematical relations and meaning apart from inferences concerning the individual learner or problem solver. Internal representations give us the framework for describing individual knowledge structures and problem solving process” (Goldin & Kaput, 1996, p. 406).

Pape and Tchoshanov (2001, p. 119) illustrated the relationship between internal and external representations as a child develops the concept of the number 5 (see Figure 3.1). They utilise three circles that demonstrate the interplay, represented by four double-ended arrows between the external representations as an outer circle with the mental or internal representation of the child. They describe how initially the child is playing with several toys and names them using numbers e.g., the first is one, the second is two and so on. In this instance, the words are symbols for the position of the toys, not necessarily how many there are. Later on the child develops an understanding that the last number of toys represents how many there are in the set. Consequently once this is learnt external representations like the numeral and its name (e.g., 5 and five) become the conventions for the internal abstraction, the number of elements in the set. Thus, “the number name, five, and the numeral, 5, are the external representations that act to stimulate an image, the internal representation, of a set of five objects” (Pape & Tchoshanov, 2001, p. 118).
This constructivist approach acknowledges that mathematical knowledge is actively constructed by learners (Goldin & Kaput, 1996). In fact, the *Principles and Standards for School Mathematics* published by the National Council of Teachers of Mathematics (NCTM, 2000) has introduced representation as a process standard, an attribute for children to learn and acquire. Within the NCTM (2000) document, “the term representation refers both to a process and to product…to the act of capturing a mathematical concept or relationship in some form and to the form itself” (p. 67). Therefore if these two forms of representation are viewed as necessary to understanding the full extent of a child’s mathematical knowledge, why then are standardised assessments concentrated exclusively on external representation i.e., words, numerals, equations, symbols and signs, graphs, diagrams, charts and tables (Matteson, 2006)? This question is crucial and as a result the research has been designed for students to have the
opportunity to describe and explain their problem solving strategies to gain insight into how their internal representation is reflected in the external representation.

This opportunity to verbalise their internal mental representation is not only beneficial in revealing the students’ problem solving strategies, but also allows them to identify aspects of the item that may be hindering their understanding. This is particularly relevant when measuring the effectiveness of graphics, often viewed as pictures without consequences. Yet perhaps the inclusion of graphics may be having detrimental effects on a child’s ability to solve a task particularly if the graphic is taking on an unintentional role. According to Lowe (1993), “an individual’s mental representation of a particular diagram is the basis for the way that diagram is processed and ultimately understood (or not, as the case may be)” (p. 5).

Therefore in this inquiry the methods selected to collect and analyse the data needed not only to provide insight into ‘how’ students answered the questions but ‘why’. Therefore a mixed method design has been developed.

**Utilising the Strengths of Quantitative and Qualitative Approaches**

According to Bergman (2008) “mixed method design, i.e., the combination of at least one qualitative and at least one quantitative component in a single research project or program, has experienced a tremendous increase in popularity...” (p. 1). The combinations of these paradigms are often the result of the researcher’s belief and values as well as the purpose of the inquiry. In Greene, Caracelli and Graham’s (1989) empirical review of 57-mixed method studies they identified five mixed method purposes. These
included triangulation, complementary, development, initiation and expansion.

A complementary mixed-method study involves the use of qualitative and quantitative methods “to measure overlapping but also different facets of a phenomenon, yielding an enriched, elaborated understanding of that phenomenon” (Greene et al., 1989, p. 258). The methods used in the current study were not just to complement each other but also to guide and interact with one another. As Stake (2010) notes, “‘mixed methods’ is using multiple methods interactively, not just using them somewhere in the same study” (p. 125). Such a complementary approach to this inquiry provided two measures to assess similar, as well as different, aspects of the same phenomenon. In the current study, students’ reported results are documented as well as their processes on how these answers were reached. A complementary approach is required to examine these two representations.

As previously mentioned, standardised mathematics assessment results are often represented in a purely quantitative light i.e., right or wrong answers. And while this analysis still played a vital role within this inquiry, this form of analysis offered no real insight into the ‘answer behind the answer’. It was for this reason that qualitative methods were utilised in an explanatory manner to complement and explore the more formal quantitative results. As Cooper (1998a) similarly noted in his study on the United Kingdom mathematics curriculum:

A central concern is to understand how pupils interpret and respond to test items in the context of actual testing. However, while the distribution of right and wrong answers is available from such settings it is only possible to infer, and even then with great difficulty, the meanings pupils give to items from their answers and the limited ‘working out’ which accompanies some
answers. To gain fuller access to children’s interpretation of items it is necessary to relax the formal testing condition ... work through a selection of items – as if in a test – but with the difference that at the end of every item they were asked to give an account of their understanding and procedure (p. 23).

Cooper’s research aims allowed the researcher to explore the very nature of test items and provide insight into the strategies and knowledge needed by students to successfully complete them. The NAPLAN test also only provides “the distribution of right and wrong answers”. Within this study the qualitative interviews provided “fuller access to children’s interpretation of items” (Cooper, 1998a). These research aims allowed the researcher to explore the nature of test items and provide insight into the strategies and knowledge needed by students to successfully complete them.

In this study a mixed-method design provided the opportunity to analyse the data from multiple perspectives and at depths. Thus quantitative techniques were used to identify patterns and relationships but also differences between the performances of students. These patterns then allowed for a more in depth analysis of student processing of information.

Subsequently the research almost resembled an experiment, with the results of each phase dictating the direction of the next.

As noted by Teddlie and Tashakkori (2009):

> Sequential mixed designs answer exploratory and confirmatory questions chronologically in a prespecified order. Though still challenging, these designs are less complicated to conduct by the solo investigator than are the parallel mixed designs, because it is easier to keep the strands separate and the studies typically unfold in a slower, more predictable manner (p. 153).

Specifically, the use of a sequential mixed design allowed four strands of data collection to occur chronologically (QUANTITATIVE ➔
QUALITATIVE ➔ QUANTITATIVE ➔ QUALITATIVE) to help establish a story from within the data. This allowed the conclusion based on the results of the first strand to lead to the formulation of design components for the next strand. Subsequently the second strand of the study was conducted either to confirm or disconfirm inferences from the first strand or to provide further explanation for its findings.

The Site and the Participants

The inquiry took place in four Australian primary schools in the state of New South Wales. The schools were diocesan Catholic primary schools for children aged 5-12 years. Situated in a regional city with a population of over 60 000, the four primary schools varied in size and socioeconomic status. The largest school had 3 classes while the smallest school had only one.

The students were in Year 5, and were aged 10 or 11 years. Altogether, approximately 143 students were involved over a three-year period. These students and schools were selected for the following reasons:

- Time constraints. The use of the Catholic Diocese resulted in quick ethics approval compared to an often lengthy and exhaustive procedure through the affiliated government departments;
- The Catholic schools represented varying socioeconomic status of the population. One was clearly located in a disadvantaged area compared to another situated within a new development with many large houses and high-income families; and
- The combined schools population was well within the capabilities of the researcher.
Forty of these students were utilised further as a nested relation to the quantitative component of the data collection with one-on-one interviews (see Johnson & Christensen, 2008). These forty students were purposely selected by their teachers to be representative of the varying ability within their classrooms. Therefore the sample included children who struggled, excelled or maintained a class average in mathematics. The benefit of using this version of “maximum variation sampling” (McMillan, 2008, p. 12) meant that students were seen as being representative of any classroom making the findings more applicable for schools and teachers.

**Implementation of Qualitative and Quantitative Instruments**

Initial contact was made with the Director of the Wagga Wagga Catholic Diocese, who recommended ethical procedures as well as possible participants. A letter seeking approval to conduct the research project was drafted and sent to the four Principals of the selected schools. Included with the letters were information packages for the teachers as well as the parents/caregivers of the students involved. All information packages included a letter of consent outlining the nature of the research project, the responsibilities of the participants and the obligations that entailed. The information packages and consent forms that were used are presented in Appendices B-F.

Once consent was granted, the test was administered by the researcher over a period of a month. The students completed the NAPLAN Item Test which was compiled of 15 items taken from the 2010 Year 5 NAPLAN which the students sat in May 2010. The Australian Curriculum, Assessment and Reporting Authority (ACARA) released the results and test in September
2010, and data collection for this inquiry commenced at the end of September and beginning of October 2010.

The test was completed under exam conditions to replicate those of the NAPLAN. This included the researcher being unable to read the question for the students or provide any assistance in understanding what the question was asking. The students’ desks were separated to discourage cheating and each class was allocated 40 minutes to complete the test. The researcher encouraged the students to manually record any working out on the text booklet and emphasised the need for students to carefully check answers, and if uncertain, to either guess or leave the question blank.

Following the month of quantitative data collection, 37 students were purposively sampled by their school’s executive teachers to participate in the qualitative component. These teachers were asked by the researcher to provide 10 students of mixed ability from within the school’s Year 5 cohort. It was emphasised that there was to be a proportionate amount of low, average and high achieving students. The students involved in the interviews were not included in the mass testing component. These students were required to answer the questions within the interview in order to immediately evaluate and record their mathematical reasoning. Their absence from the mass testing also prevented the students from becoming familiar with the items prior to the interview.

The purpose of such a design was to provide more comprehensive and in depth information from a smaller number of carefully selected cases (Teddlie & Tashakkori, 2009). It also utilised the expert judgment of the executive teachers on the abilities of their students. It was important that the interviews were as representative as possible of the testing cohort thus
taking on a nested sequential sampling design. Consequently “participants selected for one phase of the study represent(ed) a subset of those participants who were selected for the other phase of the study” (Johnson & Christensen, 2008, p. 246). The students’ responses were digitally recorded and later transcribed by the researcher. Each interview lasted no more than 40 minutes.

Using the results from the NAPLAN Item Tests as well as the coded interview transcripts, the Modified Test was produced and the large cohort were tested again almost a month after sitting the NAPLAN Item Test. The interviews were conducted straight after the mass testing at the end of October 2010 and again provided the students with the opportunity to explain their mathematical thinking and problem solving. Once again, these responses provided important insight into children’s mathematical processing as well as information about the reasons for observed differences in performances across item types (original versus Modified). As noted by Kettler, Elliot and Beddow (2009):

> The use of concurrent think-aloud protocols and follow-up questioning allows researchers to ‘unpack’ unexpected results ... recording students’ concurrent verbalizations while solving the item in question, as well as questioning students following completion of the task, may illuminate item features that contribute to the observed results (p. 542).

The interviews followed the same protocol as those used in the NAPLAN Item Test interviews making it easy to cross reference responses. The short time frames between all stages of data collection were essential to ensure limited teaching exposure between testing regimes which may have affected the validity of results.

The methods used to explore and tell such a story will now be examined.
Methods

The methods used in this investigation were a combination of quantitative and qualitative instruments to achieve the research criteria. Due to the sequential nature of the project each strand purposely directed the next. Stake (2010) referred to this as ‘progressive focusing’, and noted stages in the investigation to observe, inquire further and then seek to explain. Utilising the work of Parlett and Hamilton (1977), he described the use and relationship of the stages as they inform and direct the research:

Obviously the three stages overlap and interrelate functionally. The transition from stage to stage, as the investigation unfolds, occurs as problem areas become progressively clarified and redefined. The course of the study cannot be charted in advance. Beginning with an extensive data base, the researchers systematically reduce the breadth of their enquiry to give more concentrated attention to the emerging issues. (p. 15, cited in Stakes, 2010, p. 130).

This was particularly relevant in this study because of the uncharted territories it was attempting to explore. Although there were “pivotal ideas, anticipations (and) frameworks for understanding what to do next” (Stakes, 2010, p. 130) there was a degree of the unknown that really shaped the research as issues emerged from within the data. It was for this reason that the first stage was a pilot study which was conducted while the participants were in Year 3. The purpose of this stage of the study was to assist in establishing protocols and recognising any possible gaps or problems within the research design.

The pilot study involved 169 students who were initially tested on selected items from the 2008 NAPLAN. From within this cohort, 40 students were purposively sampled for the interview strand. Following the analysis of the first strands of quantitative and qualitative data, the NAPLAN Item Test
was modified, and 169 students were tested on the Modified Test, with the same 40 students interviewed on their mathematical reasoning.

Following the success of the pilot study, the same students who were now in Year 5 were again tested, but this time on selected items from within the 2010 NAPLAN. Due to population shifts within the participating schools, only 143 students of the original 169 participated in the second stage of the study. This also resulted in a smaller number of participants in the interview strand following the NAPLAN Item Test. Once again the items were modified as a result of analysing the data sets and students were tested on the Modified Test. The final strand of the second phase was the interview data surrounding the NAPLAN Item Test results.

Therefore the investigation took on the following design (Figure 3.2).

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NAPLAN Item Test n=143 (Quantitative)</td>
<td>NAPLAN Interviews n=40 (Qualitative)</td>
</tr>
<tr>
<td>NAPLAN Item Test n=169 (Quantitative)</td>
<td>NAPLAN Interviews n=40 (Qualitative)</td>
</tr>
</tbody>
</table>

*Figure 3.2. Research design.*
The methods and instruments used within the inquiry were chosen to allow the researcher to:

1. Gain insight into the type of questions that students struggle with in standardised assessment;
2. Consider the extent to which assessment reflects practice and represents sense making and mathematical understanding;
3. Explore the relationship between graphics, language, contexts and item features;
4. Examine the way in which items are constructed and how that impacts on a student’s capacity to make sense of mathematics; and
5. Investigate elements of an item that best contribute to our understanding of student knowledge and understanding.

The first strand of this investigation was a pilot study.

*The Pilot Study*

Two years prior to the data collection process for this inquiry, a pilot study provided the opportunity to chronologically modify and build the scope of the design in an iterative manner. Data was analysed as the conceptual framework of the study was forming. The connection between the theoretical underpinning of the study and the initial analysis of data was substantiated. This evolving process fitted appropriately the design of the study and the scaffolded links established through the formation of the research questions.

In the pilot study, 20 items from the Year 3 2008 NAPLAN were administered to 169 children Year 3 students from within 4 regional Catholic schools. These students were the same participants that would be
utilised further within this inquiry. The use of the data collection techniques and analysis process are demonstrated through the use of the following item (see Figure 3.3) taken from within the pilot test.

![Figure 3.3. Pilot test item.](image)

From the initial quantitative analysis, it was revealed that 56% of students answered this question incorrectly with the majority of the cohort choosing answer C². By reading this data alone it could be inferred that over half the students were unable to correctly read a graph or draw conclusions from it. However when questioned within the one-on-one interviews on how they drew their conclusion it revealed that nearly all students could successfully read the graph but simply did not understand the terminology, in particular the term ‘fewer’. This can be seen in the following examples:

Jane: How did you get your answer?

---

² To aid in the recognition of answers throughout the thesis a system of A, B, C and D will be utilised to describe student responses.
SJ1: 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’s response indicated that she believed ‘fewer’ meant more. HT5 also demonstrated the same misunderstanding:

Jane: How did you get your answer?

HT5: 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.

Jane: So there are fewer sheep than goats. How many sheep are there?

HT5: There are 6.

Jane: And how many goats are there?

HT5: 4.

Jane: What’s another way of saying that?

HT5: There are more sheep than goats.

These responses were then coded according to whether the graphic, language, item features or context had the most significant influence on the student’s answers, both correct and incorrect. In this instance the responses were coded ‘L’ as the language clearly impacted on the way the students answered. With the use of the codes it became apparent that 48% of the interview cohort were classified ‘L’. Subsequently the item was modified accordingly with the word ‘fewer’ replaced with the term ‘less’ (see Figure 3.4).
This item became one of the ten created for the Modified Test booklet that was then administered to the same 169 students. The quantitative results from the third strand of data collection revealed that only 5% now answered this question incorrectly. The two research tests were investigated further through an analysis of variance (ANOVA) to determine whether there were statistically significant differences between the mean scores of the NAPLAN item and the modified item. The ANOVA revealed statistically significant differences between the mean scores on each of the items: F(1,336)=155.36, p=.000. Table 3.1 presents the means (and standard deviations) for the two items.
Table 3.1
Means and standard deviations between test results

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAPLAN Item Test</td>
<td>169</td>
<td>.44</td>
<td>.498</td>
</tr>
<tr>
<td>Modified</td>
<td>169</td>
<td>.96</td>
<td>.200</td>
</tr>
</tbody>
</table>

In order to ascertain the reason behind this dramatic change, the 40 students were once again interviewed on their problem solving and mathematical thinking. An example of their mathematical reasoning was as follows:

Jane: How did you get your answer?

HT5: 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.

The use of such a research design enabled the researcher a greater insight into the relationship between test item construction and students results.

By reading the results of the original NAPLAN Item Test, with no insight into children’s mathematical thinking, it could be assumed that over half the students were unable to read a graph correctly. However through the use of the interview data it was revealed that 95% could successfully complete the mathematical component of the question but were unable to access the item due to literary restraints. These results verified the use of the sequential mixed method research design that was then later reflected in the actual inquiry.
Quantitative Testing

The quantitative data obtained was used to:

1. Assess students’ problem solving techniques and methods;
2. Determine relationships or trends between students’ performance test items and the variation in item design; and
3. Ascertain common errors in student responses and determine causes for these error patterns.

Therefore the quantitative component of the investigation addressed the first two major research questions, specifically:

1. From a selected number of often used mathematics tasks, which do students find most difficult to solve?
2. To what extent does assessment reflect sense making and mathematical understanding?

The NAPLAN Item Test

The first strand of this inquiry began with the administration of 15 selected items from the 2010 NAPLAN. The NAPLAN Item Test was used as a measure of what items students answered correctly and to identify any themes or common trends. According to Tashakkori and Teddlie (2003), tests are commonly used in quantitative research to measure performance of research participants. The use of a standardised test as a research tool guaranteed that the validity of the test had already been measured. As noted by Connolly (2011):

All NAPLAN items undergo a trialling process. Post-trial items are analysed using the Rasch model (Wright, 1980). The Rasch model is a mathematical model of test scores for tests that satisfy a number of important pre-conditions. Using the Rasch model to analyse tests allows for sensible comparisons of test scores between
different year groups, cohorts and test papers. Use of the Rasch model requires that the items in the test are:

one-dimensional: the items all test the same underlying skill or ability;

locally independent: individual items should not affect the probability of answering other items correctly; and

uniformly discriminating: the chance of answering an item correctly should increase in a uniform way for students of increasing ability.

These conditions insure that students are assessed against a standard that is both consistent and stable across a series of tests (p. 909).

Utilising the NAPLAN items as a means of testing within research is becoming more common in Australian contexts since its introduction in 2008 (see Deizmann, 2008; Greenlees, 2010; Hill, 2011; Nisbett, 2011; Morley, 2011).

The test was therefore created using the 2010 Year 5 Mathematics NAPLAN. Apart from the consideration of validity, there were a number of other reasons why the NAPLAN became the basis for this inquiry:

1. Familiarity of item design for the students. The current participants were the first cohort to complete national testing in 2008 when the NAPLAN was introduced and were therefore familiar with its format.

2. Reduction of student anxiety. The test format is well known to the participants as many items are practiced, rehearsed and part of classroom practice and therefore would carry less anxiety than unfamiliar material that the researcher may have introduced.

3. Opportunity to reflect upon current national assessment practices. NAPLAN is a high-stakes mathematics assessment which has
increased teacher and school accountability with a huge impact on classroom practice. For this reason it needs to be scrutinised.

4. Items have contextual and curricula relevance and have been purposively designed for students at particular developmental stages. It is important to determine whether item design of this powerful assessment tool is appropriately matched to mathematics curricula and stage of the students.

It is important to note that the numeracy items from the NAPLAN tests were used as ‘representative’ mathematics items suitable for students in primary schools. As such the purpose of the inquiry was not to dispute the validity of the NAPLAN but rather to enlighten teachers on the structure of such items and the impact this may have on their teaching practice. For example, one of the results from within the pilot test was the mismatch between the terminology used in the assessment and clearly that which was used in the classroom with the use of the word ‘fewer’ instead of ‘less’. The focus here is not on the test per se, but rather the student experience manoeuvring through the complexities of the test items, the NAPLAN test items merely being their vehicle.

The main disadvantage of using the NAPLAN as the research tool was that the test contained 40 items. This quantity was problematic especially in regard to collecting the qualitative data through one-on-one interviews. The researcher was conscious of not wanting to exhaust the participants with extensive and long interviews which could have resulted if all 40 items were used. It was for this reason that 15 items were purposively sampled to make up the NAPLAN Item Test. The following criterion was used in selecting the items:
1. They contained two or more of the four components of mathematics assessment (graphics, language, item features, situations);

2. It was easy to identify ways the item could be modified without changing the complexity of the question; and

3. There were varying degrees of difficulty between the questions.

As previously noted, by using items taken from a standardised test such as the NAPLAN, the psychometric properties of the test had already been determined i.e., their reliability and validity (Teddle & Tashakkori, 2009). However, Teddle and Tashakkori (2009) also warned of the possibility of cultural bias and how this may affect the test score of minority groups. Yet despite the advantages and disadvantages of the use of such a test, these NAPLAN items were the very questions students were contending with and staff being appraised on. It was therefore most useful to examine participants on these items to gain a greater understanding of what NAPLAN results are indicative of. The items used are presented in Appendix A.

**The Modified Test**

In accordance with the increased scrutiny being placed on mathematics assessment, the concept of modifying test items to make them more accessible has emerged (see De Corte, Verschaffel, & De Win, 1985; Vicente, Orrantia, & Verschaffel, 2007; Beddow, Kettler, & Elliott, 2008). In particular, these modifications have been used to increase the child’s ability to solve the item. This has been predominantly beneficial for students with disabilities. This study demonstrates the benefit of modifications for many learners and should not be restricted in its use.
According to Kettler, Elliott and Bedlow (2009), “This emerging Modification Paradigm offers a conservative and systematic approach to improving the accessibility of an item, without changing the difficulty to the point that it would no longer be an indicator of a grade-level content objective” (p. 531). It also resulted in the development of an evaluation tool for appraising the accessibility of an item known as the Test Accessibility and Modification Inventory (TAMI) (Beddow, Kettler, & Elliott, 2008).

Beddow, Kettler and Elliot (2008) argued that the goals of TAMI and the modification process were to:

1. Increase access for all test takers;
2. Remove extraneous material;
3. Maintain the same depth of knowledge;
4. Improve efficiency; and
5. Increase the validity of inferences from test results (Beddow, Kettler, & Elliott, 2008, p. 3).

Within this model they claimed that the first step in the modification process was to evaluate the original item for accessibility. This is accomplished by “appraising the constituent elements of the item, attending specifically to barriers that could affect some of the target population” (p. 532). A close analysis of every item is necessary to identify potential obstruction to student performance.

In this inquiry this step was achieved through the use of the interview data. It became apparent through the transcription and coding process that there were certain elements that limited students’ access to solving the item. Kettler, Elliott and Beddow (2009) define this access as:

the opportunity for test takers to demonstrate proficiency on the target construct of a test (e.g.,
language arts, mathematics, or science) or a test item ... In essence, complete access is manifest when a test-taker is able to show the degree to which he or she knows the tested content. Access, therefore, must be understood as an interaction between individual test-taker characteristics and features of the test itself (p. 530).

The questions raised at the end of the coding process became the basis for the modification of test items. For example if a graphic was particularly influential in a child’s response it was altered, similarly with the language, item features and context. However it was essential that the modifications did not affect the level of difficulty or the item construct being measured. To demonstrate and justify the modification process the following four items will be examined (see Figures 3.5, 3.6, 3.7 and 3.8). Each of the items was modified according to graphics, language, features used or situations. Participant’s voice appears here to illustrate its use in shaping the study. Items were modified in direct response to the student’s verbal explanation of their mathematical reasoning. These strategic examples are used to explicate the methodology used within this study.

**Graphics Modification**

The following item has been included as an example of a graphics modification. As highlighted previously, many of the NAPLAN items contained graphics, and in particular information graphics. Justifiably, this resulted in many interview responses being coded ‘G—Graphic’ because the information they were using to solve the item they needed to obtain from within the graphic. By definition this is the role of information graphics and students need to have the confidence and skill in utilising such data. However, while it is understandable that information graphics would elicit ‘G—Graphic’ responses, it should not be the case with contextual examples
as these graphics do not provide a source of information necessary for children to solve the task. It was therefore concerning when the NAPLAN Item Test interview data revealed a high proportion of students being coded as ‘G—Graphic’ for items such as the Shoe item (see Figure 3.5).

The purpose of the Shoe item was to assess a child’s conception of centimetres and length by estimating the length of a real shoe. To aid their understanding of what a shoe is (an item most children would wear everyday), an illustration of a shoe is provided. In this instance there was no need for children to make reference to the picture within their interview responses as it was not a necessary component to the solution of the task. Yet it became apparent by analysing the children’s reactions that the use of the shoe as a contextual graphic was unnecessarily relied upon. This can be seen in Ethan’s response when he answered C (75cm):

Jane: How did you get your answer?

