Improving the function of multiple-choice items in the assessment of the skills necessary for the development of proportional reasoning

Joan Martina Burfitt

Bachelor of Science (Honours), Diploma of Education, Graduate Diploma of Applied Science (Computer Studies), Master of Education (Honours)

This thesis is presented for the degree of Doctor of Philosophy of The University of Western Australia
Graduate School of Education
September 2019
THESIS DECLARATION

I, Joan Martina Burfitt, certify that:

This thesis has been substantially accomplished during enrolment in this degree.

This thesis does not contain material which has been submitted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

In the future, no part of this thesis will be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree.

This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text and, where relevant, in the Authorship Declaration that follows.

This thesis does not violate or infringe any copyright, trademark, patent, or other rights whatsoever of any person.

Approval: RA/4/1/7791.

Written student and parental consent have been received and archived for the research involving student interviews reported in this thesis.

This thesis contains only sole-authored work, some of which has been published and/or prepared for publication under sole authorship.

Signature: [Redacted]

Date: 6 September 2019
ABSTRACT

The information that multiple-choice items (MC) provide about student learning is perceived to be limited, partly because the scores may be contaminated by guessing, and also because the items test a narrow range of student skills and understanding. The aim of this study is to improve the quality of information about student achievement that may be gathered from MC items by attending to their construction and to the analysis of responses to these items.

To achieve this aim, Rasch Measurement theory was applied during the analysis of three sets of responses from lower secondary students in tests of skills associated with the development of proportional reasoning. First, the results of analysing the constructed responses from students for questions on ratios and fractions were used to identify and to confirm the skills and understandings that are associated with partial knowledge of some proportional reasoning concepts. Second, the analysis of the responses of students to NAPLAN numeracy tests provided evidence that these tests contain MC items which can detect partial knowledge, even though the items were not created to identify and score this partial knowledge.

The third set of responses was collected from an online test conducted with Year 8 students in Western Australia in 2016. The test contained six blocks of 10 MC items and all students were offered the same block to start the test. The students were allocated to one of two test designs: a non-adaptive design, where blocks were allocated at random; or an adaptive design, where blocks were allocated according to prior success. One of the distractors (incorrect options) in each item was purposefully written to provide an opportunity for the students to show their partial knowledge of the concept. Selection of this informative distractor attracted a score of 1 while the correct response scored 2 and the other distractors scored zero.

The results confirmed that it is possible to create options which can detect partial knowledge, and that giving such options a partial score adds to the information gathered about student learning. The collection of more information can increase the efficiency of a test without adding to the demands on the responding students. With two designs for the test it was possible to confirm the presence of guessing in the non-adaptive design and to show that removing responses that were likely to have been guessed provides more precise measures of the difficulty of the items and the proficiency of the students. This process of accommodating guessing can be used with a variety of ages and subject areas, and it can reduce any advantage that weaker students might gain in tests containing considerable numbers of MC items.
ACKNOWLEDGEMENTS

Throughout the time that this research was conducted, several groups and individuals have provided invaluable support. With the benefit of their financial support, time, talents, and wisdom, it has been possible for me to achieve many of the planned and unplanned outcomes that have emerged from this research.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship. The funding provided by the Pearson Psychometric Laboratory was most welcome and it supported the financing of opportunities for study and collaboration with Australian and international colleagues. The provision of financial support has enabled me to undertake a project of this quality and magnitude.

Some of the data used in this publication are sourced from the Australian Curriculum, Assessment and Reporting Authority (ACARA) and are available from ACARA in accordance with its Data Access Protocols. Other data from empirical studies in the United Kingdom were provided by Professor Jeremy Hodgen (The University of Nottingham) and Professor Robert Coe (Durham University).

I would like to express my gratitude to the students and teachers who supported this project and to the system administrators who provided permission for the research to occur. Many teachers provided the opportunities for their students to participate in this research and the involvement of the students was substantial, thus allowing considerable data to be collected for analysis. I am particularly grateful to the teachers and students who participated in the cognitive interviews, and to the parents who provided their permission for this process. Other teachers in Western Australia provided commendable support for the project, including Judith Cumpsty, Bev Hine, Wendy Pero, and Ann Shugg.

At times, there were individuals who gave their inspiring suggestions or their time to respond to my questions. People whose generosity I would like to acknowledge are Lorraine Day, Derek Hurrell, Per Kind, Caroline Long, Diana Ng, Ray Philpot, Pauline Rogers, Di Siemon, Vicki Steinle, Elena Stoyanova, and Stavroula Zoumboulis.