Ethan: I put 75cm ‘cause a shoe wouldn’t be 5cm and I just used the picture.

Many students similar to Ethan regarded the illustration to be of an adult’s shoe and subsequently distorted their concepts of length and the actual size of an adult’s foot. Many automatically assumed that an adult’s foot would be much larger than their own and they answered the item with this in mind. They therefore opted for the answer which equated with a greater length. However their responses demonstrated poorly developed number sense especially at the thought of a person’s foot being close to a metre long.

Using the item analysis construct outlined in the TAMI (Beddow et al., 2008), it is argued that visuals should only be included when necessary, should be relevant to essential item content, and clearly represent intended
images. In this instance the use of a shoe had caused unnecessary confusion as to what it represented and why it was utilised. It was for these reasons that within the modification process the graphic was taken away (see Figure 3.5).

The removal of the graphic also reduced the amount of language required as the graphic no longer needed to be introduced. The word ‘your’ was identified as an essential word and therefore bolded similar to the use of bolding ‘real’ in the original question to facilitate identification (Beddow et al., 2008).

The NAPLAN Item Test

This is a picture of a shoe.

Which of these is closest to the length of a real shoe?

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5cm</td>
<td>25cm</td>
<td>75cm</td>
<td>100cm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Modified Test

Which of these is closest to the length of your shoe?

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5cm</td>
<td>25cm</td>
<td>75cm</td>
<td>100cm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.5. Graphics modification.
With so many information graphics being used in the NAPLAN it is understandable why students like Ethan would feel it necessary to include the picture as having significance and value in obtaining the correct answer. This raised the question of how Ethan would respond if the contextual graphic was taken away?

Language Modification

The following item has been strategically included as an example of a language modification. According to Vincente, Orrantia and Verschaffel (2007), the use of language in mathematics problems has for a long time attracted the attention of researchers, both cognitive psychologists and mathematics educators. They argued that “a number of empirical studies have shown how small changes in the wording of the problem texts may have a dramatic (positive) impact on children’s solution processes and skills” (Cummins, 1991; Cummins, Kintsch, Reusser, & Weimer, 1988; Davis-Dorsey, Ross, & Morrison, 1991; De Corte, Verschaffel, & De Win, 1985; Stern & Lehrndorfer, 1992).

The intent of the Pie graph item in the NAPLAN Item Test (see Figure 3.6) was to measure a child’s ability to interpret a graphical representation and draw appropriate conclusions. Following the analysis of children’s responses to this item it became apparent that students were influenced either positively or negatively by the language. An example of this can be seen in Billy’s response when he answered B:

Jane: How did you work out your answer?

Billy: Cause it’s sort of something good to do and that’s probably why she plays it each week and the others they’re like one day, they only do the three things on one day (looking at the answers).
Billy’s answer was confusing and lacked insight into what the question was asking. An incorrect response similar to Billy’s may be interpreted as evidence that the entire concept was too hard and they were unable to accurately read a pie-graph to obtain certain information. Or the way the question was worded was problematic and led to confusion. It was this latter hypothesis that led to the modification of the language (see Figure 3.6).

The NAPLAN Item Test

Hannah made a pie graph to show the number of hours she spent on different activities over 24 hours on Monday.

Which information can be found using this pie graph?

The number of

- meals Hannah eats on this day.
- hours Hannah plays sport each week.
- hours Hannah watches TV on Tuesday.
- hours Hannah spends awake on this day.

The Modified Test

This pie graph shows the number of **hours** on Monday Hannah spent doing different activities.

Which information can be found using this pie graph?

The number of

- meals Hannah eats on this day.
- hours Hannah plays sport each week.
- hours Hannah watches TV on Tuesday.
- hours Hannah spends awake on this day.

*Figure 3.6. Language modification.*
Beddow et al. (2008) outlined in the TAMI construct that for an item to be accessible to all students, the item stem or question must only include words essential for responding to the item, with “minimal extraneous verbiage” (p. 2). Hadalyna (1999) noted that often this excessive verbiage is the result of attempting to make the item more realistic and thus provide some substance to it. However he argued that unless the purpose of the item is for students to be able to separate the useful information from the useless, this extra information is nothing more than ‘window dressing’.

Therefore in comparison to the original language used in the NAPLAN Item Test, the modified version was more specific and defined with essential words bolded to once again facilitate identification. Of course this raised the question of how Billy would respond if the language was more comprehensible?

**Item Features Modification**

The following example is indicative of an item features modification. In the Baby mass item (see Figure 3.7), the options used in the answer choices were to measure children’s abilities to compare the mass of kilograms and grams. This was achieved by recording the weights of babies in both forms of measurement. Therefore, the options used in the multiple choice were crucial in determining which was the heaviest baby. The interview process, however, revealed that sometimes children answered correctly, although this correct answer was not based on a sound mathematical knowledge of kilograms and grams conversion, but rather on the assumption that kilograms were greater. They disregarded the weights measured in grams as not relevant and focused solely on those measured in kilograms. This can be
seen in responses similar to Lauren’s where a correct response of A was provided but not based on appropriate mathematical reasoning:

Jane: What was your answer?

Lauren: Simon because grams is smaller than kilograms and Mia is 3.05 and that one is smaller than 3.5.

According to the item analysis put forward by Beddow et al. (2008) in the TAMI, these answer choices were plausible and subsequently fair and accessible. It was only through the privilege of the interview transcript that the impact of the item construction was realised. It was for this reason that the greatest mass was presented in grams instead of kilograms to gain real insight into children’s measurement understanding in the Modified NAPLAN Item Test (see Figure 3.7).

The NAPLAN Item Test

Those babies were born on the same day. Which baby has the greatest mass?

<table>
<thead>
<tr>
<th>Baby</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simon</td>
<td>3.6 kg</td>
</tr>
<tr>
<td>Jennifer</td>
<td>3.45 kg</td>
</tr>
<tr>
<td>Mia</td>
<td>3.06 kg</td>
</tr>
<tr>
<td>Oscar</td>
<td>3.09 kg</td>
</tr>
</tbody>
</table>

The Modified Test

Those babies were born on the same day. Which baby has the greatest mass?

<table>
<thead>
<tr>
<th>Baby</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simon</td>
<td>3.4 kg</td>
</tr>
<tr>
<td>Jennifer</td>
<td>3.405 kg</td>
</tr>
<tr>
<td>Mia</td>
<td>3.05 kg</td>
</tr>
<tr>
<td>Oscar</td>
<td>3.08 kg</td>
</tr>
</tbody>
</table>

Figure 3.7. Item Features modification.
This Baby mass item highlights the benefit of the mixed methods design and the interactive nature of the approaches that were utilised. In this instance it could be assumed from the NAPLAN Item Test quantitative data that students did have an accurate understanding of the relationship between kilograms and grams, but yet the interviews provided voices that did not support this assumption. This form of analysis (the taking things apart) and synthesis (the putting things back together) (Stakes, 2010) provided much needed detail into the construction and modification of mathematics assessment and the mathematical reasoning behind students’ results. It led to questions such as what answer Lauren would have given if the baby with the greatest mass was presented in grams?

**Situation Modification**

The final example is provided to exemplify and justify the modification of a situation. As previously discussed, the use of real-life situations within mathematics assessment has been gaining momentum as a reflection of changing classroom practice. However the use of situations within assessment can be problematic when considering the accessibility for all students. According to Beddow et al. (2008), an item must not require “construct-irrelevant knowledge and skills that may make the item more difficult for some test-takers regardless of the construct being measured” (p. 3).

The Scales item was included to measure a child’s understanding of fractions and decimals (see Figure 3.8). However, the use of grocery scales resulted in the creation of a situation that a child may or may not be familiar with. Therefore, 100% of student responses were classified S-Situations, as the situation either negatively or positively impacted on students’
understanding depending on their familiarity of the situation used. An example of this can be seen in Elizabeth’s response to the item where she answered D:

Jane: How did you get your answer?

Elizabeth: I times 6 by 10, ‘cause there’s 10 oranges and I got 60.

Jane: So did you have to use the pictures at all to help work out your answer?

Elizabeth: Not really.

The scales used in this item could be considered as being very specific in nature, function and appearance and is an example of how some students may be disadvantaged by its inclusion. Unless a student has specifically used the scales in a ‘real-world’ setting, or the teacher has incorporated it into their classroom teaching, some students may be unfamiliar with its use or relevance. In this instance, it is not just exposure to the use of scales but also the way they are utilised when calculating the cost of fruit and being able to relate the weight to cost per kilogram.

Subsequently the item was modified by removing the graphic and associated situation to create a word problem (see Figure 3.8). However it is important to note that the modified question would still be regarded as a situational one as it is encased within a ‘real-life’ circumstance, yet the absence of the scales made it more accessible for students who were unfamiliar with their use.
With the increasing presence of situations within high-stakes tests, the results following the modification of the Scales item will be particularly insightful into their impact on students’ results and mathematical reasoning. It poses the question of how would Elizabeth respond if the graphic creating the situation was removed?

Following the analysis of all 15 items contained within the NAPLAN Item Test, only 10 were modified and included within the Modified test (see Appendix I).
Five items were excluded from the Modified Test for either one of two reasons:

1. Anticipated modifications would alter the difficulty of the item; or
2. The item was considered easy and therefore accessible.

Of the 10 items included within the Modified Test, 3 were modified according to the graphic (either inserted or removed), 1 was modified according to the language, 4 modified according to the item features and 2 items were altered in regard to the situation. This test was then administered to the larger cohort of students only one week after the conclusion of the NAPLAN Item Test interviews.

The resulted instruments were used to gather the data that was subsequently analysed using both quantitative and qualitative methodologies.

**Quantitative Analysis**

At the conclusion of each stage of quantitative data collection the results were analysed using a variety of methods. Following the NAPLAN Item Test and Modified Test, student results were recorded simply as correct or incorrect. These were represented as ‘smiley’ faces and were used to compare results (see Appendix J). Individual students were awarded a smiley face for a correct response and a grey face for an incorrect response. This enabled the researcher to add up all the correct responses for each item and compare them between both tests. It also illustrated changes in individual children’s performance within items and across the tests. This basic analysis provided a general indication as to the effect of the modifications on student results. A more in depth analysis was required to determine the extent of statistical significance.
This was achieved through two forms of data analysis—chi-squares and analyses of variance (ANOVAs). By utilising both forms of tests, the researcher was able comment upon the frequencies, percentages, proportions and means of correct and incorrect responses to items contained within the NAPLAN Item Test and the Modified Test. This commenced with 10 chi-squares presented as cross-tab contingency tables. The tables identified the relationship between the changes to students’ correct responses in the NAPLAN Item Test to the number of correct responses in the Modified Test. Items were considered to be of relevance if the changed variance between the tests was greater than 25%. That is, students who either originally answered correctly (standard item) or incorrectly (standard item), now had incorrect (modified) or correct (modified) responses in the second instrument. Although this form of analysis was useful in establishing an initial picture of the relationship between the results of the NAPLAN Item Test and the Modified Test, it was also necessary to identify any significant differences between the means of the two tests. For this reason a number of ANOVAs were used.

As outlined by McMillan and Schumacher (2006), this form of analysis is useful in recognising significant differences between groups through the calculation of an $F$ value. McMillan and Schumacher noted that “if the $F$ value that is calculated is large enough, then the null hypothesis (meaning there is no difference among the groups) can be rejected with confidence; the researcher is correct in concluding that at least two means are different” (p. 301). The use of ANOVAs assisted in identifying any statistically significant differences between the means of the items contained within the NAPLAN Item Test and the Modified Test. An ANOVA was also utilised to
measure the effect of gender upon results. However, all quantitative data were external representations of student results and did not reveal any internal representations. It was for this reason that the results were analysed concurrently with the qualitative data to provide a rich and compelling story compared to the presentation of raw quantitative data alone.

**Qualitative Interviews**

Qualitative techniques of analysis were used to:

1. Undertake an in depth analysis of students’ sense making when solving mathematical items from both the original and modified item designs;
2. Examine the relationship between graphics, language, situations and item feature; and
3. Identify common themes between students when solving certain tasks.

The two phases of qualitative data collection were utilised to gain insight into the following:

1. To what extent does assessment reflect sense making and mathematical understanding?
2. How does item construction impact on a students’ capacity to make sense of mathematics?
3. What effects will these findings have on classroom teaching?

**One-on-one interviews**

In the present investigation interviews were conducted in order to gain a deeper appreciation and insight into the strategies students used when solving mathematical assessment tasks. With the intention of remaining
non-judgemental (Tashakkori & Teddlie, 2003), an interview protocol was established to help ensure all students were treated fairly and equitably (see Appendix I). This form of interview has been categorised by Paton (2002, cited in Teddlie & Tashakkori, 2009), as a standardised open-ended interview. This format allowed the exact wording and sequence of the questions to be prepared in advance. Subsequently all interviewees were asked the same basic questions in the same order, however they were still open-ended in nature. Some of the general questions included:

- Can you tell me how you worked out your answer?
- Tell me what you did to work out the answer.
- Was there anything in the question that helped you work out your answer?
- Did the picture help you at all?

The purpose of these questions was to ascertain the elements within the item and how they impact on a child’s mathematical reasoning. It was the open-endedness of this inquiry that allowed the students to explain their thinking processes without the fear of getting the answer ‘right’ or ‘wrong’. From the beginning, the investigator reinforced the importance of ‘getting inside’ the students’ heads rather than the correctness of their responses. This gave the interviewees the assurance to be confident and honest in their responses. It also assisted in developing trust between all parties concerned.

The interview data were analysed to identify common themes between the students’ responses and to ascertain which component of the item had most influence, both positive and negative, on a child’s understanding.
Qualitative Analysis

Coding

Transcriptions were coded according to the component of the item which overly influenced either a right or wrong response. This was in an effort to answer the third research question “How does item construction impact on a students’ capacity to make sense of mathematics?”

According to Miles and Huberman (1984), coding is a form of analysis. It is:

To review a set of field notes, transcribed or synthesized, and to dissect them meaningfully, while keeping the relations between the parts intact ... This part of analysis involves how you differentiate and combine the data you have retrieved and the reflections you make about this information (p. 56).

Therefore the emergent themes in the qualitative data were identified using codes. These codes were predetermined in light of the literature review, theoretical underpinnings and research questions and focused on the four components of an assessment item. A particular code was attached to a response based on what component of the item was most influential i.e., graphics (G), language (L), item features (IF) or situations (S). A sample of a coding sheet is presented in Appendix H.

Miles and Huberman (1984) also highlighted the fact that there was the need to provide clear operational definitions of the codes used so “they can be applied consistently by a single researcher over time and multiple researchers will be thinking about the same phenomena as they code” (p. 63). While there was only one chief investigator involved, these definitions were useful to maintain coding consistency between periods on which the
data was analysed. Figure 3.9 is a list of definitions for the codes used within the study.

<table>
<thead>
<tr>
<th>Graphics (G)</th>
<th>The student made mention of the graphic within their initial response to the notion of what helped them solve the item.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language (L)</td>
<td>The student seems confused by what the question is asking of them or lacks comprehension of the words used.</td>
</tr>
<tr>
<td>Situation (S)</td>
<td>The student fails to acknowledge and recognise the way the graphic, language and item features work together to create a situation or do not understand the use of the situation in which the mathematics concept is being presented.</td>
</tr>
<tr>
<td>Item Feature (IF)</td>
<td>Students are unintentionally distracted by the use of particular answer options or the positioning of particular aspects of the question.</td>
</tr>
</tbody>
</table>

*Figure 3.9. Definitions of Codes.*

There were a number of advantages to coding the qualitative data. These were that coding:

1. Identified common themes amongst student responses;

2. Assisted in communicating the frequency of students’ reliance upon the use of graphics, language, item features and situations in solving the assessment items; and

3. Increased awareness of future modifications.

Miles and Huberman (1984) recommended the use of coding qualitative results throughout the course of the research as a means of reducing data as well as “noting regularities, patterns, explanations, possible configurations,
causal flows, (and) propositions” (p. 24). It also provided a data set that could be analysed and a means of building a logical chain of evidence. In order to counteract the challenge of utilising a coding structure a number of reliability measures were adopted. First the pilot study data with a N=169 was carefully analysed in order to establish a coding framework. Once the framework was established a usability trial was established in order to determine the rigour of these protocols. Three practising teachers were given a random series of questions and asked to analyse these data according to the protocol framework. The reliability coefficient measured by Cronbach $\alpha$ revealed reliability co-efficients greater than 0.9. Such reliability measures are considered to be more than satisfactory. Of course there was the disadvantage of intracoder reliability where the consistency of the coder could be questioned; however, as the sole investigator was the only one coding the data, this reliability is somewhat strengthened (Johnson & Christensen, 2008).

**Quality and Rigour of the study**

In order to enhance the methodological rigor of the study, the research findings need to be presented in a manner that achieves trustworthiness. According to Graneheim and Lundman (2004), the notions of credibility, dependability and transferability are interrelated aspects of trustworthiness which contribute to the rigour of a study. In particular, they argue that these three aspects of trustworthiness are necessary for studies that utilise content analysis. The rigour of a study is also related to the ethical conduct of the study, with particular consideration of the rights of the participants and the responsibilities and role of the researcher.
Credibility

Credibility refers to confidence in how data and processes address the research questions. One of the ways Graneheim and Lundman (2004) suggest that credibility is enhanced is when participants in the study include both males and females. They also suggest that observations should be considered from different perspectives. In this study, the 143 participants of the quantitative study and the 37 participants involved in the qualitative component were a mix of boys and girls, from varying academic backgrounds, and a reflection of the socioeconomic makeup of the large rural city in one state of Australia. This makeup of participants helped to provide diversity to this study. The one-on-one interviews at both stages of data collection (after the NAPLAN Item Test and the Modified Test) provided opportunities for students to express their understandings and articulate their mathematical reasoning about the test items. The use of digital recordings in the interviews helped to capture these verbal explanations and allowed for repeated access. These recordings were used alongside the more formal quantitative measures to give this data set a voice and offer meaning behind the test scores. This helped to maximise the accuracy of the content analysis process because it provided opportunities to analyse the data from different perspectives, and thus enhance the credibility of the study. Additionally, Graneheim and Lundman (2004) suggest that another critical issue for achieving credible results within the content analysis process comes from selecting the most appropriate and meaningful codes that reflect the context of the situation. Initial coding of data was taken from the theoretical underpinnings of test item construction that were interpreted in light of the interview data. This coding provided the means
and justification behind the modification process and established credible
protocol for future study.

**Dependability and Transferability**

The notions of dependability and transferability relate to the possible
replication of the study and the consistency of the procedures for keeping
thorough notes and records of activities (Hittleman & Simon, 2006).
Graneheim and Lundman (2004) suggest that clear descriptions of culture,
context, and a thorough detailing of a design provide an opportunity for
replication. In this study, dependability was enhanced through verbatim
accounts of student voice to clearly present the link between the theoretical
underpinnings and the analysis process. The digital recordings were
accurately transcribed, and thus, precise descriptions of students’ responses
were analysed. In addition, there was a sole investigator who was therefore
very familiar with the context of the project and the interview sessions and
data collections. To facilitate transferability, specific features of the
participants, the data collection and the analysis process were explicitly
detailed outlining the design of the study.

**Ethics**

According to Flick (2007), there are four principles of research ethics. They
are:

- autonomy—respecting the rights of the individual;
- beneficence—doing good;
- non-maleficence—not doing harm;
- justice—particularly distributive justice or equity (p. 123).
Bearing these in mind, a number of ethical issues were considered in this investigation, specifically ethical issues of confidentiality, the use of pseudonyms in reporting, storage of data, and ethical issues related to the conduct of the researcher (Kumar, 1996).

From the beginning of the study, considerable attention was paid to adhering to the rights of the participants involved. As Kumar (1996) stipulated “it is important to ensure that research is not affected by the self-interest of any party and in not carried out in a way that harms any party” (p. 192). This project aimed not to harm any participants through the research process and addressed the ethical issues in the following way. In terms of confidentiality, all participants were provided with full anonymity. Pseudonyms reflective of the participant’s gender were used in qualitative findings when student voice was required and data records of test results are held solely by the investigator and will not be disclosed to any other parties. Ethical clearance was obtained through Charles Sturt University’s Human Research Ethics Committees as well as the Catholic Diocese. Permission was also sought from school principals and parents of the children involved. In accordance with the Australian code for the responsible conduct of research (Australian Government, 2007) electronic data will be stored on hard drives which are password protected, while the tangible tests will be stored in locked cabinets for 5 years. At all times, the research was conducted in an ethical manner based on the principles of the Australian Association for Research in Education’s (1997) Code of ethics.
Chapter Summary

In this chapter the design of the study, and the methods used to collect and analyse the data have been reported. The research design was framed within a sequential mixed-method design that utilised both quantitative and qualitative data collection and analyses. Due to the innovative nature of this design, the findings of a pilot study were utilised to establish protocols for both data collection and the modification process.

The quality and rigour of this study were achieved by addressing issues of credibility, dependability and transferability. This study conducted the research in an ethically sound manner, taking into consideration issues such as confidentiality, data storage, and the use of pseudonyms in reporting. Ethical approval was obtained to undertake this study from the Charles Sturt University’s Human Research Ethics Committees as well as the Catholic Diocese.

The presentation of results follows in the next two chapters. In Chapter 4 the quantitative data is comparatively analysed with data obtained through qualitative interviews. Three case studies are also presented. In Chapter 5 the results are used to justify the creation of a model based on the contexts surrounding mathematics assessment.
Chapter 4: Presentation and Analysis of Data

This chapter presents the quantitative and qualitative results of the NAPLAN Item Test and Modified Test experimental design. The presentation of each data set is not viewed in isolation due to the nature of the sequential mixed-method design and thus compliments and informs each another. Therefore, the experimental aspects of the design are analysed through multiple analyses of variance (ANOVAs) in order to determine performance differences of students across standard and modified items. Specific performance patterns and relationship are identified through chi square and cross-tab analysis. Student interviews (across the experimental design) are analysed using student voice and interview exerts. Student profiles, through case studies, are then analysed. Due to this form of analysis the chapter is divided into three sections in order to determine:

1. the difficulty of items within the NAPLAN Item instrument;
2. differences between student performance on standard and modified assessment items; and
3. the impact of the modification process.

These matters will be closely examined through three case studies of participants who were overly influenced by modifications to the graphic, language or item features.

Student Performance of Initial NAPLAN Item Instrument

Quantitative Data

This study attempted to investigate the influence of the four elements of item construction on a child’s ability to solve an assessment task. The
NAPLAN Item instrument was used to firstly provide insight into the difficulty of items but also as a process for measuring the outcome of modifications to the graphics, language, item features or situation. Table 4.1 records the percentage correct and standard deviations (SD) in difficulty order of the 15 items presented to the larger cohort of 143 students.

Table 4.1

<table>
<thead>
<tr>
<th>Item</th>
<th>N</th>
<th>% Correct</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting schedule</td>
<td>143</td>
<td>24</td>
<td>.43</td>
</tr>
<tr>
<td>Paper fold</td>
<td>143</td>
<td>25</td>
<td>.47</td>
</tr>
<tr>
<td>Garden plan</td>
<td>143</td>
<td>33</td>
<td>.47</td>
</tr>
<tr>
<td>Scales</td>
<td>143</td>
<td>40</td>
<td>.49</td>
</tr>
<tr>
<td>Fruit juice</td>
<td>143</td>
<td>43</td>
<td>.50</td>
</tr>
<tr>
<td>Boxes</td>
<td>143</td>
<td>60</td>
<td>.49</td>
</tr>
<tr>
<td>Baby mass</td>
<td>143</td>
<td>64</td>
<td>.48</td>
</tr>
<tr>
<td>Pie graph</td>
<td>143</td>
<td>68</td>
<td>.47</td>
</tr>
<tr>
<td>Shoe</td>
<td>143</td>
<td>78</td>
<td>.41</td>
</tr>
<tr>
<td>School camp</td>
<td>143</td>
<td>83</td>
<td>.38</td>
</tr>
<tr>
<td>Pegs</td>
<td>143</td>
<td>83</td>
<td>.38</td>
</tr>
<tr>
<td>Liquorice allsort</td>
<td>143</td>
<td>92</td>
<td>.28</td>
</tr>
<tr>
<td>Coins</td>
<td>143</td>
<td>92</td>
<td>.28</td>
</tr>
<tr>
<td>Street map</td>
<td>143</td>
<td>93</td>
<td>.26</td>
</tr>
<tr>
<td>Spinner</td>
<td>143</td>
<td>96</td>
<td>.20</td>
</tr>
</tbody>
</table>

These results indicated that a number of the items were clearly more challenging than others. It was apparent that the Meeting schedule question
was the hardest with only 34 of the 143 (24%) students answering it correctly, closely followed by the Paper fold item with only 36 students (25%). However there were several items that were also revealed to be quite easy for the students including the Spinner item with 96% correct and the Street map item with a 93% success rate.

The students found 5 of the 15 items much more challenging than the other items presented in the first phase of the study in the NAPLAN Item Test. These 5 items produced means of less than 50% correct and were approximately 20% more difficult than the sixth most difficult item (43% to 60%). Each of these five difficult items required the students to both analyse and interpret a graphic (Paper fold, Garden plan, Scales and Fruit juice) or visualise a graphic (Meeting schedule). With respect to analysing the graphic quite distinct types of processing were required and this demonstrated the importance of decoding and encoding in problem solving.

The Paper fold, Garden plan, Fruit juice and Scales items all contained an information graphic and this graphic needed to be interpreted correctly in order to solve the task. The Scales and the Fruit juice tasks required students to access measurement data (specifically the mass and capacity of objects). For both these items the students needed to recognise the need to incorporate the given information contained within the graphic. The Paper fold and the Garden plan required the students to not only decode information embedded within the graphic but also add to this information by either establishing the rotation of an image (in the case of the Paper fold item) or inserting measurement quantities into the graphic (in the case of the garden task). The other challenging task (Meeting schedule) required the
students to encode a graphic (i.e., imagine a calendar representation in their mind’s eye).

In summary, the students found it most challenging to engage with the decoding or encoding of graphics. By contrast, the 4 easiest items (with 92% or greater) required either the simple recognition of graphic representations (in the case of Liquorice allsort, Coins and Spinner) or the location of a coordinate point (in the case of Street map). Unlike the difficult items the students did not have to apply the information from the graphic to other mathematics or literacy demands. They simply needed to identify an object as a one step process. Thus, the graphics did not require any utility or engagement.

These results are consistent with the Year 5 New South Wales (NSW) results released by ACARA (2010), with the Spinner item included with some of the easier items, with 90% of the state answering correctly, while the Meeting schedule question was categorised as the hardest question in NSW with only 17% getting it correct. This close relationship between the cohort’s results to those achieved by the students of NSW demonstrates them to be representative of the state’s performance. Therefore what the cohort found difficult was aligned to the difficulties experienced by the state and what the state found easy was reflected in the cohort’s results.

Yet while these results revealed the difficulty of the 15 items they failed to identify the students’ mathematical understanding. It was, therefore, necessary to analyse this data in light of the mathematical thinking and understanding that went behind it. As previously addressed, a correct answer does not necessarily indicate sound mathematical knowledge or an incorrect answer failure to use an appropriate mathematical strategy.
(Clements & Ellerton, 1996). It was for this reason that 37 students (approximately 10 from each school) were randomly selected for a interview by way of a more detailed understanding of how students interpret and solve these 15 NAPLAN items.

**Qualitative Data**

In order to identify themes and commonalities between the 37 interview responses, codes were used. Student responses were categorised according to which element of the item, Graphics (G), Language (L), Item Features (IF) or Situation (S), overly influenced the way they answered the question (see Table 4.2).
Table 4.2
Coding analysis of interviews by item component

<table>
<thead>
<tr>
<th>Item</th>
<th>Code (37)</th>
<th>Graphic (G)</th>
<th>Language (L)</th>
<th>Item Features (IF)</th>
<th>Situations (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coins</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pie graph</td>
<td></td>
<td>13</td>
<td>14</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Pegs</td>
<td></td>
<td>28</td>
<td>8</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Liquorice allsort</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boxes</td>
<td></td>
<td>29</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Garden plan</td>
<td></td>
<td>34</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paper fold</td>
<td></td>
<td>25</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spinner</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Street map</td>
<td></td>
<td>34</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>Fruit juice</td>
<td></td>
<td>28</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shoe</td>
<td></td>
<td>11</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School camp</td>
<td></td>
<td>36</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baby mass</td>
<td></td>
<td>1</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meeting schedule</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37</td>
</tr>
</tbody>
</table>

The use of such codes provided insight into the children’s reasoning and problem-solving strategies. It also most importantly substantiated the modification process. Therefore if the answering of an item was overly influenced by the graphic, it was this component that was modified. Over half of the questions were overwhelmingly influenced by only one of the four elements. Consequently, the four item classifications had relatively
discrete attributes that provided insight into student understanding. However two or more attributes predominant on six of the items: namely; Pie Graph, Pegs, Boxes, Paper fold Fruit juice and Shoe). With the Pie graph item, for example, the responses were almost evenly distributed between the graphic, language and item features. However as the language was slightly more influential it was this aspect that was modified.

The coding process also revealed items that could not be modified as it would alter the complexity of the question. These included the Pegs, the Paper fold, the Spinner, the Fruit juice and the School camp items. The majority of the responses for these items were influenced by the graphics, yet a modification of this aspect of the question resulted in a dramatic change in the mathematical understanding that was being measured. It was for this reason that these items were not included in the Modified Test.

The examination of both quantitative and qualitative results for the NAPLAN Item Test provided an analysis of difficult items, influential components, and opportunities as well as restrictions to the modification process. The next section presents the comparative results of the NAPLAN Item Test and the Modified Test.

**Standard versus Modified Instruments—A Tale of Two Tests**

There were three ways the results between the NAPLAN Item Test and Modified Test were analysed. These included a basic pictorial representation, a descriptive cross-tab analysis and analysis of variance (ANOVA) for accuracy of data entry (Tabachnick & Fidell, 2007). The purposes of all sets of analysis were to determine the impact the modification of item design had on student results. These results were then
applied to the interview data received from the NAPLAN Item Test and Modified Test to gain further insight into the responses to the modifications that right or wrong results could not provide.

The first form of quantitative analysis was descriptive in nature (the ‘smiley face’ approach) in which correct and incorrect responses were recorded and analysed. This descriptive analysis provided a visual representation of individual performance across the two instruments by item.

**Analysis 1—Descriptive “Smiley Faces” Approach**

At the completion of both the NAPLAN Item Test and the Modified Test, the results were originally tabulated using smiley faces as a first level of descriptive analysis (See Appendix J). A correct response was recorded as a white smiley face (😊) and a wrong response a grey face (😢). This descriptive analysis provided the initial snapshot on how the modifications impacted on student results. Using these tables, a combined smiley analysis was created and obvious differences were noted (see Table 4.3).

**Table 4.3**

Combined smiley analysis

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Test A (143)</td>
<td>131</td>
<td>97</td>
<td>131</td>
<td>86</td>
<td>46</td>
<td>57</td>
<td>133</td>
<td>112</td>
<td>91</td>
<td>35</td>
</tr>
<tr>
<td>Correct Test B (143)</td>
<td>126</td>
<td>112</td>
<td>119</td>
<td>111</td>
<td>66</td>
<td>63</td>
<td>129</td>
<td>126</td>
<td>98</td>
<td>77</td>
</tr>
<tr>
<td>Difference</td>
<td>-5</td>
<td>+15</td>
<td>-12</td>
<td>+25</td>
<td>+20</td>
<td>+6</td>
<td>-4</td>
<td>+13</td>
<td>+7</td>
<td>+42</td>
</tr>
</tbody>
</table>

From this table it was easy to ascertain the items that had either a dramatic positive or negative change from Test A (NAPLAN Item Test) to Test B (Modified Test). However a descriptive cross-tab analysis of each question
was necessary to establish the relationship between the paired observations on the two variables (Burns, 2000).

**Analysis 2—Cross Tab Contingency Tables**

A cross tab contingency table was produced for each of the ten items. These tables provided descriptive statistics of student performance across the two instruments (through frequency counts). In each table, two cells (diagonally opposite) are highlighted to reveal changed performance from standard and Modified Tests.

For six of the ten items, the changed variance between the tests was greater than 25%. That is, students who either originally answered correctly (standard item) or incorrectly (standard item) now had incorrect (modified) or correct (modified) responses in the second instrument. These included in order of the greatest change Questions 10, 4, 5, 3, 8 and 2. Each of these modifications will now briefly be explored.

Of the 108 students who were unsuccessful with the Question 10 NAPLAN item, 49 (34.3%) got the modified item correct when an information graphic was added. Consequently, of the 35 who were successful with the NAPLAN item, only 7 (4.9%) were unsuccessful with the modified item (Table 4.4).
Table 4.4
Contingency Table for Cross Tab Analysis for performance on NAPLAN Item Test and Modified Test Question 10

<table>
<thead>
<tr>
<th>Q10 Modified</th>
<th>Incorrect</th>
<th>Correct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q10 Naplan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>Count</td>
<td>59</td>
<td>49</td>
</tr>
<tr>
<td>% of Total</td>
<td></td>
<td>41.3%</td>
<td>34.3%</td>
</tr>
<tr>
<td>Correct</td>
<td>Count</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>% of Total</td>
<td></td>
<td>4.9%</td>
<td>19.6%</td>
</tr>
<tr>
<td>Total</td>
<td>Count</td>
<td>66</td>
<td>77</td>
</tr>
<tr>
<td>% of Total</td>
<td></td>
<td>46.2%</td>
<td>53.8%</td>
</tr>
</tbody>
</table>

Table 4.5 highlights that of the 57 students who answered Question 4 incorrectly in the NAPLAN Item Test, 31 (21.7%) responded correctly to the Modified Test when the answer format was changed to multiple choice. While of the 86 students who answered the NAPLAN question correctly only 6 (4.2%) were unsuccessful in the Modified Test.

Table 4.5
Contingency Table for Cross Tab Analysis for performance on NAPLAN Item Test and Modified Test Question 4

<table>
<thead>
<tr>
<th>Q4 Modified</th>
<th>Incorrect</th>
<th>Correct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q4 Naplan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>Count</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>% of Total</td>
<td></td>
<td>18.2%</td>
<td>21.7%</td>
</tr>
<tr>
<td>Correct</td>
<td>Count</td>
<td>6</td>
<td>80</td>
</tr>
<tr>
<td>% of Total</td>
<td></td>
<td>4.2%</td>
<td>55.9%</td>
</tr>
<tr>
<td>Total</td>
<td>Count</td>
<td>32</td>
<td>111</td>
</tr>
<tr>
<td>% of Total</td>
<td></td>
<td>22.4%</td>
<td>77.6%</td>
</tr>
</tbody>
</table>

As a result of moving a graphic to the centre in Question 5 of the modified test, 27 (19%) of the 96 who answered incorrectly in the NAPLAN Item
Test were now correct (see Table 4.6). While of the 46 who correctly answered the NAPLAN Item Test, only 7 (4.9%) responded incorrectly.

Table 4.6

Contingency Table for Cross Tab Analysis for performance on NAPLAN Item Test and Modified Test Question 5

<table>
<thead>
<tr>
<th>Q5 Naplan</th>
<th>Q5 Modified</th>
<th>Incorrect</th>
<th>Correct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect</td>
<td>Count</td>
<td>69</td>
<td>27</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>48.6%</td>
<td>19.0%</td>
<td>67.6%</td>
</tr>
<tr>
<td>Correct</td>
<td>Count</td>
<td>7</td>
<td>39</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>4.9%</td>
<td>27.5%</td>
<td>32.4%</td>
</tr>
<tr>
<td>Total</td>
<td>Count</td>
<td>76</td>
<td>66</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>% of Total</td>
<td>53.5%</td>
<td>46.5%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

The cross tab analysis on Question 3 (Table 4.7) revealed that of the 12 who answered incorrectly on the NAPLAN Item Test, only 5 (3.5%) were correct in the Modified test when an information graphic was removed. This was the only question that saw a significant negative effect of the modifications between the two tests. Consequently of the 131 correct responses in the NAPLAN Item Test, 17 (11.9%) now answered incorrectly in the Modified Test.
When a contextual graphic was removed from Question 8 in the Modified test, 20 (14%) of the 31 who answered incorrectly in the NAPLAN Item Test were now able to give the correct answer (see Table 4.8). The modification also resulted in only 6 (4.2%) of the 112 who originally were correct now calculating a wrong answer.

**Table 4.8**
Contingency Table for Cross Tab Analysis for performance on NAPLAN Item Test and Modified Test Question 8

<table>
<thead>
<tr>
<th>Q8 Naplan</th>
<th>Incorrect</th>
<th>Correct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect</td>
<td>11</td>
<td>20</td>
<td>31</td>
</tr>
<tr>
<td>% of Total</td>
<td>7.7%</td>
<td>14.0%</td>
<td>21.7%</td>
</tr>
<tr>
<td>Correct</td>
<td>6</td>
<td>106</td>
<td>112</td>
</tr>
<tr>
<td>% of Total</td>
<td>4.2%</td>
<td>74.1%</td>
<td>78.3%</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>126</td>
<td>143</td>
</tr>
<tr>
<td>% of Total</td>
<td>11.9%</td>
<td>88.1%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

In Question 2 (see Table 4.9), of the 46 students who answered incorrectly in the NAPLAN Item Test, 25 (17.5%) answered correctly in the Modified test. While of the 97 students who answered correctly in the NAPLAN Item
Test, only 10 (7%) reached a wrong conclusion when the language was simplified in the Modified test.

Table 4.9
Contingency Table for Cross Tab Analysis for performance on NAPLAN Item Test and Modified Test Question 2

<table>
<thead>
<tr>
<th>Q2 Naplan</th>
<th>Incorrect</th>
<th>Correct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2 Modified</td>
<td>21</td>
<td>25</td>
<td>46</td>
</tr>
<tr>
<td>% of Total</td>
<td>14.7%</td>
<td>17.5%</td>
<td>32.2%</td>
</tr>
<tr>
<td>Correct</td>
<td>10</td>
<td>87</td>
<td>97</td>
</tr>
<tr>
<td>% of Total</td>
<td>7.0%</td>
<td>60.8%</td>
<td>67.8%</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>112</td>
<td>143</td>
</tr>
<tr>
<td>% of Total</td>
<td>21.7%</td>
<td>78.3%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

The purpose of the cross tab analysis was to appropriately document the results revealed within the initial “smiley faces” analysis. Using this data it was found that 6 of the 10 items noted changes of more than 25%. The remaining 4 items noted difference of less than 25% between the two tests. This was because any positive changes in student’s results due to a modification were cancelled out by similarly negative student’s results following the modification. Although these changes were not as significant as the other items it was still necessary to report these results to add depth to the research findings as a whole. The items that recorded only slight variances between the two tests were Questions 9, 6, 1 and 7 and will now be explored.

When the item features of Question 9 were modified, 33 of the 52 who answered incorrectly in the NAPLAN Item Test were now correct. However this modification also caused 26 students who answered correctly in the NAPLAN Item Test to answer incorrectly in the modified version.
In Question 6 when an information graphic was removed 20 (14%) students who answered incorrectly in the NAPLAN Item Test were able to correctly calculate Question 6 in the Modified Test. This modification also caused 14 (9.8%) correct students in the NAPLAN Item Test to answer incorrectly in the alternate test.

Table 4.11
Contingency Table for Cross Tab Analysis for performance on NAPLAN Item Test and Modified Test Question 6

<table>
<thead>
<tr>
<th>Q6 Naplan</th>
<th>Incorrect</th>
<th>Correct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect</td>
<td>66</td>
<td>20</td>
<td>86</td>
</tr>
<tr>
<td>% of Total</td>
<td>46.2%</td>
<td>14.0%</td>
<td>60.1%</td>
</tr>
<tr>
<td>Correct</td>
<td>14</td>
<td>43</td>
<td>57</td>
</tr>
<tr>
<td>% of Total</td>
<td>9.8%</td>
<td>30.1%</td>
<td>39.9%</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>63</td>
<td>143</td>
</tr>
<tr>
<td>% of Total</td>
<td>55.9%</td>
<td>44.1%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Of the 12 students who were unsuccessful in question 1 of the NAPLAN Item Test, 11 (7.7%) got the modified item correct when an information graphic was taken away. It was also found that of the 131 who answered
correctly in the NAPLAN Item Test, 16 (11.2%) were unsuccessful with the modified item (Table 4.12).

Table 4.12
Contingency Table for Cross Tab Analysis for performance on NAPLAN Item Test and Modified test Question 1

<table>
<thead>
<tr>
<th>Q1 Modified</th>
<th>Incorrect</th>
<th>Correct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 Naplan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>1</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>% of Total</td>
<td>.7%</td>
<td>7.7%</td>
<td>8.4%</td>
</tr>
<tr>
<td>Correct</td>
<td>16</td>
<td>115</td>
<td>131</td>
</tr>
<tr>
<td>% of Total</td>
<td>11.2%</td>
<td>80.4%</td>
<td>91.6%</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>126</td>
<td>143</td>
</tr>
<tr>
<td>% of Total</td>
<td>11.9%</td>
<td>88.1%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

In Question 7, when the answer options were changed from multiple choice to short answer, 6 out of the 10 students who answered incorrectly in the NAPLAN Item Test were now correct (see Table 4.13). This modification also resulted in 10 correct students in the NAPLAN Item Test now being marked wrong.

Table 4.13
Contingency Table for Cross Tab Analysis for performance on NAPLAN Item Test and Modified test Question 7

<table>
<thead>
<tr>
<th>Q7 Modified</th>
<th>Incorrect</th>
<th>Correct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q7 Naplan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>% of Total</td>
<td>2.8%</td>
<td>4.2%</td>
<td>7.0%</td>
</tr>
<tr>
<td>Correct</td>
<td>10</td>
<td>123</td>
<td>133</td>
</tr>
<tr>
<td>% of Total</td>
<td>7.0%</td>
<td>86.0%</td>
<td>93.0%</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>129</td>
<td>143</td>
</tr>
<tr>
<td>% of Total</td>
<td>9.8%</td>
<td>90.2%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
Although these four items did not demonstrate as much variance between the results of the two tests as the first six they are still an important part of the result story. This is especially relevant when one of the research questions of the study is how does the use of graphics, language, situations and item construction impact on a child’s capacity to make sense of mathematics? This research question focuses on the implications for the individual child on the use of the 4 components of mathematics assessment, and these results may not always be reflected in the larger cohort. It was for this reason that items that may not have been statistically significant in the quantitative data are still reported upon due to their significance within the qualitative data.

The next quantitative analyses performed on the results between the two tests were a series of analyses of variances (ANOVAs) to identify the statistical variance between the two tests.

**Analysis 3—Analysis of Variances**

A series of analyses of variance (ANOVAs) were used to determine if mean differences existed between the NAPLAN items and Modified NAPLAN items. As Burns (2000) notes, the purpose of the ANOVA is to “decide whether the differences between samples is simply due to chance (sampling error) or whether there are systematic treatment effects that have scores in one group to be different from scores in other groups” (p. 294).

An ANOVA revealed a statistically significant difference between the mean scores of students across NAPLAN items and Modified NAPLAN items \[F(1,285) = 10.75 \ p = 0.001\]. The performance of students was significantly higher on the modified items.
By contrast, there was no statistically significant difference between males and females on the variables of the NAPLAN Item Test \( F(1,142) = 1.07, p > 0.05 \) and Modified items \( F(1,142) = 1.75, p > 0.05 \) (see Table 4.14).

**Table 4.14**

Gender means and standard deviations

<table>
<thead>
<tr>
<th>Test</th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=67</td>
<td>N=76</td>
<td>(df 1,142)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAPLAN</td>
<td>6.43</td>
<td>6.61</td>
<td>6.28</td>
<td>.302</td>
</tr>
<tr>
<td></td>
<td>(1.93)</td>
<td>(1.87)</td>
<td>(1.98)</td>
<td></td>
</tr>
<tr>
<td>Modified</td>
<td>7.18</td>
<td>7.41</td>
<td>6.98</td>
<td>.188</td>
</tr>
<tr>
<td></td>
<td>(1.96)</td>
<td>(1.95)</td>
<td>(1.96)</td>
<td></td>
</tr>
</tbody>
</table>

In terms of the significant main effect statistic, subsequent ANOVAs were conducted to determine where the differences were. Of the six statistically significant items, two were graphic modifications, two were item features modifications, one was the result of language modification and one the result of changing the context (see Table 4.15).

**Table 4.15**

Means, standard deviations and univariate statistics for the items on the NAPLAN Item Test and Modified test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN</th>
<th>Modified</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=143</td>
<td>N=143</td>
<td>(df 1,285)</td>
<td></td>
</tr>
<tr>
<td>1 – Coins</td>
<td>.90</td>
<td>.92</td>
<td>.88</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td>(.30)</td>
<td>(.28)</td>
<td>(.33)</td>
<td>p = .329</td>
</tr>
<tr>
<td>2 – Pie graph</td>
<td>.73</td>
<td>.68</td>
<td>.78</td>
<td>4.03*</td>
</tr>
<tr>
<td></td>
<td>(.44)</td>
<td>(.47)</td>
<td>(.41)</td>
<td>p = .046</td>
</tr>
<tr>
<td>3 – Liquorice allsort</td>
<td>.87</td>
<td>.92</td>
<td>.83</td>
<td>4.62*</td>
</tr>
<tr>
<td></td>
<td>(.33)</td>
<td>(.28)</td>
<td>(.38)</td>
<td>p = .032</td>
</tr>
<tr>
<td>4 – Boxes</td>
<td>.69</td>
<td>.60</td>
<td>.78</td>
<td>10.50**</td>
</tr>
<tr>
<td></td>
<td>(.46)</td>
<td>(.49)</td>
<td>(.42)</td>
<td>p = .001</td>
</tr>
<tr>
<td>5 – Garden plan</td>
<td>.40</td>
<td>.33</td>
<td>.47</td>
<td>5.91*</td>
</tr>
<tr>
<td>Component</td>
<td>Mean</td>
<td>Std Dev</td>
<td>t-value</td>
<td>p-value</td>
</tr>
<tr>
<td>-------------------</td>
<td>------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Scales</td>
<td>.42</td>
<td>.40</td>
<td>.44</td>
<td>.51</td>
</tr>
<tr>
<td>Street map</td>
<td>.92</td>
<td>.93</td>
<td>.90</td>
<td>.72</td>
</tr>
<tr>
<td>Shoe</td>
<td>.83</td>
<td>.78</td>
<td>.88</td>
<td>4.96*</td>
</tr>
<tr>
<td>Baby mass</td>
<td>.66</td>
<td>.64</td>
<td>.69</td>
<td>.761</td>
</tr>
<tr>
<td>Meeting schedule</td>
<td>.39</td>
<td>.24</td>
<td>.54</td>
<td>28.27**</td>
</tr>
</tbody>
</table>

* p = < 0.05; ** p = < 0.001

The results of these analyses provided the opportunity for further in-depth analysis at a qualitative level. Under each of the four components, student voice and interview data are analysed in relation to those items that revealed statistically significant difference across the standard and modified forms. In addition, each of the four non-significant results are also reported to establish an understanding of why the modifications did not increase student performance at a statistically significant level. It should be noted that student performance changed in a consistent manner for each of these four non-significant items. That is, the statistically significant graphic item (Liquorice allsort item) decreased across the standard and modified tasks when the information was taken away. Similarly, the Coins item had an information graphic removed across the two instruments which also resulted in a decrease in student performance but in this case the change was not statistically significant. Consequently further qualitative analysis was justified.
Graphics Quantitative Analysis

Results of the ANOVAs revealed that two of three items that were modified according to the graphic were statistically significant to p < 0.05 (see Table 4.16).

Table 4.16
Univariate analysis and modifications of graphics items

<table>
<thead>
<tr>
<th>No.</th>
<th>Item Name</th>
<th>Modification</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Coins</td>
<td>Information graphic removed</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p = 3.29</td>
</tr>
<tr>
<td>3</td>
<td>Liquorice allsort</td>
<td>Information graphic removed</td>
<td>4.62*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p = 0.032</td>
</tr>
<tr>
<td>8</td>
<td>Shoe</td>
<td>Contextual graphic removed</td>
<td>4.96*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p = 0.027</td>
</tr>
</tbody>
</table>

* p = < 0.05

The ANOVAs indicated that there was a statistically significant difference in items where an information or contextual graphic was taken away. It was also found that in another context, the removal of an information graphic did not result in a statistically significant change. The qualitative data was therefore utilised and presented in two ways. Firstly, the two statistically significant items were examined to gain insight into the reason behind the change. Secondly, the qualitative data was used to comment upon why a similar modification impacted differently within items. Once again the reporting of all the data (both significant and insignificant) strengthened the research study as a whole and revealed more about the impact of the modifications on students sense making.
The Impact of Modifying Graphics on Student Performance

*Information graphic removed*

The Liquorice allsort item was modified by excluding the information graphic and replacing it with written data (see Figure 4.1).

The NAPLAN Item Test

This lolly is made with equal layers. The layers are white or black.

What fraction of the lolly is made of black layers?

\[
\begin{array}{cccc}
\frac{2}{5} & \frac{1}{2} & \frac{2}{3} & \frac{3}{5} \\
\circ & \circ & \circ & \circ
\end{array}
\]

The Modified Test

A cake is made with five equal layers. The layers are two white and three pink.