Inspirational beyond imagination were my two supervisors, Professor Peter Merrotsy and Professor David Andrich. With their many suggestions and considerable feedback, I have been able to achieve much more than I thought possible.
AUTHORSHIP DECLARATION: SOLE AUTHOR PUBLICATIONS

This thesis contains the following sole-authored work that has been published and prepared for publication.

<table>
<thead>
<tr>
<th>Details of the work:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Location in thesis:</td>
<td>This article presents a very brief overview of some aspects of the material covered in Chapters 3, 7 and 9. It is not referenced in the thesis.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Details of the work:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Location in thesis:</td>
<td>This article presents some of the findings which are reported in Chapter 9. It is not referenced in the thesis.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Details of the work:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Location in thesis:</td>
<td>The first part of Chapter 7 describes in detail the process and results of identifying partial knowledge in constructed-response items. The article is referenced in the thesis.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Details of the work:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Location in thesis:</td>
<td>Chapter 8: Item review</td>
</tr>
</tbody>
</table>

Signature: [Redacted] Date: 6 September 2019
**TABLE OF CONTENTS**

**CHAPTER 1: Introduction**
1.1 Rationale .................................................................................................................. 1
1.2 Assessment of mathematical understanding .............................................................. 3
1.3 Concerns with the current use of MC items ............................................................... 5
1.4 Planning for the improved function of MC items ..................................................... 6
1.5 Significance of this study .......................................................................................... 8
1.6 Overview of the dissertation .................................................................................... 9

**CHAPTER 2: Assessment of mathematical achievement**
2.1 Assessment ............................................................................................................. 11
  2.1.1 The nature of assessment .................................................................................. 12
  2.1.2 The purpose and outcomes of conducting assessments .................................... 13
2.2 Types of assessment ............................................................................................... 15
2.3 Assessments for secondary students ....................................................................... 17
2.4 Developing an assessment framework ................................................................... 20
  2.4.1 TIMSS framework ......................................................................................... 21
  2.4.2 PISA framework ............................................................................................ 21
  2.4.3 National testing in the United States of America ......................................... 22
  2.4.4 National testing in the United Kingdom ......................................................... 23
  2.4.5 SOLO Taxonomy ............................................................................................ 24
  2.4.6 Bloom’s Taxonomy ....................................................................................... 24
  2.4.7 Designing a test framework ......................................................................... 27
2.5 Mathematical achievement .................................................................................... 28
2.6 Item types ............................................................................................................... 30

**CHAPTER 3: Multiple-choice items**
3.1 Types of items: MC and CR .................................................................................... 32
3.2 Attraction of MC items ......................................................................................... 35
3.3 Creation of MC items ............................................................................................ 36
3.4 Concerns with the use of MC items ....................................................................... 42
  3.4.1 Guessing in MC items ................................................................................... 42
  3.4.2 Low level thinking in MC items ..................................................................... 48
  3.4.3 Consideration of partial credit in MC items .................................................. 50
  3.4.4 Quality of the distractors in MC items ......................................................... 52
  3.4.5 Perception of gender bias .............................................................................. 54
3.5 Improving the function of MC items ....................................................................... 57

**CHAPTER 4: Proportional reasoning**
4.1 Importance ............................................................................................................ 58
4.2 Defining proportional reasoning ........................................................................... 60
4.3 Behaviours associated with proportional reasoning ........................................... 64
4.4 Skills and understandings needed for sound proportional reasoning .................. 65
4.5 Development in proportional reasoning ............................................................... 69
4.5.1 Aspects of the development of proportional reasoning ........................................... 70
4.5.2 Development pathways ............................................................................................. 75
4.6 Challenges for learners ................................................................................................ 84

CHAPTER 5: Partial knowledge
5.1 Selection of options ..................................................................................................... 85
5.2 Informative distractors ................................................................................................. 87
5.3 Development of knowledge ......................................................................................... 90
5.4 Partial knowledge in proportional reasoning ............................................................. 95
5.5 Rewarding partial knowledge in MC items ................................................................. 100
5.6 Research questions .................................................................................................... 103

CHAPTER 6: Rasch measurement theory
6.1 Research using Rasch measurement theory ................................................................. 106
6.1.1 Developmental scales in mathematics .................................................................. 106
6.1.2 Factors affecting performance .............................................................................. 108
6.2 The Rasch model ........................................................................................................ 111
6.2.1 Determining the estimates .................................................................................... 112
6.2.2 Reading the scale ................................................................................................. 113
6.2.3 Identification of item difficulty ............................................................................ 114
6.2.4 Polytomous scoring ............................................................................................... 116
6.2.5 Targeting ............................................................................................................... 117
6.2.6 Linking ................................................................................................................... 118
6.3 Fit of data to the model ............................................................................................... 119
6.4 Differential item functioning ..................................................................................... 125
6.5 Accommodating guessing .......................................................................................... 127
6.6 Partial credit and distractor analysis .......................................................................... 130

CHAPTER 7: RESULTS: Identification of partial knowledge
7.1 Partial knowledge in constructed-response items ....................................................... 136
7.1.1 Constructed-response items in the ICCAMS tests ............................................... 136
7.1.2 Partial credit in constructed-response (ICCAMS) items ....................................... 137
7.1.3 Results of awarding partial credit ......................................................................... 140
7.1.4 Type of partial knowledge identified in ICCAMS tests ......................................... 144
7.1.5 Further outcomes ................................................................................................. 146
7.2 Partial knowledge in multiple-choice items ............................................................. 147
7.2.1 Conjectures about existing MC items .................................................................... 147
7.2.2 The NAPLAN tests ............................................................................................... 148
7.2.3 Analysis of MC items ............................................................................................ 149
7.2.4 Effects of awarding partial credit in MC (NAPLAN) items ................................... 156
7.2.5 Implications of recognising partial knowledge in MC (NAPLAN) items ............. 160
7.3 Merit of awarding partial credit for partial knowledge .............................................. 169
CHAPTER 10: Further discussion and conclusion

Research questions answered ................................................................. 239
10.1 Research activity .............................................................................. 240
10.2 MC items to address higher-order thinking ......................................... 242
10.3 Controlling for guessing in MC items ................................................... 244
10.4 Informative distractors ...................................................................... 247
10.5 Partial knowledge ............................................................................. 248
10.6 Partial credit for partial knowledge ..................................................... 250
10.7 Perceived gender bias ....................................................................... 251
10.8 Further outcomes ............................................................................ 251
10.9 Concluding remarks ........................................................................ 253

REFERENCES

References ............................................................................................. 254
FIGURES

Figure 2.1 Framework design for PISA test of mathematical literacy ........................................ 22
Figure 2.2 Framework (in part) for national testing in the United Kingdom .................................. 23
Figure 2.3 Part classification of content and behaviours in mathematics ...................................... 25
Figure 2.4 Overview of Revised Bloom’s Taxonomy ........................................................................ 26

Figure 3.1 Example of a poorly-written MC item ............................................................................. 37
Figure 3.2 MC item deemed to test higher-order thinking ............................................................... 48
Figure 3.3 MC item where selection of distractor may indicate partial knowledge ....................... 50
Figure 3.4 Achievement in NAPLAN Numeracy in Years 7 and 9: 2015–2017 ............................ 54

Figure 5.1 MC item addressing different fine-grained skills ......................................................... 89
Figure 5.2 MC item for proportional reasoning (1) ......................................................................... 92
Figure 5.3 MC item for proportional reasoning (2) ......................................................................... 96

Figure 6.1 Category probability curve (CPC) ................................................................................. 113
Figure 6.2 Item characteristic curves (ICC) for Items 4, 19, and 20 ............................................. 114
Figure 6.3 ICC with observed responses for Item 2 ......................................................................... 115
Figure 6.4 Category probability curve (CPC) for Item 12 ............................................................... 116
Figure 6.5 Person-item threshold distribution ................................................................................. 118
Figure 6.6 ICC for ill-fitting items .................................................................................................. 121
Figure 6.7 Items with different discrimination ................................................................................. 123
Figure 6.8 ICC showing no DIF by gender ..................................................................................... 125
Figure 6.9 ICC showing DIF by year group ................................................................................... 126
Figure 6.10 Observed responses indicative of guessing ................................................................. 127
Figure 6.11 Distractor selections indicative of guessing ................................................................. 128
Figure 6.12 Category probability curves for Items 17 and 13 ......................................................... 130
Figure 6.13 Distractor curves and content structure for Item 60 .................................................... 131
Figure 6.14 Graph showing location of maximum information ..................................................... 133