What fraction of the cake is made of white layers?

\[
\begin{array}{cccc}
\frac{2}{5} & \frac{1}{2} & \frac{2}{3} & \frac{3}{5} \\
\circ & \circ & \circ & \circ
\end{array}
\]

*Figure 4.1. The Liquorice allsort item.*

The ANOVA as presented in Table 4.17 revealed that such a change was statistically significant $p < 0.05$. This analysis also revealed that the mean score between the NAPLAN Item Test and the Modified Test items significantly decreased by 11%. This meant that in this context, students
struggled when the information was presented in written form rather than as a graphic.

Table 4.17
Univariate analysis of Liquorice allsort item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN N=143</th>
<th>Modified N=143</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 - Liquorice allsort</td>
<td>.87</td>
<td>.92</td>
<td>.83</td>
<td>4.62*</td>
</tr>
<tr>
<td></td>
<td>(.33)</td>
<td>(.28)</td>
<td>(.38)</td>
<td>p = .032</td>
</tr>
</tbody>
</table>

The interview data revealed that such a decrease in the success of students in the Modified Test was due to the inability to visualise the lolly using the information given and inappropriately focusing on the numbers. This was particularly evident in Kyle’s responses to both items where he originally answered A in the NAPLAN Item Test and then C in the Modified Test:

Jane: How did you get your answer?

Kyle: Well at first I thought it was 2 out of 3 because there’s 2 black layers but then I looked at the lolly and saw there was 5 layers not 3 so then I chose 2 out of 5.

Although Kyle exhibits appropriate mathematical knowledge of fractions in the NAPLAN Item Test, the absence of the information graphic in the Modified Test revealed some flaws in his conceptual understanding when he answers C:

Jane: How did you get your answer?

Kyle: I chose 2 out of 3 because there’s 2 white layers and 3 pink and that’s how I worked it out.

In his response to the NAPLAN Item Test, the information graphic actually prompted Kyle to get the right answer. It was therefore more powerful to represent the information as a graphic rather than contained within words.
This was in spite of the fact that there was no need for any further calculation in the Modified Test compared to the NAPLAN items that required the student to count out the different colours to achieve the answer. By having the figures already written down and not requiring any further analysing made Kyle complacent, negating the need to double check his response.

The Modified Test also required the student to read every aspect to gain an overall understanding as to what the question was asking, while Kyle’s attention was narrowed only to specific aspects. It would seem that the more reading that is required, the more chance there is of children skimming and picking out the information they consider important resulting in incorrect solutions and the creation of ineffective mathematical strategies. Kyle’s responses reveal that he is capable of calculating fractions when represented pictorially but struggles to understand the concept when no picture is attached. While Kyle’s answer was correct on the NAPLAN Item Test it could be argued that his knowledge of fractions is not comprehensive as yet.

Although the removal of an information graphic resulted in a statistically significant change in the Liquorice allsort item, this modification was not as significant when applied to the Coins item (see Figure 4.2).
The NAPLAN Item Test

Gina has only these coins.

She buys a magazine for $1.95.
How much money does Gina have left?

<table>
<thead>
<tr>
<th>$1.00</th>
<th>$1.10</th>
<th>$2.00</th>
<th>$2.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Modified Test

Gina has $4.05.
She buys a magazine for $1.95.
How much money does Gina have left?

<table>
<thead>
<tr>
<th>$1.00</th>
<th>$1.10</th>
<th>$2.00</th>
<th>$2.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.2. The Coins item.

Table 4.18 highlights that although the number of correct responses to the Coin item between the NAPLAN Item Test and the Modified Test declined similar to the Liquorice item it was not as dramatic. This is particularly significant as the student performance in both items in the NAPLAN Item Test were identical (.92). It therefore could not be argued that one item was more difficult than the other.

Table 4.18
Univariate analysis of Coins item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN N=143</th>
<th>Modified N=143</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - Coins</td>
<td>.90</td>
<td>.92</td>
<td>.88</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td>(.30)</td>
<td>(.28)</td>
<td>(.33)</td>
<td>p = .329</td>
</tr>
</tbody>
</table>

134
To understand why a similar modification would impact differently on performance within items of the same difficulty it was necessary to go back through the qualitative data. Kyle’s response to the Liquorice allsort item revealed that many of the children were focusing on the numbers included in the question and were not focusing on the part-whole relationship of the fraction (and thus were not visualising the layers in the lolly). In the Modified coins item it appeared that the children were attempting to mentally calculate the sum in their head and were very reluctant to use the numbers on the page. In fact, only half of the 37 students interviewed utilised an algorithm even though this would have been the most straightforward way of working out the answer.

The interview data revealed that the slight decrease in the success of students in the Modified Test was due to an inability to mentally calculate a subtraction algorithm. This can be seen in Sarah’s responses where she goes from a confident description of her mathematical reasoning in answering D in the NAPLAN Item Test to a confused and illogical explanation of answering B in the Modified Test.

Jane: How did you get your answer?

Sarah: I went to the $1 first and then I counted 95 out of the silver coins and I got $2.10 left.

In this instance Sarah is able to effectively utilise the picture contained within the question to calculate her answer. The absence of the information graphic in the Modified Test requires Sarah to utilise another problem solving strategy in which Sarah attempts to visualise the sum:

Jane: How did you get your answer?

Sarah: Well I went for these 2 (the 5’s on the end of $4.05 and $1.95) and I got 10 so you put a 0 and a 1
there and the 0 and the 9 it would be 10 because of the 5 and 5 and that would be 19 but I looked at the answer and there wasn’t any 9’s on any of them so I just, the closest one was this one...

Jane: So did you work it out in your head?

Sarah: Yes.

There appears to be an internal confusion as to when the best strategy is to visualise and when it is more beneficial to extract the numbers from within the question and create an algorithm. This confusion is reinforced in current assessment design with minimal space provided for working out. By Year 5 most students would be quite proficient in using algorithms and clearly Sarah has a bit of an idea about their format as she starts with the numbers at the end. However her initial instinct is to attempt to solve the item within her head. If the item had encouraged working out through the provision of blank space or even the instructions to ‘please show working out’ Sarah may have been more confident in exercising this strategy than to attempt it in her mind’s eye.

Although this item was not statistically significant it does demonstrate a similar confusion to those witnessed in the Liquorice allsort item, but in this instance the consequences were not as dramatic.

*Contextual graphic removed*

The Shoe item was also modified by taking away a graphic, however in this instance it was a contextual one (see Figure 4.3). This was because 31% of the students made reference to the picture of the shoe despite its irrelevance to obtaining the answer.
The NAPLAN Item Test

This is a picture of a shoe.

Which of these is closest to the length of a real shoe?

5 cm  25 cm  75 cm  100 cm

The Modified Test

Which of these is closest to the length of your shoe?

5 cm  25 cm  75 cm  100 cm

Figure 4.3. The Shoe item.

This meant that the graphic did not contain any information needed to solve the item. The ANOVAs revealed in this context that the change was also statistically significant $p < 0.05$ (see Table 4.19), however the change was a statistical improvement upon the NAPLAN with the mean scores increasing by 10 between the two tests.
Table 4.19

Univariate analysis of Shoe item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN N=143</th>
<th>Modified N=143</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 - Shoe</td>
<td>.83</td>
<td>.78</td>
<td>.88</td>
<td>4.96*</td>
</tr>
<tr>
<td></td>
<td>(.37)</td>
<td>(.41)</td>
<td>(.33)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p = .027</td>
</tr>
</tbody>
</table>

The interview data for the Shoe item was quite insightful as to the possible reason behind such a significant change. It appeared that the picture of the shoe was distracting the children from analysing the question logically. In the NAPLAN Item Test question it is asking the children to estimate the length of a ‘real’ shoe by providing a picture of a shoe that is not close to any dimensions of a ‘real’ shoe. The Modified Test item is also asking them to estimate the length of a ‘real’ shoe but this time providing an accurate representation of a shoe by directing them to look at their own. It was the representations of shoes in the two items that was affecting students’ mathematical thinking and reasoning, as noted in Mikayla’s responses.

Mikayla’s response to the Shoe item in the NAPLAN Item Test was C—75cm. When asked how she reached this conclusion she replied:

Mikayla: Well 5cm is too small for a real shoe and 25cm is sort of a bit small too. 75cm would probably be the size of a real shoe and 100cm would be too big.

Jane: So did the picture of the shoe help you work out your answer?

Mikayla: Yeah.

Jane: How did it help you?

Mikayla: Well if it had of been a picture of a baby shoe it would have been 5cm but because it was a picture of an adult shoe it was 75cm.
In discussing the Modified Test response of her Answer B (25cm), Mikayla provided the following description:

Mikayla: Because 5cm is too small to be my shoe and 75cm is too big and same as 100cm and 25cm was just about the right size.

In Mikayla’s response to the NAPLAN Item Test she refers several times to the term ‘real’, trying to ascertain who the shoe belongs to according to the picture. The picture however is purely a contextual one that plays no part in the solution process, yet Mikayla inappropriately attempts to utilise it as part of her mathematical reasoning. She equates a picture of an adult shoe with a larger measurement and therefore places a large, unrealistic length on it.

Comparatively, the Modified Test item also asks the student to estimate the length of a real shoe but this time by actually looking at a real shoe, rather than using an unrealistic, distracting picture. As a result, Mikayla appropriately applies successful mathematical strategies and carefully eliminates each option till the right answer is reached. Her thought processes are a lot more logical and methodical than previously.

It could be perceived from her original response to the NAPLAN Item Test that Mikayla did not have a sound understanding of measurement by considering a shoe to be close to 75cm. However when investigated further it was revealed that there were aspects of the question that were negatively impacting on her mathematical reasoning. In contrast, the Modified Test revealed that Mikayla’s measurement knowledge was in fact quite acceptable once the distraction was removed and a more accurate context created.
Language Quantitative Analysis

Following the coding of the interview data, it was found that the solution for only one item was overly influenced by the language used and classified as ‘L’. Results from the ANOVAs exposed a statistically significant difference (p < 0.05) when the language in this item was modified (see Table 4.20). While there were other questions where the language was modified, it was as a result of modifying the situation. These modifications were referred to as a decontextualising of the language rather than a result of only changing the language involved and therefore were not classified as language items.

Table 4.20
Univariate analysis and modification of language item

<table>
<thead>
<tr>
<th>No.</th>
<th>Item Name</th>
<th>Modification</th>
<th>F (df 1,285)</th>
<th>p = 0.046*</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Pie graph</td>
<td>Decontextualising language</td>
<td>4.03</td>
<td></td>
</tr>
</tbody>
</table>

* p = < 0.05

The ANOVAs revealed that when the language in an item was simplified there was a significant change in student results. The qualitative data was again utilised to assist in uncovering the reason behind such a change.

The Impact of Modifying Language on Student Performance

Decontextualising Language

The Pie graph item was modified by decontextualising the language used in the stem of the question as well as highlighting the words ‘hours’ and ‘Monday’ (see Figure 4.4). This became apparent in the NAPLAN Item Test qualitative data when 38% were influenced either positively or negatively by the language. The univariate analysis presented in Table 4.21 revealed
that such a change increased the mean score from the NAPLAN Item test to the Modified Test by 10.

**The NAPLAN Item Test**

Hannah made a pie graph to show the number of hours she spent on different activities over 24 hours on Monday.

Which information can be found using this pie graph?

- The number of
  - meals Hannah eats on this day.
  - hours Hannah plays sport each week.
  - hours Hannah watches TV on Tuesday.
  - hours Hannah spends awake on this day.

**The Modified Test**

This pie graph shows the number of hours on Monday Hannah spent doing different activities.

Which information can be found using this pie graph?

- The number of
  - meals Hannah eats on this day.
  - hours Hannah plays sport each week.
  - hours Hannah watches TV on Tuesday.
  - hours Hannah spends awake on this day.

*Figure 4.4. The Pie graph item.*
It was these changes that warranted the term ‘decontextualising’. This concept is comparable to the notion used in contextual graphics where the graphic is used solely to assist in creating a situation. Similarly in this instance descriptive language was used to create a situation. Therefore only the language necessary to solve the item was included and any concepts not necessary in determining an answer were removed. The language was decontextualised. At the same time, these modifications did not impact on the mathematical concept being measured i.e., a child’s ability to extrapolate information from a pie graph.

Table 4.21
Univariate analysis of pie-graph item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN</th>
<th>Modified</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N=143</td>
<td>N=143</td>
<td>(df 1,285)</td>
</tr>
<tr>
<td>2 - Pie graph</td>
<td>.73</td>
<td>.68</td>
<td>.78</td>
<td>4.03*</td>
</tr>
<tr>
<td></td>
<td>(.44)</td>
<td>(.47)</td>
<td>(.41)</td>
<td>p = .046</td>
</tr>
</tbody>
</table>

The interview data uncovered that the increase in student success rate was due to the improved clarity that the Modified Test provided. By removing the extraneous aspects of the question, in particular the notion of Hannah making the pie graph and the use of 24 hours, resulted in a simplified and more direct approach to the question. The succinctness in the question certainly provided clarity within the answers as noted in Celine’s responses when she answered B in the NAPLAN Item Test:

Celine: Well it says “meals Hannah eats on this day”, it doesn’t actually say “how many meals”. And for the hours watching TV on Tuesday, this is Monday. I was a little bit stuck on awake that day, I thought that one (Answer B) was a little more direct.
However when asked to comment on her answer of D to the Modified Test, Celine is able to confidently refute the ineligibility of the other possible answers:

Celine: I said that because I just looked at all the answers and I made sure they were on Monday specifically and so that was the only one that made sense.

Jane: Why didn’t you think it was the first one?

Celine: Because it doesn’t say the number of meals she eats it just says the hours.

Jane: And what about the second one?

Celine: It says each week.

Jane: And the third one?

Celine: That’s on Tuesday instead of Monday.

It is interesting to note that in Celine’s response to the NAPLAN Item Test, she highlights the importance of Monday in ascertaining her answer in one aspect but not in another. She makes reference to the data in eliminating answer C but does not apply the same logic when choosing answer B. Comparatively in the Modified Test, when questioned on all possible correct responses, her elimination process was correctly justified.

In this instance it appears that the length of the question in the NAPLAN Item Test required too much working memory to be able to retain the original understanding so by the time she reached her conclusion it was clouded and incorrect. However in the Modified Test, by using only the necessary information verbatim, Celine was able to process all the information correctly and methodically. Therefore it could be said that this item was not necessarily testing Celine’s ability to read a pie graph but rather how much information she could retain and utilise.
Item Features Quantitative Analysis

Four out of the 10 Modified NAPLAN questions were modified according to the features included in the item. These features involved examining the answer stem, location of any graphic or positioning of the question. The modifications therefore involved changing the answer type from multiple choice to constructed-response and vice versa, repositioning the graphic within the item and changing the answer options provided in a multiple choice item. However the ANOVAs identified only two of these changes as statistically significant, one at p < 0.05 and the other highly significant at p < 0.001 (see Table 4.22).

Table 4.22
Univariate analysis and modifications of item features

<table>
<thead>
<tr>
<th>No.</th>
<th>Item Name</th>
<th>Modification</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Boxes</td>
<td>Changed from constructed-response to <strong>multiple choice</strong></td>
<td>10.50** p = .001</td>
</tr>
<tr>
<td>5</td>
<td>Garden plan</td>
<td><strong>Repositioning of graphic</strong> within the item</td>
<td>5.91* p = .016</td>
</tr>
<tr>
<td>7</td>
<td>Street map</td>
<td>Changed from multiple choice to constructed-response</td>
<td>.72 p = .395</td>
</tr>
<tr>
<td>9</td>
<td>Baby mass</td>
<td>Changed <strong>multiple choice options</strong></td>
<td>.761 p = .384</td>
</tr>
</tbody>
</table>

* p = < 0.05; ** p = < 0.001

These subsequent univariate results revealed statistically significant differences between two of the four items; namely the Boxes item (p=.001) and the Garden plan item (p=.016). It also revealed that there was no
statistically significant change to the Street map and Baby mass items due to their modifications. Further analyses of the two significant items followed by the non-significant items are explored in the following sections utilising qualitative data.

**The Impact of Modifying Item Features on Student Performance**

*Changed from constructed-response to multiple choice*

The Boxes item was originally coded as ‘G’ (graphic) according to the interview data as understandably most students relied on the graphic to calculate their answer. However in this instance it was difficult to actually change the graphic in any way without changing the nature of the task. Yet this item was one of the few that were represented in constructed-response (short answer) compared to the more popular multiple choice in the NAPLAN Item Test. It was for this reason that the answer format was changed (see Figure 4.5).
The NAPLAN Item Test

![Image of the NAPLAN Item Test]

How many small boxes can fit in the carton altogether?

12 16 20 24

The Modified Test

![Image of the Modified Test]

How many small boxes can fit in the carton altogether?

12 16 20 24

Figure 4.5. The Boxes item.

As a result, the change was highly significant $p < 0.001$ with the mean score from the NAPLAN Item Test improving by 18% (see Table 4.23).

Table 4.23
Univariate analysis of Boxes item in NAPLAN and Modified NAPLAN

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN N=143</th>
<th>Modified N=143</th>
<th>F (df 1, 285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 – Boxes</td>
<td>.69</td>
<td>.60</td>
<td>.78</td>
<td>10.50**</td>
</tr>
<tr>
<td></td>
<td>(.46)</td>
<td>(.49)</td>
<td>(.42)</td>
<td>p = .001</td>
</tr>
</tbody>
</table>
Presumably, the benefit of modifying an item into multiple choice is that the probability of correctly answering the question greatly improves. It also means that the correct answer is provided and facilitates the opportunity to strategically work from the answers rather than the question for the result. While the quantitative results support these hypotheses, they also resonated loudly in the qualitative data with responses similar to Harrison’s when he was asked how he calculated his answer of 11 in the NAPLAN Item Test:

   Harrison: My reason being because altogether in there there are 5 and there are spaces and up the top there is 1 space and there are 3 spaces up from the one below and 4 boxes going up from the bottom and 3 boxes going up from the first block.

The inclusion of the multiple choice options prompts Harrison to go beyond counting to 11 to include the missing 5 boxes and to answer 16 in the Modified Test:

   Harrison: As it says there are 5 small boxes already in there so I put 5 and there’s no box at the bottom of one of the pictures so I worked out it was 4 going up, then 3, then I then 3 and that equals 11 and altogether that equals 16.

In both the NAPLAN Item Test and the Modified Test, Harrison comments on the presence of the five boxes already in the carton. Yet in the NAPLAN Item Test he fails to add them to the total giving him an incorrect answer of 11. However it seems providing the multiple choice options in the Modified Test serves as a reminder to include the original 5.

The strategies Harrison uses to calculate the number of boxes in both tests are almost identical so the modification did not affect the way he attempted to solve the items. It was just the final calculation at the end that resulted in an incorrect response to the NAPLAN Item Test and the assumption that he was unable to visualise and calculate the number of boxes necessary to fill
the carton. Once again it would seem that without the qualitative data, an incorrect conjecture about Harrison’s spatial understandings would have been made.

**Repositioning graphic within the item**

In the Garden plan item, the presentation and layout of the question was modified by centring and rotating the graphic to a more prominent position (see Figure 4.6).

---

**The NAPLAN Item Test**

This is the plan of a garden.

What is the perimeter of the garden?

- 36 m
- 64 m
- 68 m
- 72 m

**The Modified Test**

This is the plan of a garden.

What is the perimeter of the garden?

- 36 m
- 64 m
- 68 m
- 72 m

*Figure 4.6. The Garden plan item.*
Such a change to the item features was statistically significant and improved the mean score of the NAPLAN Item Test by 14 (see Table 4.24).

Table 4.24
Univariate analysis of Garden plan item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN N=143</th>
<th>Modified N=143</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 - Garden plan</td>
<td>.40</td>
<td>.33</td>
<td>.47</td>
<td>5.91*</td>
</tr>
</tbody>
</table>


Table 4.24 (continued)

<table>
<thead>
<tr>
<th></th>
<th>(49)</th>
<th>(47)</th>
<th>(.50)</th>
<th>p = .016</th>
</tr>
</thead>
</table>

The analysis of the interview data revealed that the students seemed to overlook incorporating all sides into calculating the perimeter on the NAPLAN item, possibly due to its close positioning to the question stem and answers. This was in spite of the fact that when asked the definition of a perimeter all students correctly described it as the distance all around the shape. An example of this can be seen in Elise’s response to the NAPLAN Item Test when questioned on how she got the Answer A:

Elise: 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.

Jane: What does perimeter mean?

Elise: The outside of an object.

By moving the graphic away from the clutter in the Modified Test, the qualitative data exposed a heightened awareness to include all sides in the equation as evident in Elise’s response when discussing her answer of D:

Elise: 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.
According to Elise’s responses, the format of the NAPLAN item clearly impacted on her ability to solve the task. By placing the graphic too close to the question, Elise fails to notice the need to include those sides that are blank. However when the sides of the graphic that need to be calculated are clearly visible, she can effectively determine their value and include them within the perimeter. It can be concluded, therefore, that the location of the graphic within an assessment item impacts on a child’s ability to solve the task. This is particularly relevant with the prevalence of graphics in mathematics assessment tasks.

*Changed from multiple choice to constructed response*

Since one of the modifications carried out within the study was the changing of item responses from short answer to multiple choice (the Boxes item) it seemed beneficial to explore the alternative and any possible changes in results. It was for this reason that the Street map item was changed from multiple choice to short answer response (see Figure 4.7)
It has been found that when an item was changed from short answer to multiple choice students’ results improved (the Boxes item). If this is the case it could be assumed that the opposite should occur when the alternate modification is carried out with a decrease in students’ performance. The
univariate results for the Street map item revealed this to be true but only marginally with no statistically significant change (see Table 4.25).

Table 4.25
Univariate analysis of Street map item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN (N=143)</th>
<th>Modified (N=143)</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 - Street map</td>
<td>.92</td>
<td>.93</td>
<td>.90</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>(.28)</td>
<td>(.26)</td>
<td>(.30)</td>
<td>p = .395</td>
</tr>
</tbody>
</table>

Although there was only a slight decrease in student performance, it was a decrease, and assists in substantiating the data surrounding the use of multiple choice versus short answer response. That is, multiple choice items could be considered an easier answer option to written responses. The reason behind only a slight change in this instance could be that the Street map item was already considered to be a relatively easy question. In the original NAPLAN Item Test data it was found to be the second easiest item out of the 15 with a mean of .93. This could indicate that students were already exhibiting a sound mathematical understanding of grid references and the modification of answer type was not going to impact too dramatically. On the very few that it did impact on it can be seen that without the multiple choice options students are more prone to make careless mistakes as they have no opportunity to check if their answer is correct. This can be seen in Amy’s responses to the Street map item. In a few words she was able to describe her answer of D in the NAPLAN Item test:
Amy: I chose Summer and Fox because C4 stops there and they’re the 2 streets on the corner.

In the Modified Test Amy started at D4 with her fingers but went crooked and arrived at the streets contained within C4. This could have been a careless error or maybe an indication that she had remembered the NAPLAN version of the question which asked for the streets within C4 and not D4 like in the Modified. It could be argued that either way the absence of the multiple choice options meant she was unable to check her answer and consequently she reached the wrong conclusion.

Although this analysis was on an item that did not yield statistically significant data, it has to be useful in substantiating and confirming those items that did yield such results. It was for this reason that the following Baby mass item results have also been presented within the analysis chapter.

**Changed multiple choice options**

Although this was the only example of modified multiple choice options used in this research, the findings are considered important as 98% of the students noted the use of this feature. The results surrounding the Baby mass item were unexpected. This was primarily based on the reasons behind the modification. As noted in Chapter 3, the Baby mass item was modified due to an obvious mismatch between children’s correct external representation and their incorrect internal reasoning and thinking processes. Children were getting the correct answer not because of sound mathematical understanding but because they viewed the kilograms as larger than grams despite the number that was in front. It was for this reason that the multiple choice options were modified (see Figure 4.8).
The NAPLAN Item Test

These babies were born on the same day.
Which baby has the greatest mass?

Simon 3.5 kg
Georgia 3.450 g
Max 3.05 kg
Oscar 3.05 g

The Modified Test

These babies were born on the same day.
Which baby has the greatest mass?

Simon 3.4 kg
Georgia 3.490 g
Max 3.05 kg
Oscar 3.050 g

Figure 4.8. The Baby mass item.

It was anticipated that because students were getting the right answer with incorrect reasoning, the modification would result in a poorer student performance because their reasoning would no longer match a correct response. However it was found that in the slight change of performance between the two tests, students improved in the Modified Test. This is represented in Table 4.26.
Table 4.26
Univariate analysis of Baby mass item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN N=143</th>
<th>Modified N=143</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 – Baby mass</td>
<td>.66</td>
<td>.64</td>
<td>.69</td>
<td>.761</td>
</tr>
<tr>
<td></td>
<td>(.47)</td>
<td>(.48)</td>
<td>(.47)</td>
<td>p = .384</td>
</tr>
</tbody>
</table>

It is difficult to explain why this improvement between the two tests occurred. In this instance the qualitative data provides no themes or commonalities between the students’ reasoning. There were some stories such as Lucy’s that exemplified the anticipated outcome but it cannot be considered reflective of the cohort. Lucy was one of the students who answered the NAPLAN Item Test correctly, but not based on sound mathematical understanding as demonstrated in her explanation:

Lucy: I chose Simon.

Jane: Why did you think Simon?

Lucy: Well Georgia and Oscar are only grams and I didn’t think Mia because she was 05 and Simon was 35.

When the NAPLAN was modified, Lucy’s internal representation was appropriately represented externally with an incorrect response of A:

Lucy: Well Georgia and Oscar were grams and Mia was 05 and Simon was .4 so I chose Simon.

Although definitive reasons to the overall results of Question 9 cannot be provided, Lucy’s story does highlight the ambiguity of NAPLAN results. This is particularly relevant in today’s context of high-stakes testing and increased accountability. If students’ results are going to be used as a means for assessing schools and teachers, we need to ensure that they are indicative...
of what a child does or does not know. If this is not the case, or is impossible to achieve, then maybe they do not warrant their current status and importance. Although quick to acknowledge the relevance of a national assessment in schools today, it is more the emphasis that is placed upon it that is concerning, especially in light of stories such as Lucy’s.

**Situation Quantitative Analysis**

To create a situation within a question involves the use of graphics and language. Therefore to alter the situation would involve modifying either of these components. It was for this reason that an information graphic was added or removed to change the situation of two questions in the NAPLAN Item Test. Only one of these changes resulted in a highly statistically significant result of p < 0.001 (see table 4.27).

*Table 4.27*

Univariate analysis and modification of context items

<table>
<thead>
<tr>
<th>No.</th>
<th>Item Name</th>
<th>Modification</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Scales</td>
<td><strong>Information graphic removed</strong> to alter the situation</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p = .474</td>
</tr>
<tr>
<td>10</td>
<td>Meeting schedule</td>
<td><strong>Added information graphic</strong> to alter the situation</td>
<td>28.27**</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p = .000</td>
</tr>
</tbody>
</table>

** p = < 0.001

The addition of an information graphic in the Modified Test dramatically altered student responses to Question 10. However the altering of the situation created by the removal of a graphic did not significantly impact the results in Question 6. Both of these questions will be analysed through the qualitative data.
The Impact of Modifying the Situation on Student Performance

*Added information graphic to alter the situation*

The Meeting schedule item was modified by the inclusion of an information graphic (see Figure 4.9).

The NAPLAN Item Test

A meeting is held on the first Tuesday of each month. There was a meeting held on 6 March. What is the date of the April meeting?

April

The Modified Test

A meeting is held on the first Tuesday of each month. There was a meeting held on 6 March. What is the date of the April meeting?

April

![Figure 4.9. The Meeting schedule item.](image)

This change resulted in the mean score of the Modified Test improving by 30 from the NAPLAN Item Test (see Table 4.28)

*Table 4.28*

Univariate analysis of Meeting schedule item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN</th>
<th>Modified</th>
<th>F</th>
</tr>
</thead>
</table>

157
The Meeting schedule item was one of the hardest items within the NAPLAN Item Test. While trying to reflect a ‘real-life’ scenario, the reality is that not many adults would encounter such a situation without the use of a calendar. This resulted in students utilising inappropriate strategies with many of the students like Zachary calculating the answer to be the 7th with the following explanation:

Zachary: I worked out that April is, I mean March is one with 31 days so I added that to, I added 30 days and then added 1.

It was for this reason that a calendar was added to the Modified Test. Such a change resulted in more appropriate problem solving strategies as demonstrated within Zachary’s interview when he correctly calculated the 3rd as the answer:

Zachary: If it’s on the first Tuesday I went to the 31st which is the last day and then I added all the ones up to Tuesday.

In the NAPLAN Item Test it appeared many students like Zachary attempted to work out the answer using any relevant information they knew about months in the year. As a result several children also calculated an answer of the 7th because March has 31 days but April only has 30. This meant that they added one on from the 6th. This inappropriate use of their understanding of time indicates that many of the students had not encountered such a situation within the classroom. Although as previously
discussed, this circumstance is not one that many would encounter without the presence of a calendar.

This item also required the students to encode rather than decode a graphic. It could be assumed that within the classroom more time would be dedicated to teaching students how to retrieve information from or decode a calendar than imagining a calendar representation in the mind’s eye. Subsequently when a calendar is added to the question in the Modified Test, student success rate rose as well as the use of appropriate problem solving strategies. These findings indicate a mismatch between the emphasis and use of graphics within the classroom and the expectations of students’ use with graphics within high-stake testing. If children are required to encode graphical representations it cannot be assumed that this knowledge will simply be innate.

It could be argued in this instance that the inclusion of the calendar as a modification to making the situation more realistic has made the item easier. However while there was a significant statistical difference between the two tests, the Modified Test item was still challenging enough for only half the cohort to answer correctly (see Table 4.15). This also indicates an increased likelihood of the Modified Test item being similar to a situation addressed within or outside the classroom making it accessible for more students to demonstrate their understanding of time and highlight those who are struggling.

In the Meeting schedule item, an information graphic was added to assist in creating a new and familiar situation which resulted in a statistically significant change in students’ results. Because of this it was also anticipated that the removal of an information graphic in the Scales item to
alter the situation would also result in a statistical change but this was not the case. This item will now be explored along with possible arguments to aid in enlightening these results.

**Removed information graphic to alter the situation**

The situation created within the Scales item was to resemble buying fruit at the supermarket and using the scales and prices provided to calculate the cost. In order to alter the situation the scales were removed in the Modified Test (see Figure 4.10).

---

**The NAPLAN Item Test**

The price of oranges is $6 per kilogram (kg).
The cost of 10 oranges is closest to

- $6
- $15
- $25
- $60

---

**The Modified Test**

One apple weighs 0.250 kg.
If the price of apples is $6 per kilogram (kg), the cost of 10 apples is closest to

- $6
- $15
- $25
- $60

*Figure 4.10. The Scales item.*
As previously discussed in Chapter 3, the modification of the Scales item was the result of students’ inattentiveness to utilising the information contained in the graphic in their problem solving. It was anticipated that students’ results would improve in the Modified Test if the details contained within the information graphic were comprehensively detailed within the question. Although students’ results did improve slightly, it definitely was not within the anticipated capacity (see Table 4.29).

Table 4.29
Univariate analysis of scale item in NAPLAN Item Test and Modified Test

<table>
<thead>
<tr>
<th>Item type</th>
<th>Total</th>
<th>NAPLAN N=143</th>
<th>Modified N=143</th>
<th>F (df 1,285)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 - Scales</td>
<td>.42</td>
<td>.40</td>
<td>.44</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>(.49)</td>
<td>(.49)</td>
<td>(.50)</td>
<td>p = .474</td>
</tr>
</tbody>
</table>

The cross-tab analysis of the Scales item (Table 4.11) does shed some light on the activity surrounding the question when attempting to explain the results. It reveals a large number of students went from wrong to right responses between the two tests but almost just as many students went from right to wrong. Consequently, any changes in student performance cancelled each other out. However it is still necessary to acknowledge these changes as they are in very nature to changes to children’s performance and aids in answering the research question to what extent does assessment reflect sense making and mathematical understanding?

Even though the quantitative data reveals no statistical significance between the results of the test, the qualitative data definitely exposes differences in children’s thinking as a result of the changes. This was particularly evident
in Kate’s responses. In the NAPLAN Item Test Kate’s explanation mirrored others when she answered D:

Kate: I times 6 by 10 and got 60 and added the dollar sign.

Jane: Did you use the picture to help work out the answer?

Kate: No.

However when the information contained within the graphic was incorporated into the question, Kate was able to appropriately utilise it and calculate the correct response of B:

Kate: I know that 250 is a quarter of 1000 so I just went like 250 + 250 + 250 + 250 which is 1000. So I just did that another time which makes it up to 8 apples and that’s just 2 more apples left over and I added all the $6 together with an extra 3 and got $15.

Kate’s transformation in the way she calculated her answer is a clear example of the dramatic differences that occurred across the respective instruments—that is, the removal of the graphic to alter the situation resulted in fundamentally different understandings of the task. Kate’s conceptual understanding had not changed across the standard and modified instruments—what had changed was her interpretation of what the task was asking. Thus, these qualitative episodes provided an important level of analysis that was not exposed if the quantitative results had been judged only on significant probability changes.

**Summary of Quantitative and Qualitative Data of the Two Tests**

The multiple ANOVAs on the two tests revealed statistically significant results when:
1. An information graphic was removed;

2. A context graphic was removed;

3. The language was simplified;

4. A constructed-response item was replaced with multiple choice;

5. An information graphic was moved; and

6. An information graphic was added.

These quantitative results of the larger cohort were viewed in light of the interview data. A number of further issues also emerged from the qualitative interview data.

The first was the beneficial use of information graphics in assessment items. It was found that students were more inclined to rush over reading vital aspects of a question compared to when this information was presented as a graphic. It also decreased the likelihood of the students incorrectly decoding the information contained within the question when it was represented as a picture. With the high prevalence of information graphics in assessment items today, in fact in the classroom and society as a whole, students seem to have developed a familiarity and confidence with their use in mathematics. This was reinforced with the positive results received when an information graphic was added to make an item more realistic and practical. However it was their discernment between those graphics that was important. This included an inability to note the importance of information graphics compared to the insignificance of contextual graphics which proved problematic.

Although the presence of contextual graphics in assessments is minimal, it appears that when they are utilised students find them confusing and
distracting. The interview data revealed that students placed unwarranted importance on contextual graphics and failed to understand their use within the question. Understandably, when attempting to describe a ‘real’ item, it was more beneficial for the student to draw their attention to an actual representation of the real item compared to a picture that may have a number of connotations.

The decontextualising of the language used was also to the child’s benefit. It appeared a more clearly defined question resulted in a more articulate and well thought out response. It also decreased the possibility of measuring a child’s working memory or cognitive load rather than their mathematical understanding.

The results also revealed an overwhelming improvement when questions used a multiple choice format compared to a constructed-response item (short answers). It was found that the multiple choice option served as a self-checking tool for students as it allowed for double checking responses with the answers provided. Subsequently, this resulted in careful computations and minimised miscalculations and errors.

The interviews also highlighted the importance of the location of the graphic within an assessment item. From the students’ responses it was noted that less attention was paid to the graphic if it was too closely aligned to the question and the answer options. It appeared that despite a concrete understanding of the definition and concept of perimeter, the placement of the graphic negatively impacted on their ability to solve the task. Centreing the graphic away from superfluous information resulted in positive results that reflected their mathematical understanding.
All of these qualitative insights into the quantitative data have provided a story behind the modification process of each of the 10 NAPLAN items. Due to the nature of the research, it also provided data for the creation of three case studies based on students who were overly influenced by one of the features of mathematics assessment. These cases will be examined to provide more sophisticated insights into the impact of particular aspects of test item design on children’s mathematical reasoning.

**Student Case 1 (Max): The challenges of decoding graphics**

Max was a student who exhibited a substantial improvement between tests. In the NAPLAN Item Test Max answered 4 out of 10 questions correctly while in the Modified Test he received a score of 7 out of 10. When investigated further, the three items that Max improved upon were all modified according to the graphic or the positioning of a graphic within the question.

**The Coin Item**

The first of the three items which Max improved upon was the use of the coins in Question 1 of the NAPLAN Item Test (see Figure 4.11).

![Figure 4.11. Max—Coins.](image)
It is important to note that when explaining his answer of C Max never indicated how much money he thought was represented in the coins:

Max: ‘Cause she has that much money and she spends $1.95 and she ends up with $2.

Jane: So what did you use in the question to help get your answer?

Max: The picture helped.

Jane: How did you use the picture?

Max: If you didn’t know how much money she had you couldn’t work it out.

This resulted in a non-specific explanation of how he calculated his answer and exhibited no clear mathematical thinking or reasoning. This also made it quite difficult to ascertain the exact cause of his incorrect answer, whether it was the result of miscalculating the amount of money represented or an error in the subtraction process. Either way, taking the graphic away and presenting the algorithm numerically resulted in a more sophisticated and logical response in the Modified Test (see Figure 4.12).

![Graphic](https://via.placeholder.com/150)

**Figure 4.12.** Max—Modified coins.

The absence of the graphic (see Figure 4.12) allowed Max to logically work his way through the algorithm in his head by using valid mathematical strategies. The absence of the picture decreased the possibility of miscalculation as it eliminated the initial step of counting the coins, thus allowing Max to focus only on the algorithm involved and answering $2.10:
Max: Because if you take away the 5c first you’ll have $4 and then you take the $1 off that, that leaves $3 and then you take 90c off that and you get $2.10.

It could be argued that eliminating this step made the item easier, however it was found in the quantitative data that the mean score for the Modified Test decreased (see Table 4.4). While this change was not statistically significant it can be used to verify that the modification in fact did not make the item easier for the majority of students as there was a decline in correct responses. For Max, however, the item was more accessible without the pictures. Similarly, Max found Question 8 from the NAPLAN Item Test easier when a contextual graphic was removed (see Figure 4.13).

The Shoe Item

Max’s response to Question 8 in the NAPLAN Item Test was unlike others involved in the study as it did not indicate that he was picturing an adult’s shoe when he incorrectly answered C (although he originally answered B but crossed it out):

Max: Because if a real shoe is 5cm that’s very little. If it was 75 it would be pretty close to a real shoe.

Jane: What kind of shoe are you thinking about?
Max: Maybe runners.

Jane: That belong to?

Max: Me.

Jane: Did the picture help you at all?

Max: No not really.

Jane: What were you picturing in your head?

Max: Nothing really.

Jane: Why did you change your mind?

Max: I just like forgotted a real shoe, I thought it was like a fake shoe.

He was however confused as to what kind of shoe it was, a real shoe or a fake one. He believed that the picture was that of a fake shoe, yet the question asked him the length of a real shoe. While his answer does not define what constitutes a ‘fake shoe’ he does mention the word ‘real’ repeatedly throughout his response. Attempting to utilise the picture in his answer creates conflict between what is being asked in the question and what is being represented in the picture.

While his answer may not be exactly the same as others examined in the study, in both circumstances it is their over reliance and inappropriate importance placed on the contextual graphic that impacts on their mathematical reasoning.

Similar to other interview responses, when directed to look at his own ‘real’ shoe (see Figure 4.14), Max is able to logically apply appropriate lengths.
Max acknowledges looking at his ‘real’ shoe as he clarifies his thinking when questioned on why he thought B was the correct answer in the Modified Test:

Max: Cause I just looked at my shoe and thought it would be 25. It wouldn’t be 5cm cause that’s a bit too short, 75cm is a bit too big and 100 cm is way too big.

This exemplifies the importance of incorporating a context that a child can relate to. Max’s response highlights the benefits of representing a real shoe as realistically as possible rather than as a small, unrealistic, out of proportion graphic. It also calls for the re-examination of contextual graphics included in mathematics assessment.

**The Garden Plan Item**

Also similar to the larger cohort, Max benefited from the repositioning of the graphic in question 5 of the NAPLAN Item Test (see Figure 4.15).
Figure 4.15. Max—Garden plan.

The close proximity of the graphic to the other elements of the item diminishes Max’s capacity to view all sides of the garden. This resulted in Max quickly arriving at the answer of 36 and providing the following explanation:

Max: Because I went to 16 first and there’s another 16. And then I counted all of them and then I pushed that one out and then I just worked them 4, 2, 6 and that.

From Max’s responses, he appears to be a quick worker who immediately assesses the necessary and relevant information. In this instance he is quick to identify the sides containing measurements as well as the one that does not require any further calculation (the opposite side to 16). However in his calculation he only includes the numbers that are visible, especially when they add up to one of the multiple choice options. Yet when the graphic is moved and the need to calculate the sides is more obvious, he takes the time to work out the relevant information rather than assuming it is all contained within the question (see Figure 4.16).
The difference in Max’s logical reasoning between the NAPLAN Item Test and the Modified Test in this item is astounding as he correctly arrives at the answer of 72:

Max: I added up these two and I got 32 and I put it down there and then I added 6+4+2 and then I got 24 cause and then I just added these 2 and I got 16 and then I just added them all together \((32+24+16=72\) written on page).

By moving the graphic it emphasises the inclusion of all the sides and the possible need for further calculations. These findings resonate with those of the larger cohort who also benefited from the repositioning of the graphic in this question.

**Summary**

These results indicate that Max is particularly influenced by the quality and use of graphics within assessment items. Of particular interest is his
preference for word problems devoid of information graphics contrary to the findings of the larger cohort.

His results therefore significantly improved when the graphics, both information and contextual, were taken away. His responses suggest that he favours questions that involve one-step analysis as it decreases the possibility of miscalculation and he can quickly ascertain the answer. Max appears to be a very literal thinker in the sense that he tries to incorporate every aspect of the question into his answer even if it is inappropriate. It is for this reason that Max benefits from questions that only include the relevant information that is easily identified and appropriately presented. Results like Max’s highlight the need for careful thought and analysis into the presentation and necessity of graphics within assessment items.

This is also applicable to the language that is used in mathematics assessment as Emily’s results indicate.

**Student Case 2 (Emily): The challenges of literacy demands**

There was only one item within the NAPLAN Item Test that was classified as being modified according to the language used. However the modification of a context within another item could also inadvertently change the role of the language and therefore the child’s ability to solve the task. Emily’s results indicated that language was particularly pertinent in achieving positive results. Her score from the NAPLAN Item Test to the Modified Test improved by two as a result of modifications to the language used within two items.
The Pie Graph Item

Emily’s response to question 2 (Figure 4.17) was similar to that of the larger cohort.

Before even analysing Emily’s response to the NAPLAN Item Test it is evident that the question contains unnecessary information not pertinent to solving the task. The use of a situation makes the question lengthy and complicated. This is reflected in Emily’s response where she is unable to apply the information contained within the question to the answer options. By the end of her explanation she is confused and admits defeat and answers B:

Emily: Because it shows how many hours and meals she eats this day it doesn’t show how many meals it shows hours. And I didn’t do hours Hannah watches TV on Tuesday because it doesn’t say Tuesday. And I didn’t do hours Hannah spends awake on this day because it says here, hang on. Um oh 24hrs is oh no it has sleep down, I don’t know I just did that one.
It seems that the use of the term 24 hours to assist in creating a situation perplexes Emily as she tries to apply it within her explanation. Once again this highlights the benefit of decontextualising the language involved in mathematics assessment so students have the opportunity to demonstrate accurately their mathematical understanding. This is evident in Emily’s response to the equivalent Modified Test item (see Figure 4.18).

![Figure 4.18](image)

**Figure 4.18.** Emily—Modified pie graph.

It seems that the uncomplicated structure of the question is reflected in the concise way Emily applies it to all of the answer options. She is able to logically dismiss inappropriate answers leaving her with the only suitable response of D:

Emily: And I didn’t do meals Hannah eats on this day because it says the number of hours not meals. And the hours Hannah plays sport each week, each week and it says only on Monday and the hours Hannah watches TV on Tuesday and it says only Monday. So I thought it had to be hours Hannah spends awake on this day.

Emily appears to be a child very capable of applying her mathematical understanding provided the question is straightforward and she can quickly
ascertain what is being asked of her. This was also evident in the other language item that Emily improved upon.

**The Scales Item**

The Scales item (Figure 4.19) was modified according to the situation as it was coded ‘S’ (see Table 4.2) from the interview data. To change the situation used in the item the information graphic was removed and replaced with a word description.

![Figure 4.19. Emily—Scales.](image)

The reason behind such a change was that students were ignoring the role the graphic played within the situation and therefore were not utilising it in their solution process. In this instance the information graphic was being viewed as a contextual one and subsequently deemed unnecessary.

An example of this can be seen in Emily’s response to the item. It can be seen from Emily’s response that she appears ignorant of the weight measured on the scales and the role it plays in solving the problem. While acknowledging that the graphic may have been important initially, she later dismisses it when she is unable to recognise how it could be utilised in the solution process as she answers D:
Emily: Because it was $6 per kilogram and the cost of 10 oranges is closest to, I did 6x10 yep and I got 60 and 60 was down there.

Jane: Did you need to use the picture at all to help you?

Emily: No. At first I used it and then I didn’t think I needed it.

Jane: Why did you think at first you may need it?

Emily: I thought it might have had something to do with it.

Jane: Have you seen something like that in the picture?

Emily: Yeah at the shops, and to weigh fruit for how much money you need to pay.

It is interesting to note that when questioned on the purpose of scales in a real life scenario she identifies their role but yet is unable to use that understanding when faced with a similar situation in a testing condition.

However when the information contained within the graphic is represented as a word problem (Figure 4.20), Emily is able to effectively incorporate it.

![Figure 4.20 Emily—Modified scales.](image)

While the explanation of her reasoning in the Modified Test is a little confusing and hard to follow, the process she uses is a lot more appropriate than those recorded in the NAPLAN Item Test. The language is
uncomplicated, direct and structured, clearly giving her a place to start and work from as she calculates the answer of $15:

Emily: 25 is like, 50 is half a kilogram and so 50 would be $3 and then 50 + 50 is $6 and then ... (starts again). 4 apples would be $6 and then 50 would be $3 which makes that $12 and then plus another 50 makes that $15.

These findings again reinforce Emily’s preference for questions that are straightforward and not encased within a situation.

**Summary**

In both instances language had been used to modify the NAPLAN items; however it had been used in two different ways but with similar results. The first instance was the *decontextualising of the language*. This involved getting rid of any superfluous information contained within the question and resulted in Emily being able to identify the key elements and effectively utilise them.

The second instance was the *decontextualising of a situation through the use of language*. Due to Emily’s failure to recognise the role an information graphic played in the creation of a situation, the situation was removed by substituting the graphic with language. Once again the straightforward nature of the language used assisted in Emily being able to logically apply sound mathematical understanding.

Emily’s student profile highlights the important role language plays in mathematical assessment but also recognises some of the problems associated with the use of situations. These findings could bring Emily’s NAPLAN results into disrepute as not a true reflection of her mathematical understanding. In this instance she could have been incorrectly labelled as
struggling with certain mathematical concepts, namely data and mass. Yet the interview data revealed elements of the item that were impacting Emily’s capacity to access the item.

Also of concern are students whose results represent a sound understanding of mathematical concepts but yet the modification process reveals inappropriate strategies and perceptions. This has not only been evident in Sophie’s student profile but in previous students already discussed e.g., Lucy.

**Student Case 3 (Sophie): The composition of item design**

Sophie’s results improved due to changes made within the features of the item. This included the modification of the answer stem from short-answer to multiple choice as well as changes to the answer options within a multiple choice item.

**The Boxes Item**

As previously discussed, there is the notion of multiple choice items being regarded as easier than constructed-response items. This was due to a number of reasons, including the opportunity to be able to work backwards from the answers, a higher percentage of guessing the correct answer, and the ability to prompt the student through the provision of the correct answer contained within the options. Sophie’s results seem to support these conjectures when answering the box item in the NAPLAN Item Test (Figure 4.21).
In this item Sophie acknowledges the carton to be three boxes high and then added it four times giving her an answer of 12:

Sophie: I chose 12 because there are 3 and then I added the next three and the next three and I added down there.

Although her answer is starting to resemble the formula for measuring volume of LxBxH, her miscalculation of the height of the boxes results in a wrong response. However when the answer of 16 is provided in the multiple choice option (Figure 4.22) Sophie is a lot more accurate.

**Figure 4.21.** Sophie—Boxes.

**Figure 4.22.** Sophie—Modified boxes.
It is hard to ascertain what prompted Sophie to reconsider the height of the carton to be 4 boxes, but the multiple choice option obviously assisted. The formula she uses is more sophisticated than the one she used in the NAPLAN Item Test by acknowledging the need to multiply the two values to calculate the answer of 16:

Sophie: Cause there’s 4 down and then I just did 4 times 4.

Jane: So how did you know to times by 4?

Sophie: Because there’s 4 makes a square and there’s 4 up.

Similar to the findings of the larger cohort within the study, Sophie’s results expose a need to evaluate the use of constructed-response versus multiple choice options in assessment tasks. This includes the impact they have on the mathematical understanding that is supposedly being measured. This is particularly important with such a high prevalence of multiple choice items contained within mathematics assessment and therefore also makes analysing the answer options just as vital. This is highlighted in Sophie’s response to the following item.

**The Baby Mass Item**

As previously discussed in the modification chapter, Question 9 (Figure 4.23) was altered due to an inappropriate use of the item features to obtain the correct answer.
It was found that students were dismissing the weights that were measured in grams simply because grams are smaller than kilograms without actually looking at the number in front. Despite this inaccurate understanding of mass, students were getting the answer correct. However in Sophie’s case, her lack of mathematical knowledge was suitably represented with an incorrect response when she answers B:

Sophie: I chose Georgia cause it looks like a bigger number to all the rest.

Jane: Was there anything else that helped you work out your answer?

Sophie: No.

Jane: What does mass mean?

Sophie: The weight.

Sophie was distracted by the largest number representing the heaviest baby despite what measurement was placed behind it. Yet in the Modified Test (Figure 4.24) not only does she get the right answer but also exhibits sound mathematical understanding.
In Sophie’s account of the Modified Test she demonstrates a clear understanding of the conversion of grams to kilograms and therefore appropriately identifies the heaviest baby as Answer B:

Sophie: Cause it had the bigger number and Simon has 3.4kg which mean 3400g.

Jane: Why didn’t you think it could be Oscar?

Sophie: Cause he only had 3090g which is a smaller amount.

Jane: What about Mia?

Sophie: She only had 3005 and she was the smallest.

It is questionable as to what prompted or assisted Sophie’s newly applied knowledge yet it does reveal the ambiguity of NAPLAN results. With only a couple of weeks in between each test this is arguably not a long time for any classroom teaching to take effect. Sophie may have been having a hard day on the day of the test, forgotten to eat her breakfast, was running late to school, was anxious with the testing conditions, was distracted by a noise outside or a number of other things that could have all affected the way she originally answered the question. These are issues that cannot be ignored.
but rather must be acknowledged as real possibilities that could affect student performance.

Summary

Sophie’s data raises a major issue of multiple choice versus constructed-response, that is, who decides what answer format is going to be used and for what purpose? If the amount of constructed-response items is purely a result of a required percentage then is that a valid reason for their inclusion, and similarly for multiple choice? While there appears to be no public access to such information in regard to the NAPLAN, the justification for both forms of question needs to be established. These findings indicate that multiple choice questions are a more equitable and accessible style which needs to be considered when designing high-stakes assessment tasks.

In regard to the second item, the interview data alone makes it difficult to ascertain the exact reason behind Sophie’s change in response. However, in light of the literature highlighting the impact of individual students’ backgrounds on test performance, perhaps it is appropriate to consider the possibility of outside influences on Sophie’s test experience. It cannot be assumed that a child is capable of switching off the things that are troubling them and appropriately focus on the assessment item at hand. Nor can it be assumed that children would not feel the pressure or place unnecessary stress upon themselves during a high stakes assessment. These issues are ones that are out of the teacher’s control and must be acknowledged in national testing regimes.
Summary of the Case Studies

The case studies provided data which allowed trends and patterns by three of the interview participants to be developed and explored. In particular, the case studies provided data with respect to: (a) the impact of graphics on a child’s mathematical performance; (b) the responsibility of language on a child’s problem solving strategies; and (c) the need for close analysis of item features.

The following issues which relate to the methods or approaches used by the three students when attempting to solve the NAPLAN Item Test and the Modified Test were identified and discussed:

1. Not all students benefit from the use of information graphics within mathematics assessment.

2. The creation of a situation in an assessment item may jeopardise the quality of the language and graphic used.

3. There needs to be consistency between answer types for improved accessibility and equity for all students.

4. Students’ beliefs, values and personalities may affect performances on and approaches used in high-stakes assessments.

5. Correct answers do not necessarily reflect sound mathematical understanding.
Chapter 5: Model Development and Implications for Practice

This chapter presents the implications for theory and practice that emerged from the results of the study. As part of the analysis a model was developed as a framework of mathematics assessment practice and design. The chapter is divided into five sections:

1. Model development;

2. Implications of the context of assessment;

3. Examining the conditions created by mathematics assessment;

4. Implications of the item context for test designers; and

5. Summary of implications for teachers.

The results chapter revealed a constant interplay between several conditions that impacted on students’ results. In order to highlight these relationships a model was created not only to assist in attempting to present the research findings but to serve as a guide for future test development. It was also developed to provide a deeper understanding for teachers in regard to the many aspects of assessment and the best way these could be addressed in the classroom.

Model Development

The initial intent of the research study was to isolate the elements of item design in order to identify and evaluate their impact on children’s mathematical understanding. This in itself was an innovative measure since much of the research literature has focused on one or the other. However, the results indicate that to gain a deeper understanding, these elements
should not be viewed through such a narrow lens, but a conditional one. It is therefore not beneficial to analyse only the language of a test item without incorporating the graphic and the situation. It is also detrimental to analyse the use of a situation without valuing the unique attributes and background knowledge of the student. It was for this reason that a model was developed to show the relationship between not only the elements of item design but the assessment itself and subsequent conditions (see Figure 5.1).
Figure 5.1. A model of the contexts of mathematics assessment.
The model portrays the connection between the context of the assessment, the context of the conditions and the item context. Often the term context has been utilised in research to describe the real-life situation that encompasses the item design (see Boaler, 1993a; Verschaffel, De Corte, & Lasure, 1994). However this situational lens limits the understanding of the term ‘context’. By definition, context refers to the situation or circumstances that are relevant to an event. It is for this reason that it is often used to describe real-life situations located within items. However the findings suggest that there are other circumstances rather than just those within the item that impact on a child’s access to an assessment task. These include considering the context surrounding the assessment (the first circle, Figure 5.2) as well as the context of the conditions (the middle circle, Figure 5.3).

As the NAPLAN was the assessment tool on which this research study was based the model reflects the contexts it created. Therefore the first circle in the model includes elements of the NAPLAN that assisted in creating the assessment context. This included:
- pencil and paper format;
- high stakes;
- high graphic content;
- cost effective; and
- real-life scenarios.

Each of these elements impact directly on the conditions it creates and the items it develops. This relationship is indicated by the arrows moving towards the other two circles (Figure 5.4).

![Figure 5.4. The use of one-way arrows within the model.](image)

It is important to note that the arrows are not interchangeable or two-way. The arrows are one-way and point out the assessment context as the source in creating the other contexts. Subsequently the NAPLAN context creates conditions including:

- high pressure;
- accountability;
- individual student profiles;
• cognitive load;
• teaching to the test; and
• majority multiple choice.

Often this middle circle of conditional context is overlooked and not considered in regard to mathematics assessment with more emphasis being placed on the ‘how’ and the ‘what’ rather than the ‘who’. The ‘who’ represents the way individuals interact with the assessment context and the conditions this creates. These conditions are particularly relevant in regard to classroom practice as each form of assessment will uniquely create a range of circumstances that will directly impact on teachers and students to varying degrees. For example, high stakes testing creates greater accountability and pressure on teachers and students compared to more informal modes of testing in the classroom, and even then this will depend on individual characteristics of the participants involved. Similarly the benefits of the use of real-life scenarios are subject to a student’s social class, gender, ability and ethnicity, as well as a teacher’s willingness and ability to use them within the classroom setting. This results in the creation of an infinite amount of conditions subject to individual teachers and students as well as principals, schools and communities. It is for this reason that the conditions of an assessment must be considered in light of students’ results, especially as a means of measuring mathematical knowledge.

The final circle (Figure 5.5) represents the original premise and understanding of mathematics assessment as outlined in the literature review of this current study.
That is, that the item context consists of:

- graphics;
- language;
- situations; and
- design.

It was initially anticipated that these components of an item held the key to cracking the code of mathematics assessment. However, analysing the impact of each element individually has assisted in understanding the volatility of the item context. This unpredictability is the result of elements contained within the first two circles. To illustrate the interplay between all three circles and the need to consider more than the item context in mathematics assessment, the case study of Emily was applied to the model (see Figure 5.6).
In this instance the NAPLAN was the assessment governing the context and determining the nature of the task and therefore its conditions created Emily’s assessment context (Figure 5.7).
One of these features included the use of real-life scenarios and consequently the need for contextual language. This was translated directly in the item context through the use of situations however the middle circle also impacted upon its use in the item (Figure 5.8).

**Figure 5.8. Emily’s conditional context.**

Individual student profiles are important conditions that need to be considered in the item context. In this instance, Emily was a student who was more successful with items that were not written within a context than those that were. This resulted in an item context that was not compatible with her individual profile (Figure 5.9).

**Figure 5.9. Emily’s item context.**
Subsequently, Emily was unable to demonstrate her mathematical understanding due to limitations placed upon her through the use of the contextualised language.

Emily’s model exemplifies the interplay of the three circles in measuring Emily’s numeracy. These circles cannot be viewed in isolation, can be immeasurably different according to the assessment and participants, and have an accumulating effect. Each circle holds the key to unlocking the code of mathematics assessment and like all good codes may require many combinations to unlock it. These combinations will now be examined further in light of the three circles.

**Implications of the Context of Assessment**

The context of assessment refers to the circumstances surrounding its use and the way it is carried out. As there can be a number of purposes behind mathematics assessment so too can there be a variety of assessment contexts. Each circumstance would therefore impact differently on student’s performance by creating alternate conditional and item contexts. The analysis of the assessment framework provides a key to unlocking the code of mathematics assessment but it will not fit all the locks, only those with similar contexts.

Because the NAPLAN was used as the assessment tool for the research study its context will now be examined.

**The NAPLAN Context**

As previously discussed the NAPLAN was introduced in 2008 to enable nationwide comparisons and the first step towards Australia’s inaugural national curriculum (Lowrie & Diezmann, 2009). It was expected that:
The results from these national literacy and numeracy tests will provide an important measure of how Australian schools and students are performing in the areas of reading, writing, spelling and numeracy. The results from the assessment program will be used for individual student reporting to parents, school reporting to their communities, and aggregate reporting by States and Territories against national standards” (Curriculum Cooperation, n.d., cited in Lowrie & Diezmann, 2009, p. 144).

However it has been argued that “standardized testing is an extremely high-stakes practice in which children’s worth is ‘measured’ by a score that is not likely to be validly interpreted” (Haladyna, Haas & Allison, 1998, p. 266). Despite such concerns the Australian Government, along with an increasing number of countries, established mandatory testing across primary and secondary schooling. With such large-scale testing came the need for cost effective measures. This included the use of pencil-and-paper formats utilising a majority of multiple choice questions.

According to Threlfall, Pool, Homer and Swinnerton (2007), assessment of mathematics through the medium of paper and pencil is a “well-established part of an officially sanctioned formal procedure” (p. 335). They noted that “this is the assessment that counts: it is seen to embody many of the approved educational values in mathematics and is a strong determinant of how teachers set about their work of teaching mathematics in mainstream classrooms” (p. 335). Regrettably, Clements and Ellerton (1996) found that in regard to pencil-and-paper examinations “politicians, bureaucrats, and indeed people from all walks of life, seem to believe that a strong competitive examination is the cornerstone of economic progress and equality of educational opportunity” (p. 144).

However recently there have been moves particularly in the US and UK to explore the possibility of testing using computers. Hargreaves, Shorrock-

there is considerable interest in the possible development of online tests and examinations being shown by a number of bodies including government and the regulatory authorities. There is little doubt that within the next five years the Awarding Bodies will be expected to deliver one or more high stakes examinations online into schools and Further Education Colleges (CCEA, 2001, p. 1, cited in Hargreaves et al., 2004, p. 30).

Similarly, in Australia, discussions have begun on the possible introduction of a digital version of the NAPLAN.

Although there are arguments for and against both forms of assessment that require further research and development, it is the unique way each one will impact on the conditions of mathematics assessment that also needs to examined.

Let us imagine a NAPLAN item reproduced in a digital form. Will the graphic be identical? How will the item context be modified because the assessment context has changed? Even if the items look ‘identical’, it may not necessarily mean that the representation is the same. As Hargreaves et al. (2004) noted, “although children are relatively familiar with computers and very familiar with pencil-and-paper assessments, this does not automatically suggest that they can integrate these skills to complete assessments in mathematics using a computer” (p. 30). For example, graphics often take on a more ‘3D like form’ on a computer screen than they do in a pencil-and-paper form. Therefore it is likely that the graphic will change. It would also be tempting for test designers to add colour to bar and pie graphs and to use more elaborate pictures with context graphics,
however once again these additions would have a flow on effect through the creation of new conditions and subsequently new item contexts.

Other important issues within a NAPLAN-like\(^3\) context are the use of real-life scenarios and subsequently the style of items it includes. Although within this study these real-life situations have been analysed through an item context, it is the assessment context that could be regarded as the original source. In particular, the test construct that is being measured must be considered. In the NAPLAN context the testing bodies emphasise that it is not a test of content but rather to test 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


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;

\(^3\) Real life scenarios are used in other national and international high-stakes testing e.g. Primary School Leaving Examination (PSLE), Singapore; Trends in International Mathematics and Science Study (TIMSS); Educational Testing Service (ETS).
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).

This is just another example of how the assessment context impacts directly on the item context and therefore neither can be viewed in isolation. In this instance, attempting to measure a child’s numeracy has produced items that include ‘meaningful’ contexts. The results of the research study have shown that attempting to assess children’s numeracy impacts directly on the item context as well as the adverse conditions it creates. It seems that with so many variables interplaying in the assessment process of this test construct, numeracy is a difficult skill to assess with current assessment techniques.

The conditions that the NAPLAN context creates will now be examined.

**Examining the Conditions Created by Mathematics**

**Assessment**

According to Lowrie and Diezmann (2009), the use of high stakes testing has become increasingly influential in teaching practice and the day-to-day decisions teachers make. This has inadvertently created a conditional context that impacts on a child’s performance. These circumstances can vary according to individual students, teachers and schools but must be taken into account in the analysis of results.

One of the consequences of standardised assessment is the pressure it may place on teachers and students. Haladyna, Haas and Allison (1998) conclude that a number of organisations believe:
Testing increases pressure and stress on children, which sets them up for failure, lowered self-esteem and potential health risks;

Testing compels teachers to spend valuable time preparing children to take tests and teaching to the test, undermining what otherwise could be sound, responsive teaching and learning;

Testing limits children’s education possibilities, which results in a mediocre curriculum and learning;

Testing discourages social and intellectual development, such as cooperation, creativity and problem-solving skills, as time is spent instead on learning exactly what appears on the test; and

Testing leads to harmful tracking and labelling of children, especially those of minority and low socioeconomic backgrounds (p. 265).

The degree of influence these issues have on students and teachers may vary but can significantly affect performance.

Along with the pressure placed on teachers is the test-based accountability. As defined by Haladyna (2006), this is when “the test score becomes the only basis for assessing a student or a group of students in a classroom, school, school district, or state” (p. 32). Although these results are only a snapshot of student performance providing information that is relatively unsophisticated and generic (Lowrie & Diezmann, 2009) they are the very ones teachers are held accountable for. It is these conditions of high accountability, increased pressure and altered teaching practices that are necessary to consider when attempting to decode mathematics assessment.

Similarly, the use of multiple choice items with high graphical content presented in a pencil and paper format may affect performance in terms of demand on working memory and Sweller’s (1994) cognitive load theory. Although Threlfall et al. (2007) recognise that Sweller (1994) “acknowledges that element interactivity depends on the knowledge of the
individual, and is not a function of the task as such, it is reasonable to compare tasks across different mediums of presentation, looking at differences in the potential for successive and simultaneous treatment in the affordances of the context” (p. 336). They argue that while these differences may not apply to all individuals, it seems likely that it will for some individuals and therefore be of importance for assessment.

A common theme throughout the conditional context is the role individual students play in its creation. Although a significant amount of research focuses on the validity and reliability of assessment when measuring a child’s numeracy, the role of the social context is often forgotten. As noted by Zevenbergen and Lerman (2001), “in these new times, numeracy is a feature of the reforms and hence serious considerations are made of what it is to be numerate, but with little consideration of the social context within which judgements about levels of numeracy are made” (p. 571). However, the use of real-life scenarios within the assessment context must be analysed in light of student’s social class, gender, ability and ethnicity (Atweh, Bleicher, & Cooper, 1998).

One of the questions that needs to be raised according to Bernstein’s theoretical model is “the use of everyday contexts (as) serving as distractions from the main mathematical underpinnings of the task” (Zevenbergen & Lerman, 2001, p. 572). They noted that while it could be assumed that students will identify mathematical assumptions built into a ‘realistic’ task, they may inappropriately treat them within an everyday context ignoring the mathematics task altogether. Similarly, Cooper and Dunne (1998) argued that middle class students perform better than their working-class peers on realistic tasks due to their ability to identify the
mathematics processes involved. Zevenbergen and Lerman (2001) identified this phenomenon as Bernstein’s (1996) recognition rule. They noted that this rule allowed individuals to “recognise the specialty of the context they are in” (Bernstein, 1996, p. 31, cited in Zevenbergen & Lerman, 2001, p. 573). This rule distinguishes that students will respond differently to the use of real-life situations in the assessment context, creating unique conditional contexts. “When students fail to identify this rule...that are unable to respond appropriately” (Zevenbergen, 2000b, p.13).

An example of the way the conditions of the assessment impacted on a child’s ability to solve a task can be seen within Celine’s response outlined in the results chapter of this study. Celine struggled to comprehend the Pie graph question due to its length, as noted in her transcript where she incorrectly answered B:

Celine: Well it says “meals Hannah eats on this day”, it doesn’t actually say “how many meals”. And for the hours watching TV on Tuesday, this is Monday. I was a little bit stuck on awake that day. I thought that one (Answer B) was a little more direct.

Once again, the context of the assessment—the NAPLAN—was the original source in dictating the conditions it created. This again highlights the one-way arrow originating at the assessment context within the model. In this instance the use of realistic tasks produced items that included contextualised language and subsequently extraneous and detailed information. However it appeared that the length of the question required too much working memory for Celine to be able to retain the original understanding resulting in a confused and incorrect response. This exemplifies the way cognitive load may impact on a child’s ability to solve a task where the item ended up not necessarily testing Celine’s ability to
read a pie graph but rather how effective she was at retaining and utilising all the information contained within the question.

This brings us to the final circle contained within the model, the item context. As much of the item context has already been analysed independently in the results chapter, and the implications of their use in assessment evaluated through the students’ results and interviews, it also needs to be considered in light of the interplay between the three circles within the model. These findings are particularly beneficial in regard to test item design.

**Implications of the Item Context on Test Design**

The research findings revealed several issues significant to test designers. While these concerns were raised as a result of utilising the NAPLAN instrument, their application and relevance is universal to all mathematics assessment. Once again the aim of the research was not to scrutinise the integrity of the NAPLAN but rather the impact modifications on assessment items have on student performance. The NAPLAN was viewed as a valid representative of assessment tasks. In fact, according to Connolly (2011), the NAPLAN Numeracy assessment guidelines are a very close match to the Trends in International Mathematics and Science Study (TIMSS) assessment framework, an international test undertaken in most developed nations.

These NAPLAN item writing guidelines as noted by Connolly (2011) were produced based on “published best practice from a number of authoritative sources” (p. 914). These included:
Thomas Haladyna’s (1999) 30-point checklist for multiple-choice writing rules.


The Educational Testing Service (ETS) International Principles for Fairness Review: Guidelines (2007). According to Connolly (2011), “this major testing agency in the USA publishes a set of principles for fairness designed to avoid bias at a content level and at item review” (p. 915).

These guidelines were established to ensure validity of test-score interpretation. That is, students’ performance is a reflection of their mathematical knowledge.

Yet despite all these good measures put in place, Haladyna (2006) argued that perfection in test development and validation is unattainable. He recognised several perils of standardised achievement testing similar to those identified in the current study including student individuality, varied teacher instruction, test preparation, cheating, test development, test administration, test scoring and standard setting. Once again this reinforces the need to evaluate all three circles of context involved in mathematics assessment as they are so closely linked to one another and impact on student’s results.

Although these threats to validity exist, and are certainly reinforced through the results of the current study, they cannot be dismissed as by-products of assessment particularly when “test scores affect students’ lives and their
teachers’ careers” (Haladyna, 2006). It is for this reason that research is vital to inform test-item design and improve validity of mathematics assessment. As Haladyna (2006) noted “If the messenger of student learning is so badly flawed, where is the truth in the message?”

Using this analogy, the current study revealed several flaws in the NAPLAN ‘messenger’ which were exposed following analysis of the ‘messages’. These will now be examined further.

**The use of graphics in the item context**

The research findings reinforced the need for analysis and investigation into the inclusion of graphics in mathematics assessment. While there has been some research into the use of graphics in assessment (see Diezmann, 2008), little has focused on the impact they have on children’s performance (Lowrie & Diezmann, 2009).

With the rapidly increasing use of graphics in what could be considered a digital age, society has necessarily become more reliant on representing information in graphical and diagrammatical forms (Lowrie & Diezmann, 2009). It could be seen from the research findings that children too have developed a confidence in the use of information graphics with a decrease in test performance when an information graphic was removed and an increase when one was added. Consequently when an information graphic was utilised the students were able to effectively access the data it contained. At the same time, it highlighted the need for the inclusion of well-constructed graphics. Moreover, as noted in the previous chapter, the interplay between the graphic and the other item context features is critical (Lowrie, Diezmann & Logan, 2012).
According to Mackinlay (1999), the two criteria of good graphical design are codified as expressiveness and effectiveness. He maintained that:

Expressiveness criteria determine whether a graphical language can express the desired information. Effectiveness criteria determine whether a graphical language exploits the capabilities of the output medium and the human visual system (p. 66).

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. It may be the case that contextual graphics have no role to play in high-stakes testing since such graphics are by definition contextual and therefore not necessary to the task at hand.

The notion of ‘expressiveness’ is similar to Kosslyn’s (2006) Principle of Relevance. According to Lowrie, Diezmann and Logan (2012) this principle is based on the notion that “graphs can become ineffective if too much information or too little information is presented” (p. 171). Subsequently the first step in preparing a graphic is to determine exactly what information needs to be conveyed (Kosslyn, 2006).

In regard to ‘effectiveness’, Shah and Hoeffner (2002) identified “a viewer’s knowledge about graphs, and a viewer’s knowledge and expectations about the content of the data in the graph” (p. 47) as two major factors that influence viewer’s interpretation. As previously noted in the literature, the successful comprehension of a graphic is not only dependant on the appropriateness of the graphic but the ability and knowledge of the reader (Lowe & Promono, 2006). It is for this reason that there is a need for a close alignment between graphics design and the curriculum to ensure
students are not disadvantaged and results are a true indication of performance.

Although the inclusion of information graphics positively impacted on student performance in the current study, the use of contextual graphics were detrimental to student results. With the increase of both types of graphics within mathematics assessment, “knowing when to and how to extract information embedded in graphics can be problematic” (Lowrie & Diezmann, 2009, p. 11). In light of the findings it appears that students struggle with this exact dilemma.

Even though contextual or auxiliary representations (Lowrie, Diezmann, & Logan, 2011) are not essential to the solution as they do not contain any information, they are often added as cues to the context. It has also been discussed within the literature the possibility of their inclusion as a form of motivation and elaboration (Shimada & Kitajima, 2006). However it was found that students were inappropriately attempting to include these in their problem solving strategies. For this reason test designers need to re-evaluate their good intentions of including such a graphic and the necessity of its inclusion. In other words, it is difficult enough to develop appropriate information graphics, let alone graphics that have no role to play.

**The use of design in the item context**

The use of design in the item context is crucial to ensuring valid and reliable results. The design used must reflect the competencies that the test is intended to measure and therefore needs to be of high quality. One aspect of design that has been controversial both in research and in the current
study is the use of multiple choice questions in comparison with constructed-response formats.

According to the NMAP (2008), “many educators believe that constructed-response items (e.g., short answers) are superior to multiple choice items in measuring mathematical competencies and that they represent a more authentic measure of mathematical skill” (p. 60). However they found that when examining the literature on the psychometric properties of both forms of questions, the scientific literature does not support the assumption that each type measures different aspects of mathematics competency. These results are similar to those of Hancock’s (1994) who concluded that “for assessing objectives whose test items may take either format, multiple-choice and constructed-response formats appear comparable at sub-synthesis cognitive levels” (p. 150). Rodriguez (2003) also questioned the use of the more expensive form of constructed-response items when the multiple choice format effectively measured the same construct.

Despite these findings, Osterlind (2002) argued that “there is no developed theory to undergird item writing, nor is there a comprehensive resource identifying the distinctive features and limitations of test items, the function of test items in measurement, or even basic editorial principles and stylistic guidelines” (p. 5). Although he acknowledged the list of ‘dos’ and ‘don’ts’ that have been offered, he argued that “a simple list neither captures the complexity of the task nor conveys why certain features are requisite to producing test items of merit” (p. 5).

Despite these findings the decision of the construction and type of items used within assessment lies in the hands of influential teachers, education administrators, psychometricians, bureaucrats and politicians, all of whom
advocate the continued use of such tests on both educational and pragmatic
grounds. Therefore it is not what the test designers know and what research
findings suggest that govern the development of test item design, but rather
the overarching influence of many contributors including political agendas.
This means that despite the best intentions of test designers to make an
assessment valid and reliable more research needs to be undertaken and
acknowledged in the assessment writing process.

Another instance of test design is considering the appearance and
arrangement of the item on the page as recognised in the current research
study. According to Osterlind (2002) “an attractive appearance will
facilitate communication with examinees who may otherwise be distracted
by sloppy page layout, difficult to read type or poor quality type” (p. 193).
This page layout includes the placements of graphics, the bolding of words,
the layout of answer options and the placement of the question stem. This
also highlights the common misconception that pictures i.e., contextual
graphics, are necessary in order to keep the reader ‘focused’. However as
the findings suggest it has resulted in an inappropriate focus upon the
graphic and the inability of the child to differentiate between important
information and those only included to make an item look pretty.

Similar to the rules governing illustrative text as outlined by Anstey and
Bull (2000), the order, grouping and arrangement of words and their
presentation using a variety of fonts, sizes and layouts on the page can
provide diverse degrees of meaning. This also includes the many ways the
elements of the graphic are brought together and can generate different
meanings. It is therefore vital that the purpose behind the assessment item is
reflected in the layout and design of the question. This concept requires further research and investigation.

The use of language in the item context

The old adage ‘keep it simple stupid’ seems highly applicable when it comes to the language used within assessment items. As noted by research within the literature, the role of reading and the use of language are particularly problematic when they begin to interfere and inadvertently measure the reading ability of a child rather than the intended construct (see, for example, Abedi, Lord & Plummer, 1995; Abedi, Lord, & Hofstetter, 1998; DeCorte, Verschaffel, & DeWin, 1985). In order to make items fair and accessible to all students, test designers must ensure that the concepts being tested are the ones intended and not the literacy performance of a child. According to Clements (2004):

...at the upper primary level most errors on mathematics tests and examinations are caused by Reading, Comprehension or Transformation errors, or by Carelessness. Often, pupils are able to carry out one or more of the four operations (+, -, x, ÷) needed to answer a question, but they do not know which operations to use (Clements, 2004, p. ii).

Although these findings are substantiated within the current study, they should not be used as grounds for the exclusion of language from mathematics assessment altogether. It rather indicates that the language used should be carefully analysed according to its necessity and in particular its readability. Wiest (2003) defined readability as “all factors related to reading and comprehending written text” (p. 1) and can include such things as the format of the text, the structure of the language and the vocabulary used (Shorrocks-Taylor & Hargreaves, 1999).
Wiest (2003) claimed that one key consideration in standard readability formulas is word and sentence length. Mobely (1986, cited in Shorrock-Taylor & Hargreaves, 1999) similarly noted that strange and difficult sentence structure can cause the reader to lose “the thread of meaning” (p. 128). This was found to be particularly problematic on student performance in the research study.

Yet the readability of an item is as much a product of the characteristics of readers as of texts. Harrison (1980, p. 33, cited in Shorrock-Taylor & Hargreaves, 1999) noted:

> Readability is an attribute of text; comprehension is an attribute of readers. There is therefore a fundamental difference between the two concepts. Having made that distinction though, a moment’s reflection makes it clear that the concepts are intimately related, in that very often when we use the term readability we mean in effect the comprehensibility of a text (p. 128).

This concept supports the use of the three circles within the model and the need for test designers to acknowledge the relationship between them. For it is redundant to measure the effectiveness of the language used within an item without recognising the role individuals’ play in the process. This was evident in the results of the current study where students’ responses to the same question varied significantly based on who was reading it, not necessarily what they were reading. This also aligns with Bourdieu’s theoretical position where:

> words are never just words, language is never just a vehicle to express ideas. Rather is comes as the product and process of social activity which is differentiating and differentiated; and thus, differentially valued within fields of social activity. Language is value-laden and culturally expressive according to standards of legitimacy and opposition to them (Bourdieu, cited in Greenfell, 1998, p. 73).
Therefore test designers must ensure that the language used in a test item is not biased or favours a particular group of students. This is particularly relevant when used to link to the real world of students (Zevenbergen, 2000a).

**The use of situations in the item context**

A lot of the implications regarding the use of situations within test design are similar to those already outlined within design, graphics and language. This is because of the close relationship the situation plays in all aspects of the item context. In essence the type of situation utilised dictates the means by which it is expressed and interpreted. This is particularly important when test designers are attempting to incorporate a situation within an item.

This can be best illustrated by using an item from the 2009 Year 5 Mathematics NAPLAN (see Figure 5.10).

![Figure 5.10. A train timetable.](image-url)

This item is embedded in the situation of a train timetable. It could be assumed that this is a ‘real-life’ situation that students would be familiar with. However, as Zevenbergen (2000a) argued, “the implicit assumptions embedded in these types of questions needs to be questioned in relation to
the real lives of the students” (p. 219). For students living in remote areas the mechanics of a train timetable may be quite foreign to them compared to the understandings of their peers in inner city areas. Consequently this “creates greater cognitive demands for some students than others, thereby restricting their access to the tasks and subsequent resolution” (Zevenbergen, 2000a, p. 219).

This example once again emphasises the need for careful analysis of the use of these ‘real-life’ situations and whose life and culture we are actually valuing and whose we are choosing to ignore. It acknowledges that situations used in standardised assessment need to be authentic. Niss (1992, cited in Palm, 2009) defined an authentic extra-mathematical situation as “one which is embedded in a true existing practice or subject area outside mathematics, and which deals with objects, phenomena, issues, or problems that are genuine to that area and are recognised as such by people working in it” (p. 4). Due to the large scope of the NAPLAN to include students from varying backgrounds it would seem an incredible feat to consist of truly authentic situations. It is for this reason that Palm (2009) acknowledged that there is little empirical evidence for the effects of the use of more ‘authentic’ word problems.

Another reason is the way ‘authentic’ has been vaguely described throughout research and the subsequent lack of frameworks to guide research and synthesise research results. It is for this reason that Palm (2009) developed a framework for authentic tasks concerned with the concordance between word problems and real-world task situations. Within this framework Palm (2009) identified several key elements that must be as closely aligned to a real-world situation as possible for the item to be
considered authentic. These include the event, the kind of question, the realism of the information, the presentation mode, the plausibility of the solution strategies, the circumstances in which it is presented as well as its purpose. This theoretical framework could be utilised by test designers to analyse the use of situations within the item context to establish its attributes as an ‘authentic’ task. If an item is actually presented in an authentic manner, interpretation of what the task demands will not be dependant of social class, ethnicity, geographic location or gender. In an idealist world these would be the items included in mathematics assessment however at present seem to be absent in current assessment practices.

Implications for Teachers

One of the initial research questions of this study was what effects will these findings have on classroom teaching? This is particularly pertinent when the assessment being used to evaluate students is the same one used as a measure of schools and teachers’ performance. However it is not the intent of the research to provide teachers with a set of strict guidelines and criteria in order to obtain the best results from their students. This practice, often referred to pejoratively as ‘teaching to the test’, does not benefit the student or the school community as a whole. In fact, Haladyna (2006) refers to this practice as ‘consumer fraud’, in which students, parents and the community are deceived about how much learning has actually occurred. Alternatively, Haladyna (2006) urges teachers to be ethical in their test preparation. He noted that “the most ethical test preparation is good teaching: using the content standards, aligning instruction to these standards, assessing learning, re-teaching, and re-assessing” (Haladyna, 2006, p. 37). It is for this reason
that the research findings will be analysed according to classroom teaching rather than specific test preparation techniques.

**Graphics in the Classroom**

The most noteworthy aspect derived from the findings is the need for teachers to explicitly teach the role of graphics in mathematics. Although the study indicated that children are becoming familiar with the use of graphics there was an apparent confusion as to their purpose within questions. It appears that teachers are incorporating and emphasising the use of information graphics but in the same instance failing to explicitly teach the role or ‘non-role’ of contextual graphics. An example of this was the use of the shoe picture where children inappropriately tried to incorporate it into their mathematical reasoning. As with many aspects of mathematics we cannot simply assume that children will naturally assimilate such understanding without explicit instruction. Similarly, Lowrie and Diezmann (2005) maintained that the explicit teaching of such practices has to occur in order for students to effectively decode graphical information.

From a classroom perspective, teachers need to create situations where context is embedded into classroom practices and increasingly these contexts require the representation of graphics. Thus, teachers are challenged to represent mathematical understandings as they are being represented in the outside world. However, there are numerous graphic conventions that need to be addressed in any given graphics representation. More challenging is the fact that many graphics are inappropriately represented, so firstly teachers need to ensure they have a sound understanding of good graphic conventions before utilising graphics in classroom situations. Without this sound content knowledge teachers will
only add to the problem rather than confronting it with the development of sound structure and function. Even the seemingly simple construction of a graphic from an Excel spreadsheet needs to be comprehensively taught, otherwise students will inadvertently reinforce the use of poor graphics (e.g., using a bar chart to represent time and distance rather than a line graph).

Another strategy which may be effective is to encourage students to source graphics from advertising agencies that represent both simple and complex data in very powerful (and appropriate) ways. Students can then describe the features of the graphic and comment on the ease in which data can be interpreted or summarised. Such pedagogical approaches must be advocated as graphics become increasingly a part of our lives.

**Language in the Classroom**

The research findings highlighted the fact that students found it challenging to process both graphic information and literacy demands simultaneously. As Mayer (1989) has highlighted, students are more likely to understand graphics when they are presented with texts, however he pointed out that the text had to add value to the graphic and vice versa. As a teaching implication it is imperative that teachers do not present graphical information without appropriate text. It is also necessary to demonstrate to students the connection between the literacy and the graphic. Such practices reduce the cognitive demands on students especially when faced with novel tasks or tasks that have substantial mathematics demands. Therefore it is not just a graphic that can support students’ understanding of a task, but if represented appropriately, the relationship between the language and the graphic as well.
As mentioned above, one place to start is by exemplifying best practices. If there is a disconnect between the graphic and the literacy students will often, as indicated in the research, lose their train of thought and struggle with connecting the various elements of the task. It should not be the case that tasks are designed in order to confuse students rather they should be arranged in ways that allow students to utilise the information to demonstrate what they know. The Pie graph (Question 2) was a case in point. In this task students became confused with the unusually lengthy framing of the language and therefore were unable to apply it appropriately to the graph. Typically, language associated with graphs is precise and directly associated with graphical information rather than the increasing descriptive demands that were required to contextualise the information.

**Situations in the Classroom**

It is commonly accepted that assessment should reflect both classroom practices and societies needs and demands. As a result, assessment is usually viewed as a process which reflects what happens in the classroom and readies students for life outside of school. In terms of mathematics research the enormously influential *Realistic Mathematics* movement (originated from the Freudenthal Institute in The Netherlands) has demonstrated the benefits of contextualising mathematics within a student’s knowledge base and general societal understandings. Such approaches have worked very well in a number of mathematics situations and especially situations where real-world knowledge is beneficial (e.g., dealing with money and developing measurement concepts). However, when the intended purpose of assessment design is to embed situations into a given design task, the quality and reliability of the assessment suffers (Boaler,
1994). Sullivan, Mousley and Zevenbergen’s (2003) argument regarding the use of context in the classroom best describes the cautious approach that needs to be taken:

We are not arguing that contexts should not be used: indeed we believe that contexts have much to offer. The issue for us is that teachers need to be fully aware of the purpose and implications of using a particular context at a given time, to choose a context that is relevant to both the problem content and the children’s experience, and to have strategies for making the use of the context clear and explicit to the students (p.111).

The present study has demonstrated the reduction in strong assessment principles when too much contextual information is described within the task scenario. After all, one person’s realistic mathematics is not another’s rich context. In fact, the desire to contextualise tasks may well disadvantage as many students as it advantages. The situation does not just simply disrupt the context of any realistic intent; it can also provide an unintentional mismatch between the elements of the task. For example, in the calendar task many students were inappropriately applying their understanding of the days within a month to calculate their answers rather than monitoring and patterning the sequencing of events that would happen from this scenario. With the graphic added the situation not only became more realistic, it actually made clear what was required in the task.

The research has also highlighted the need for teachers to acknowledge the diverse range of learners within their classroom and the unique perspectives and understandings they bring to a task. The use of situations must provide avenues for all students to exemplify their mathematical understanding rather than becoming obstacles for them. As Morgan (2003) noted:

Traditional explanations of students’ failure to complete assessment tasks satisfactorily have tended to
focus on the characteristics of the individual student. He or she lacks mathematical ability, has not studied hard enough, reacts badly to the pressure of examination, and so on. However, it is possible to adopt an alternative perspective that sees success in school mathematics as participation in a particular form of discourse rather than as acquisition of a body of knowledge. Such a perspective allows us to see failure as a consequence of students’ use of the concepts, values and forms of expression of alternative discourses, different from those of the ‘target’ discourse (p. 46).

It needs to be acknowledged by teachers that the use of situations is not simply a ‘one size fits all’ circumstance. The mathematics classroom should be a place where different values, understandings and cultures are not only acknowledged but embraced. The provision of such a learning environment not only results in a more useful application of mathematics but can better prepare students for assessment tasks that are embedded within these situations. It will also allow students the opportunity to recognise the need for mathematics within their daily lives if they can see a direct correlation to it.

**Item Features in the Classroom**

While throughout this thesis there has been no suggestion of any form of teaching to the test it may be something that cannot be avoided in the attempt to expose students to the multiple-choice format. In this instance it is only through practice and intentional exposure that students will develop a confidence and familiarity with such a design. As both Kosslyn (1985) and Mayer (1989) have indicated, even the positioning and placement of features of the item can influence the manner in which a given task is solved. Although this study has highlighted flaws in the multiple-choice design, it is recognised that this type of assessment format must remain for
practical and cost-effective reasons. There has been an extensive research base which interrogates the ‘worthiness’ of this type of format and a number of agencies (e.g., the Educational Testing Service) possess a comprehensive database of ‘effective’ and reliable items. This type of format is not going to go away. However this study has highlighted the role of distractors in students’ processing and particularly the interplay between the question and answer format. It was found that capable students often use the answer options as a scaffold to construct the most appropriate solution (correct solution). However in a similar way, students inappropriately utilised the multiple choice options to reinforce wrong answers. Thus, it is again important that teachers explicitly share test strategy processes with students as a way of more equitably using the specific design features as a productive assessment tool.

One way of empowering students is to encourage individuals to actually design tasks for others to solve. By taking charge of the design process students get to see the relationship between item components, and especially the one correct answer and other distractors presented in any multiple choice format. Students can then swap design tasks with one another for others to solve. In a sense this is similar to those mathematics researchers who advocate problem posing (Silver, 1994).

**Summary of the Model and Implications**

The purpose of this chapter was to analyse the results of the study in regards to theory and practice. This was achieved through the development of a model that clearly identified the links between the various contexts involved in mathematics assessment. These links highlighted the interconnectivity of
the assessment context, the conditional context and the item context when measuring a child’s numeracy. Each of these contexts was explored in light of the research findings, focusing specifically on the implications for test designers and teaching practice. The following chapter will summarise this discussion according to further research opportunities, acknowledging limitations of the study.
Chapter 6: Conclusions

This chapter summarises findings and implications of the research questions presented in Chapter 3 and the data analysed in Chapters 4 and 5. It also explores the implications for mathematics teaching of the model developed in Chapter 5. Limitations for the study are also discussed and a summary of the implications for theory and practice described. Avenues for further research are then explored.

Findings in Relation to the Research Questions

This study investigated the manner in which the design and representation of mathematics assessment tasks influenced students’ reasoning. In particular, the investigation described the influence of specific components of items on students’ mathematical reasoning, and identified the degree to which these features produced particular assessment outcomes. As a result, the study questioned the entire design process of high-stakes testing instruments and problematised the extent to which slight variations in particular design features impacted on student performance. Although details of the research findings and implications have been explored in Chapter 5, the overall findings of the study are briefly summarised under the research questions in the following sections.

The first question to be explored is:

From a selected number of often used mathematics tasks, which do students find most difficult to solve?

The students found 5 of the 15 items much more challenging than the other items presented in the first phase of the study in the NAPLAN Item Test.
These 5 items produced means of less than 50% correct and were approximately 20% more difficult than the sixth most difficult item. Each of these five difficult items required the students to either analyse and interpret a graphic (Paper fold, Garden plan, Scales and Fruit juice) or visualise a graphic (Meeting schedule). With respect to analysing the graphic, quite distinct types of processing were required and this demonstrated the importance of decoding and encoding in problem solving. The Paper fold, Garden plan, Fruit juice and Scales items all contained an information graphic and this graphic needed to be interpreted correctly in order to solve the task. The Scales and the Fruit juice tasks required students to access measurement data (specifically the mass and capacity of objects). For both of these types of items the students needed to recognise the need to incorporate the given information contained within the graphic. The Paper fold and the Garden plan required the students to not only decode information embedded within the graphic but also add to this information by either establishing the rotation of an image (in the case of the Paper fold item) or inserting measurement quantities into the graphic (in the case of the garden task). The other challenging task (Meeting schedule) required the students to encode a graphic (i.e., imagine a calendar representation in their mind’s eye).

In summary, the students found it most challenging to engage with the decoding or encoding of graphics. By contrast the 4 easiest items (with means at 92% or greater) required either the simple recognition of graphic representations (in the case of Liquorice allsort, Coins and Spinner) or the location of a coordinate point (in the case of Street map). Unlike the difficult items, the students did not have to apply the information from the
graphic to other mathematics or literacy demands. They simply needed to identify an object as a one-step process. Thus, the graphics did not require any utility or engagement. This study demonstrated that of the items used as representative of today’s high-stakes assessment tasks, students found items that required either encoding or decoding of graphics most challenging.

The next research question was:

_How does the use of the respective item components (graphics, language, situations and item construction) impact on a child’s capacity to make sense of mathematics?_

In order to answer this research question a range of descriptive and univariate analyses were employed. Initial observations were sought to identify change patterns in student behavior based on an experimental design that isolated specific components of mathematics tasks. Relatively minor variations accompanied the modified design task and yet substantial changes took place. A series of ANOVAs were used to determine difference between student performance on the standard NAPLAN Item Test and Modified Test items. Six of the 10 items revealed statistically significant differences between performance across the two sets of respective items. For 5 of the 6 items, the modified task produced higher means, thus indicating that student performance increased based on design changes rather than content changes. For 2 of these items the removal of unnecessary contextual information improved student performance. For the Shoe item, student performance increased dramatically when a contextual graphic (i.e., the shoe) was removed. For the Pie graph item the unnecessary literacy demands were taken away.
Two other graphic alterations resulted in significant change in performance. In both instances there was a change from decoding to encoding (the removal of a graphic which represented a fraction) or encoding to decoding (the addition of an information graphic to represent a calendar). The first had a negative effect on student performance while the second produced a positive increase in performance. In both instances the students had more difficulty in encoding than they did decoding the information. These results are noteworthy since the first research question highlighted students’ inability to engage with and manipulate decoded information. This research question identified the fact that encoding the information was even more challenging and thus a hierarchy of difficulty was established.

The remaining two tasks that revealed significant variations in performance were associated with the repositioning of information within the respective tasks. For the garden problem, the graphic was simply repositioned, while for the Boxes task the item was framed within a multiple choice format. These results again highlight the critical importance of item composition (as described by Kosslyn, 1985 and Mayer, 1989).

This study has demonstrated through the use of the modification paradigm and a mixed methods approach, that a child’s capacity to make sense of mathematics is heavily influenced by the use of the graphics, language, situations and item features included in the task.

The third research question was:

*To what extent does assessment reflect sense making and mathematical understanding?*
This research question was explored through a qualitative analysis of students’ reasoning and reflections after they had solved both sets of tasks from the NAPLAN Item Test and Modified Test. These data revealed that students’ conceptual understandings were sophisticated and appropriate even in situations when they incorrectly solved the tasks. That is, the students clearly possessed the appropriate knowledge and tools to solve the tasks and yet they made errors in processing the tasks largely due to the nature of the task design. Of equal concern was the fact that students were successfully completing the tasks without adequate knowledge of the concepts being assessed. Slight modifications in item construction evoked varying levels of sense making that were particularly influenced by modifications to the graphic or the positioning of the item features within the tasks. Depending on the appropriateness of a graphic, the graphic either became a tool that helped scaffold information (when added) or forced the students to create their own image-based representation when the graphic was removed. The addition of a graphic allowed the students to decode relative information while the removal of a graphic required the students to represent (usually in their mind’s eye) an image to support their problem-solving behaviours.

Consequently a high degree of the students’ sense making was determined by the composition of the assessment task rather than the mathematics content fundamentally associated with the assessment purpose. In other words, the modifications were not made to the assessment’s purpose but only to the design. Therefore any changes in student performance were a reflection of the design modifications.
This has demonstrated how current high-stakes testing cannot be regarded as an accurate reflection of students’ sense making and understanding due to the impact of design modifications.

The final research question was:

*What effects will these findings have on classroom teaching?*

This question was addressed through the development of a practical applications model that demonstrated the relationship between not only the elements of item design, but the assessment itself, and subsequent conditions that the assessment creates. The model highlights the critical importance of item construction and also other contributing factors necessary when considering the assessment of a child’s numeracy. These are particularly relevant for classroom teachers as the findings of this study indicated a mismatch between classroom content and assessment expectations.

If teachers are aware of current assessment design they can make use of incorporating them into their classroom. Although certainly not encouraging ‘teaching to the test’ there is a need for teachers to make use of these findings in developing sound mathematical teaching strategies so students are equipped with the relevant skills necessary to achieve in high-stakes tests. These include a detailed knowledge of the use of graphics, a familiarity with the use of language and exposure to a number of different situations.

Most importantly, it will assist in bridging the gap that this research study has revealed between the impact of assessment design on a student’s ability to solve a task. Consequently, assessment results will be more closely
aligned to a student’s mathematical understanding. Unless classroom teachers possess a sound understanding of how items are constructed, and the interplay of the various components of an item, there is little chance of providing students adequate support in order to develop a ‘test awareness and readiness’.

This study also highlighted the value added to assessment through the use of student voice. It is important that teachers spend an equal amount of time analysing children’s external representations as well as their internal mathematical reasoning. Within this study, allowing students the opportunity to verbalise their mathematical thinking provided real insight into misconceptions and gaps in their understanding. Typically, that would be the role of the teacher.

This research study has identified a number of issues relevant to teaching practices today.

**Limitations of the Study**

As previously stated, there has been little research reported on the modification and use of all four elements of an assessment task. Therefore the modification process utilised within the study was partially related to the Test Accessibility and Modification Inventory (TAMI) developed by Beddow, Kettler and Elliott (2008). Although the TAMI was developed as a means of making assessment items more accessible for students with disabilities, it has proved to be an invaluable source in identifying the problematic nature of assessment items and providing guidelines to the modification process. The modifications executed in this study had a similar intention to those implemented by Beddow, Kettler and Elliot (2009) in an
attempt to “afford an entire group of students access and better measurement of their achieved knowledge” (p. 531). Although not all the modifications implemented in this study resulted in statistically significant results between the two tests for the entire cohort, there were individual stories that demonstrated improved access and change of mathematical thinking.

However, as the research literature surrounding the modification paradigm was scant, it was difficult to theoretically support the modifications utilised in the research apart from the use of the qualitative data that was collected. Consequently, modifications to items were driven by an analysis of data rather than from theoretical underpinnings. Although essentially it was the purpose of such data to assist in revealing trends and misunderstandings in children’s thinking, it did mean that a majority of the modifications were instigated from empirical data and the subsequent analyses of that data. This resulted in modifications that became new prototypes rather than from a comprehensive theoretical position. Thus, other changes may have also been appropriate if such a theoretical position had been established elsewhere and therefore the modifications could be considered reactionary. Indeed, this limitation meant that not all 40 NAPLAN questions could be investigated and modified since it would have taken so much time to run repeated pilot trials for all 40 items.

The modification process also proved problematic due to the necessary urgency in developing the modified items. In an attempt to reduce any effects of teaching on students’ results, the initial qualitative and quantitative data had to be quickly analysed to create the Modified Test as soon as possible. The main disadvantage of such a process was that of hindsight. It is easier to identify more successful ways that items could have
been more effectively modified now at the completion of the study, especially with the use of the model that was developed. For example, the Shoe item may have been more valuable without the reference to looking at their own shoe but simply worded “What is the length of a shoe?” Similarly, a more complex experiment could have been created if the standard NAPLAN question had been compared to two items with the same feature modified in different ways. Once again, for example, the Shoe item could have had the picture removed altogether or replaced with a smaller picture to ascertain whether the issue was the inappropriateness of the graphic or the inappropriateness of the size of the graphic. Thus, the modification process is a limitation that should be noted when considering the particular findings of this study especially in regard to transferability.

**Avenues for Further Research**

There are a number of research issues that have emerged from the study that warrant further exploration. The research reported in this study 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. The two main aspects of the study that need to be focused on in more detail include the extent to which: (1) modifications in item structure influence both student performance and their sense making; and (2) decoding and encoding of information fluctuate with respect to how mathematical ideas are represented and presented in mathematics tasks.

**Item Modification**

Although studies have investigated performance differences of students across standard and modified assessment items (e.g., the Educational
Testing Service) to the best of my knowledge no studies have examined
student sense making in relation to item modification. This study has
examined the influence of minor alterations in task design in relation to
student’s understanding of an assessment task. Given the dramatic changes
in students’ understanding, it is imperative that studies with larger banks of
items take place. This is now possible given the model developed in this
thesis.

Thus, more research is required in relation to the influence of graphics and
the positioning of graphics in assessment tasks. Specifically research needs
to be undertaken into the merit of having context graphics in tasks since it
could be argued that they are not necessary and in fact become distracting to
both students and the intent of the actual mathematics task. With respect to
information graphics, a stronger experimental design is required to
determine where best the graphic should be situated in a task and how the
written information that surrounds the graphic best conveys meaning.
Furthermore, additional research needs to be undertaken into how graphic
displays impact on decoding ability. For example, when requiring
information about the mass of an orange, does it matter if that text is written
or if a picture of a scale with an orange on top of it is displayed? It seems to
be the case that the latter scenario is the preferred option in the NAPLAN
context, and thus more extensive research needs to be undertaken into the
merit of such a philosophy.

**Decoding and Encoding**

As noted in Chapter 2 of the study, assessment has changed dramatically
within the last 10 years with an ever-increasing presence of complex,
‘realistic’ contextual and information graphics. As a result of this, Lowrie
and Diezmann (2007) argued that students, particularly younger children, are more likely to decode rather than encode due to the fact that the graphic needs to be interpreted. This is in contrast to traditional word problems where students are encouraged to encode or draw out the information and represent it a different way. As a result there is a high emphasis on mental computation and reduced space for students to show working out. Subsequently important heuristic strategies such as drawing a diagram are no longer required in assessment and therefore not made explicit in classroom teaching. It seems ironic that in an attempt to create mathematical thinkers through the use of graphics and ‘real life’ situations we are limiting the opportunity for them to work mathematically and problem solve.

It is imperative that studies consider the extent to which students encode both concretely and in the ‘mind’s eye’ as they solve these new assessment tasks and possibly recognise its limitations. We need to ask, for example, what is the role of visual imagery and visualization in new assessment? This is particularly relevant considering the current push towards computerised high stakes testing (Hargreaves et al., 2004). What is the relationship between the ways students encode items represented on paper in a 2D form to the way they would go about encoding in 3D environments, i.e., computers?

Furthermore studies on decoding need to progress to the way in which students interpret graphics that are more dynamic. It may not be the case that ‘quality’ 2D items follow into 3D environments or indeed beyond to dynamic 3D environments. Consequently, a traditional line graph that represented speed over time in a 2D form can also be represented in a digitised manner, where manipulations on the y-axis affect the x-axis and
vice versa. Other dynamic presentations might include weather maps where cloud cover is displayed over a 12-hour period in fast time. These types of graphic representations are currently available but are a considerable time away from assessment for young students, nevertheless it is worthwhile to begin considering these issues in relation to how students decode information.

**Validity of Mathematics Assessment**

According to Hargreaves et al. (2004), the notion of validity has always been an issue for traditional pencil-and-paper tests. That is, does assessment measure what it is supposed to measure and how well does it do this? The findings in the research suggest that this is not the case for two reasons, the first being what we are attempting to assess—a child’s numeracy. Attempting to assess a child’s ability to think mathematically and appropriately apply mathematics to real life scenarios is not something that can measured using current assessment practices. As discussed in Chapter 5 through the development of the model, this mismatch of purpose in assessing a child’s numeracy current assessment design features is something that warrants further exploration, especially when the results are held in such high esteem by society and governing bodies.

The second threat to question the validity of assessment from within the results is the change in student success and understanding when an element of an item was modified. Indeed the modification paradigm appears to be an underutilised but insightful approach in determining the validity of assessment. It definitely needs to be addressed and explored further in the future.
At the beginning of this study, high-stakes assessment was presented as a complicated code. In order to crack this code it was necessary to understand the mechanics of its composition, gather necessary tools, and if possible develop a ‘bot’ or structure with universal applications. This study has attempted to equip students and teachers as code-breakers of the future. This was achieved through a detailed analysis of the structure of high-stakes assessment tasks, the required abilities and tools students would need to unlock these items, and most importantly, the creation of a ‘bot’ or model that can be applied to many forms of mathematics assessment. This study hopefully will become one of many that will provide insights but not instructions, provide tools but not answers, provide confidence but not certainty, and create thinkers not conformists of high-stakes testing today.
References


240


Greenlees, J. (2010). The terminology of mathematics assessment. In L. Sparrow, B. Kissane & C. Hurst (Eds.), *Shaping the future of mathematics


242


Appendices

Appendix A. The NAPLAN Item Test

Looking beyond the answer: The code-breaking world of mathematics assessment

2010 NAPLAN

Name:

Grade:  5

Student ID:

Date:
Question 1  The Coins Item

Gina has only these coins.

She buys a magazine for $1.95.
How much money does Gina have left?

<table>
<thead>
<tr>
<th>$1.00</th>
<th>$1.10</th>
<th>$2.00</th>
<th>$2.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Question 2  The Pie Graph Item

Hannah made a pie graph to show the number of hours she spent on different activities over 24 hours on Monday.

Which information can be found using this pie graph?
The number of

- meals Hannah eats on this day.
- hours Hannah plays sport each week.
- hours Hannah watches TV on Tuesday.
- hours Hannah spends awake on this day.
Question 3  The Peg Item

Jess takes 2 pegs out of this bag at the same time.

Which of these is impossible?

☐ a blue peg and a black peg
☐ a red peg and a red peg
☐ a green peg and a green peg
☐ a yellow peg and a black peg

Question 4  The Liquorice Allsort Item

This lolly is made with equal layers.
The layers are white or black.

What fraction of the lolly is made of black layers?

\[
\begin{array}{ccccc}
2 & 1 & 2 & 3 & 3 \\
5 & 2 & 3 & 5 & \hline
\end{array}
\]
Question 5  The Boxes Item

There are 5 small boxes in this carton, all of the same size.

How many small boxes can fit in the carton altogether?

Question 6  The Garden Plan Item

This is the plan of a garden.

What is the perimeter of the garden?

36 m  64 m  68 m  72 m
The Paper Fold Item

Question 7

Ron paints these letters on a piece of paper:

While the paint is still wet, he folds the paper along the dotted line.

When Ron unfolds the paper, what will it look like?

Question 8 The Spinner Item

This spinner is used in a board game.

Sanjey spins the arrow.

On which number is the arrow most likely to stop?

1 2 3 4
Question 9  The Street Map Item

Rick and David met on the corner of two streets. The corner is in C4 on the map.

On the corner of which two streets did Rick and David meet?

☐ Summer and Monty
☐ Grantham and Fox
☐ Duncan and Summer
☐ Summer and Fox

Question 10  The Scales Item

The price of oranges is $8 per kilogram (kg).

The cost of 10 oranges is closest to

☐ $6
☐ $15
☐ $25
☐ $80
The Fruit Juice Item

**Question 11**

Nina mixes these different juices to make a ‘Fruit Drink’.

She uses only full bottles and uses at least one of each juice. How many full bottles of each juice does Nina use to make exactly 2L of the ‘Fruit Drink’?

<table>
<thead>
<tr>
<th>Bottle of drink</th>
<th>Number of bottles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemon Juice</td>
<td></td>
</tr>
<tr>
<td>Orange Juice</td>
<td></td>
</tr>
<tr>
<td>Pineapple Juice</td>
<td></td>
</tr>
<tr>
<td>Apple Juice</td>
<td></td>
</tr>
</tbody>
</table>

The Shoe Item

**Question 12**

This is a picture of a shoe.

Which of these is closest to the length of a real shoe?

- 5 cm
- 25 cm
- 75 cm
- 100 cm
**Question 13** The School Camp Item

This is the plan of a school camp.

Jarod walks from one of the tents and goes west to the Hall.
Which tent does he walk from?

<table>
<thead>
<tr>
<th>tent 1</th>
<th>tent 2</th>
<th>tent 3</th>
<th>tent 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

**Question 14** The Baby Mass Item

These babies were born on the same day.
Which baby has the greatest mass?

<table>
<thead>
<tr>
<th>John</th>
<th>3.5 kg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sayaka</td>
<td>3.450 kg</td>
</tr>
<tr>
<td>Max</td>
<td>3.05 kg</td>
</tr>
<tr>
<td>Oscar</td>
<td>3.000 kg</td>
</tr>
</tbody>
</table>

[Images of baby faces with mass data]
A meeting is held on the first Tuesday of each month. There was a meeting held on 6 March.

What is the date of the April meeting?

April
Appendix B. Information Package for Principals

INFORMATION PACKAGE FOR PRINCIPALS

Project title: Looking beyond the answer: The code-breaking world of mathematics assessment

During 2008 to 2011, Jane Greenlees (PhD student, Charles Sturt University), is undertaking a mathematics research project that focuses on how children solve standardised test items commonly used in schools. The purpose of this project is to examine the extent to which standardised testing represents student sense making and mathematical understanding. It will also explore the impact test design has on student performance and subsequent teaching practices. During the project, all children will complete two test booklets in Year 3 and Year 5, with a select few participating in two one-on-one interviews each year with the researcher. All interviews will be tape recorded.

As test item design has changed dramatically over the last 10 years, the results of this study will provide an understanding of the influence this has had on student success. Knowing the way in which test items are constructed and how it impacts on student’s capacity to make sense of mathematics will help teachers make informed decisions about student results as well as current teaching practices.

Children’s participation in this research study is voluntary and they can withdraw at any time. There are no out-of-the-ordinary risks associated with this research and there will be no discomfort to the children. In all reporting of the research and any publications, the identity of the children and the school will remain anonymous.

The University requires informed consent for all participants and I am seeking consent for your school to participate in the study. The tests will be conducted by me, a registered teacher, in the students’ classrooms as a normal part of school maths activities in consultation with the class teacher. The interviews will take place in a location determined by the teacher but familiar to all students.

At the conclusion of the project, a summary of research findings will be provided to the school to be made available to interested teachers and parents/guardians.

Thank you for considering your school’s participation in this study. This project has ethical clearance from CSU and has been approved by the Wagga Wagga Catholic Diocese. Any queries about the project should be directed to Jane Greenlees at j.greenlees@csu.edu.au or ph 02 6925 0215. If you have any concerns or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer:

The Executive Officer
Ethics in Human Research Committee
Academic Secretariat
Charles Sturt University
Private Mail Bag 29
Bathurst NSW 2795

Tel: (02) 6338 4628
Fax: (02) 6338 4194

Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.

If you agree to your school’s participation in this project, please complete the relevant section of the accompanying Principal’s Consent Form and return to Jane Greenlees at Charles Sturt University, RIPPLE, Locked Bag 588, Wagga Wagga NSW 2658 or fax to 02 6933 2962.

Yours sincerely,

Jane Greenlees
Appendix C. Principal Consent Form

PRINCIPAL CONSENT FORM

Project title: Looking beyond the answer: The code-breaking world of mathematics assessment

Chief Investigator:
Jane Greenlees (PhD student CSU, jgreenlees@csu.edu.au or 02 6925 0215)

Statement of consent

This form and the accompanying Information Package have been provided to advise you of the mathematics research project that is to be conducted from 2008 to 2010; to inform you of your rights about your school's participation in this project; and to invite you provide consent for your school's participation in this project.

By signing below, you are indicating that you:

- have read and understand the information package about this project;
- have had any questions answered to your satisfaction;
- understand that you can contact the Chief Investigator if you have any additional questions about the project;
- understand that children are free to withdraw at any time, without comment or penalty;
- understand that work samples of the children from the project may be used in reporting the results of this research; and
- agree to your school's participation in this project.

Charles Sturt University's Human Research Ethics Committees have approved this study. If you have any concerns about the ethical conduct of this project you may also contact:

The Executive Officer
Ethics in Human Research Committee
Academic Secretariat
Charles Sturt University
Private Mail Bag 29
Bathurst NSW 2795

Tel: (02) 6338 4628
Fax: (02) 6338 4194

I give permission for my school ___________________________ to participate in the mathematics assessment project. I understand that I am free to withdraw my consent at any time during the project and that any information or personal details gathered in the course of this research are confidential and that my name, children’s names, the school’s name nor any other identifying information will be used or published without written permission.

Principals' signature: ___________________________ Date: __________________

Please return to Jane Greenlees at:
Charles Sturt University, RIPPLE, Locked Bag 588
Wagga Wagga NSW 2658 or fax to 02 6933 2962
Appendix D. Information Package for Parents/Guardians

INFORMATION PACKAGE FOR PARENTS/GUARDIANS OR CAREGIVERS

Project title: Looking beyond the answer: The code-breaking world of mathematics assessment

During 2008 to 2011, Jane Greenlees (PhD student, Charles Sturt University), is undertaking a mathematics research project that focuses on how children solve standardised test items commonly used in schools. The purpose of this project is to examine the extent to which standardised testing represents student sense making and mathematical understanding. It will also explore the impact test design has on student performance and subsequent teaching practices. During the project, all children will complete two test booklets in Year 3 and Year 5, with a select few participating in two one-on-one interviews each year with the researcher. All interviews will be tape recorded.

As test item design has changed dramatically over the last 10 years, the results of this study will provide an understanding of the influence this has had on student success. Knowing the way in which test items are constructed and how it impacts on student’s capacity to make sense of mathematics will help teachers make informed decisions about student results as well as current teaching practices.

Children’s participation in this research study is voluntary and they can withdraw at any time. There are no out-of-the-ordinary risks associated with this research and there will be no discomfort to the children. The test items used will be familiar to the students and should not cause any undue anxiety. In all reporting of the research and any publications, the identity of the children and the school will remain anonymous.

The University requires informed consent for all participants and I am seeking consent for your school to participate in the study. The tests will be conducted by me, a registered teacher, in the students’ classrooms as a normal part of school maths activities in consultation with the class teacher. The interviews will take place in a location determined by the teacher but familiar to all students.

At the conclusion of the project, a summary of research findings will be provided to the school to be made available to interested teachers and parents/guardians.

Thank you for considering your school’s participation in this study. This project has ethical clearance from CSU and has been approved by the Wagga Wagga Catholic Diocese. Any queries about the project should be directed to Jane Greenlees at jgreenlees@csu.edu.au or ph 02 6925 0215. If you have any concerns or reservations about the ethical conduct of this project, you may contact the Committee through the Executive Officer:

The Executive Officer
Ethics in Human Research Committee
Academic Secretariat
Charles Sturt University
Private Mail Bag 29
Bathurst NSW 2795

Tel: (02) 6338 4628
Fax: (02) 6338 4194

Any issues you raise will be treated in confidence and investigated fully and you will be informed of the outcome.

If you agree to your child’s participation in this project, please complete the relevant section of the accompanying Parent/Guardian Consent Form and return it to your child’s teacher by X (date).

Yours sincerely,

Jane Greenlees
Appendix E. Parent/Guardian Consent Form

PARENT/GUARDIAN CONSENT FORM

Project title: Looking beyond the answer: The code-breaking world of mathematics assessment

Chief Investigator:
- Jane Greenlees (PhD student CSU) (jgreenlees@csu.edu.au or 02 6925 0215)

Statement of consent
This form and the accompanying Information Package has been provided to advise you of the mathematics research project that is being conducted from 2008 to 2010; to inform you of your rights about your child’s participation in this project; and to invite you to provide consent to your child’s participation in this project.

By signing below, you are indicating that you:
- have read and understand the information package about this project;
- have had any questions answered to your satisfaction;
- understand that you can contact the Chief Investigator if you have any additional questions about the project;
- understand that children are free to withdraw at any time, without comment or penalty;
- understand that work samples of the children from the project may be used in reporting the results of this research; and
- agree to your child’s participation in this project.

Charles Sturt University’s Human Research Ethics Committees have approved this study. If you have any concerns about the ethical conduct of this project you may also contact:
The Executive Officer
Ethics in Human Research Committee
Academic Secretariat
Charles Sturt University
Private Mail Bag 29
Bathurst NSW 2795
Tel: (02) 6338 4628
Fax: (02) 6338 4194

I give permission for my son/daughter ____________________________ in class ____________ to participate in the Mathematics Assessment project. I understand that I am free to withdraw my consent at any time during the project and that any information or personal details gathered in the course of this research are confidential and that my name, children’s names, the school’s name nor any other identifying information will be used or published without written permission.

Parent’s/Guardian’s Signature: ____________________________ Date: ____________
Appendix F. Information Package for Teachers

INFORMATION PACKAGE FOR TEACHERS

Project title: Looking beyond the answer: The code-breaking world of mathematics assessment

During 2008 to 2011, Jane Greenlees (PhD student, Charles Sturt University), is undertaking a mathematics research project that focuses on how children solve standardised test items commonly used in schools. The purpose of this project is to examine the extent to which standardised testing represents student sense making and mathematical understanding. It will also explore the impact test design has on student performance and subsequent teaching practices. During the project, all children will complete two test booklets in Year 3 and Year 5, with a select few participating in two one-on-one interviews each year with the researcher.

Your principal has agreed to your school’s participation in the project over the three year period 2008-2010. Year 3 students will participate in this study in 2008. The test will be conducted in your classroom as a normal part of class maths activities. The interviews will be arranged at a time convenient to your classroom teaching. The chief investigator is a registered teacher and will consult with you in regards to suitable times to conduct the test and interviews in order to minimise disruption to your teaching program.

Thank you for your anticipated cooperation with this research. As test item design has changed dramatically over the last 10 years, the results of this study will provide an understanding of the influence this has had on student success. Knowing the way in which test items are constructed and how it impacts on student’s capacity to make sense of mathematics will help teachers make informed decisions about student results as well as current teaching practices.

Yours sincerely

Jane Greenlees
Appendix G. Interview Protocol

2010 NAPLAN Item Test Interview Protocol

INTRODUCTION

Welcome the student and thank him/her for coming. Introduce yourself,

“My name is xxxxx and I work at Charles Sturt University or CSU.”

Explain the project.

“At the university we are finding out about how children work out maths tasks. We have taken these from this year’s NAPLAN test.”

Show them some examples of the tasks from the booklet.

“I am going to be tape recording whilst you do the interview. I do this to help me remember what you said and did. We will use the information you tell us to help teachers to teach their students about how to use different kinds of maths tasks.”

Explain that the interview has nothing to do with school or their reports.

Highlight the point,

“It doesn’t matter if you get the question right or wrong, what is important is that you are able to explain to me how you worked the answer out.”

The Interview

Turn on the tape recorder. Say

“This is (name) from (class) and we are doing interview 1 on (date).

Stop audio tape.

“Today we are going to be looking at different sorts of tasks. All these tasks have words and some have words and numbers. Most of them have a diagram or a picture.”

“I would like you to work out the answers one page at a time and then we will come back and have a talk about them. It is OK to draw on the page if, you can use the space under the task or on the spare page if you need to do any working out.”

When the child has completed one page, start the audio tape and say

“(name) could you read out the first task on the page and tell me how you worked it out.”
Use open questions as much as possible, for example:

- Show me how you used the diagram.
- Where did you start? Tell me why you decided to start there.
- Where did you go from there?
- What were you thinking in your head?

Try to gain as much information as possible without putting the child under any undue pressure or leading them to an answer. If the child has done any working out ask them about it and describe what they have drawn and where it is on the page.

“That’s a great explanation/idea” or something similar. “Let’s look at the second task now. How did you work this one out?”

“That is great (name). Is there anything else you would like to tell me? Let’s look at the next page.”

Continue on following this procedure throughout the booklet.

End of Interview

“Is there anything we’ve done here today that you’d like to tell me more about?”

Thank the child for coming and doing the interview and let them know we will be doing more activities with them later in the year.

Give the student their thank you gift.

Stop tape. Complete details on the cover of the booklet.
Appendix I. The Modified Test

Looking beyond the answer: The code-breaking world of mathematics assessment

2010 Modified NAPLAN

Name:

Grade:  5

Student ID:

Date:

Previously published material in this book is copied on behalf of Charles Sturt University pursuant to Part VB of the Commonwealth Copyright Act 1996.
Question 1  The Coins Item

Gina has $4.05.
She buys a magazine for $1.95.
How much money does Gina have left?

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.00</td>
<td>$1.10</td>
<td>$2.00</td>
<td>$2.10</td>
</tr>
</tbody>
</table>

Question 2  The Pie Graph Item

This pie graph shows the number of hours on Monday Hannah spent doing different activities.

- sleep
- music
- eating
- sport
- watching TV
- school

Which information can be found using this pie graph?
The number of

- meals Hannah eats on this day.
- hours Hannah plays sport each week.
- hours Hannah watches TV on Tuesday.
- hours Hannah spends awake on this day.
Question 3  The Liquorice Allsort Item

A cake is made with five equal layers.
The layers are two white and three pink.
What fraction of the cake is made of white layers?

\[ \frac{2}{5}, \frac{1}{2}, \frac{2}{3}, \frac{3}{5} \]

The Boxes Item

Question 4

There are 5 small boxes in this carton, all of the same size.

How many small boxes can fit in the carton altogether?

12, 16, 20, 24
Question 5  The Garden Plan Item

This is the plan of a garden.

What is the perimeter of the garden?

36 m  64 m  68 m  72 m

Question 6  The Scales Item

One apple weighs 0.250 kg.
If the price of apples is $6 per kilogram (kg), the cost of 10 apples is closest to

$6  $15  $25  $60

Year 5 2010 – Modified NAPLAN 3
The Street Map Item

Question 7

Rick and David met on the corner of two streets.

The corner is in D4 on the map.

On the corner of which two streets did Rick and David meet?


The Shoe Item

Question 8

Which of these is closest to the length of your shoe?

5 cm  25 cm  75 cm  100 cm

○   ○   ○   ○
Question 9  The Baby Mass Item

These babies were born on the same day.
Which baby has the greatest mass?

[Images of babies with masses: 34 kg, 38.5 kg, 30.5 kg, 200 kg]

Question 10  The Calendar Item

A meeting is held on the first Tuesday of each month.
There was a meeting held on 6 March.
What is the date of the April meeting?

[Calendar for March]

<table>
<thead>
<tr>
<th>Sun</th>
<th>Mo</th>
<th>Tu</th>
<th>Wed</th>
<th>Th</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
</tr>
</tbody>
</table>

April
# Appendix J. Quantitative Data Analysis ‘Smiley Faces’

<table>
<thead>
<tr>
<th>Code</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GR</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SM</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JW</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SK</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>😊😊😊😊😊😊😊😊😊😊😊😊</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Correct Test A (cohort + interviews) | 62 | 46 | 59 | 34 | 18 | 27 | 61 | 52 | 41 | 14 |
| Correct Test B (cohort + interviews) | 57 | 50 | 53 | 50 | 26 | 32 | 56 | 60 | 46 | 33 |
| Total Students                      | 66 | 66 | 66 | 66 | 66 | 66 | 66 | 66 | 66 | 66 |
| Difference                          | -5 | +4 | -6 | +10 | +8 | +5 | -5 | +8 | +5 | +19 |
# Appendix K. Smiley Faces Individual School Results

## School 1 Interviews 2010

<table>
<thead>
<tr>
<th>Code</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>CM</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>BM</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>AD</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>BC</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>VN</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>BT</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>AL</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>SK</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>NS</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>AM</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
</tbody>
</table>

Correct Test A (cohort + interviews): 62 46 59 34 13 27 61 52 41 14
Correct Test B (cohort + interviews): 57 50 53 50 26 32 56 60 46 33
Total Students: 66 66 66 66 66 66 66 66 66 66
Difference: -5 +4 -6 +15 +8 +5 -5 +8 +5 +19

## School 2 Interviews 2010

<table>
<thead>
<tr>
<th>Name</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>EK</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>MA</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>RW</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>BG</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>JW</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
<tr>
<td>HE</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
</tbody>
</table>

Correct Test A (cohort + interviews): 19 12 18 10 3 5 15 16 10 3
Correct Test B (cohort + interviews): 17 14 17 15 7 7 17 18 15 10
Total Students: 19 19 19 19 19 19 19 19 19 19
Difference: +2 +2 -1 +5 +4 +2 -2 +2 +5 +7
### School 3 Interviews 2010

<table>
<thead>
<tr>
<th>Name</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correct Test A (cohort + interviews): 37 28 39 32 24 19 40 33 32 11
Correct Test B (cohort + interviews): 38 33 33 30 28 17 40 34 24 24
Total Students: 41 41 41 41 41 41 41 41 41 41
Difference: +1 -5 -2 +1 -3 8 13

### School 4 Interviews 2010

<table>
<thead>
<tr>
<th>Name</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correct Test A (cohort + interviews): 13 11 17 17 2 6 17 11 8 5
Correct Test B (cohort + interviews): 13 13 14 16 6 7 17 14 12 7
Total Students: 17 17 17 17 17 17 17 17 17 17
Difference: +2 -3 +6 +4 +1 -3 +4 +1