Figure 7.1 Category probability curve (CPC) for Item 17 ............................................................... 138
Figure 7.2 Category probability curve (CPC) for Item 9 ................................................................. 139
Figure 7.3 Item locations for dichotomous items in first and final analyses ................................. 140
Figure 7.4 Item locations for all items in first and final analyses .................................................. 141
Figure 7.5 Person-item threshold distributions with and without partial credit ......................... 142
Figure 7.6 Information functions for analyses with and without partial credit ............................. 143
Figure 7.7 Distractor curves indicating various levels of information .......................................... 150
Figure 7.8 Category probability curves with ordered and disordered thresholds ....................... 153
Figure 7.9 Distractor curves indicative of guessing ....................................................................... 154
Figure 7.10 Summary of analyses of NAPLAN data ....................................................................... 155
Figure 7.11 Person-item threshold distributions with and without partial credit ......................... 157
Figure 7.12 Distractor curves for items with ordered thresholds ................................................... 162
Figure 7.13 Distractor curves for items with disordered thresholds (Items 18, 39, 43)............. 164
Figure 7.14 Distractor curves for items with thresholds ordered but not expected ..................... 165
Figure 8.1  An item designed to be a difficult item for a Year 8 student.................. 174
Figure 8.2  An item designed at a low level of difficulty for a Year 8 student............. 174
Figure 8.3  Pathway of blocks in adapted version of Year 8 test.............................. 177
Figure 8.4  First ten items of Year 8 test.................................................................. 182
Figure 8.5  Item 55 showing original and edited version........................................... 185
Figure 8.6  Item 43 before and after editing................................................................. 187

Figure 9.1  Comparison of item difficulty estimates: non-adaptive design ................. 203
Figure 9.2  Distribution of person proficiency: responses not tailored.......................... 206
Figure 9.3  Distribution of person proficiency: responses tailored............................... 206
Figure 9.4  Comparison of item difficulty estimates: adaptive design.......................... 207
Figure 9.5  Distractor curves for Item 30......................................................................... 212
Figure 9.6  Distractor curves for Item 44......................................................................... 213
Figure 9.7  CPC Item 30.............................................................................................. 214
Figure 9.8  Skills associated with full and partial knowledge........................................ 217
Figure 9.9  Comparison of item difficulty estimates: both designs.................................. 221
Figure 9.10 Comparison of threshold values: mean item anchoring......................... 225
Figure 9.11 Comparison of threshold values: individual item anchoring................. 226
Figure 9.12 Person-item threshold distribution: dichotomous scoring.................... 230
Figure 9.13 Person-item threshold distribution: partial credit given.......................... 230
Figure 9.14 Histogram of person proficiency for both test designs: no partial credit.... 232
Figure 9.15 Histogram of person proficiency for both test designs: partial credit........ 232
Figure 9.16 Histogram for person proficiency and item difficulty: guessing included ... 234
Figure 9.17 Histogram for person proficiency and item difficulty: guessing excluded... 234
Figure 9.18 Item showing DIF for gender...................................................................... 235
Figure 9.19 Item map showing the order of threshold difficulty............................... 237
TABLES

Table 2.1  MC items in student assessments ............................................................... 31
Table 4.1  Behaviours associated with proportional reasoning .............................. 64
Table 4.2  Proportion in the curriculum documents .............................................. 65
Table 4.3  Curriculum content of skills for proportional reasoning ...................... 66
Table 4.4  Proficiencies relating to proportional reasoning in Years 6 to 8 ............ 68
Table 4.5  Stages of development in the concept of ratio .................................... 78
Table 4.6  Developmental hierarchy ........................................................................ 80
Table 4.7  Student attributes in a learning progression ......................................... 83
Table 5.1  Popular incorrect responses in MC items ........................................... 94
Table 7.1  Proportions of students receiving partial and full credit ..................... 146
Table 7.2  Partial credit scoring and indications of PK from distractor curves ....... 152
Table 7.3  Threshold distances in items scored with partial credit ....................... 159
Table 7.4  Stimulus for Item 23: Year 7, 2013 ......................................................... 168
Table 8.1  Block structure for Year 8 test ................................................................. 175
Table 8.2  Block structure for Block 1: Year 8 at the low level of difficulty .......... 178
Table 8.3  Partial knowledge skills in Block 1 (Items 1 to 10) ............................... 181
Table 8.4  Numbers of participants for each block .............................................. 193
Table 8.5  Numbers of participants for different test designs .............................. 194
Table 8.6  Changes in items for different levels of tailoring ................................. 196
Table 9.1  Dichotomous scoring statistics for three test situations .................... 201
Table 9.2  Mean item difficulty: Blocks in both designs ...................................... 202
Table 9.3  Mean item difficulty: Blocks for tailored responses ............................ 205
Table 9.4  Items with ordered thresholds ............................................................... 210
Table 9.5  Number of items with ordered thresholds ........................................... 211
Table 9.6  Identifying partial knowledge from distractor curves ........................... 215
Table 9.7  Comparing fit for three sets of estimates ............................................ 222
Table 9.8  Awarding partial credit ............................................................................ 229
Table 9.9  Gender differences with guessing ....................................................... 233
ACRONYMS AND ABBREVIATIONS USED IN THIS MANUSCRIPT

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAMT</td>
<td>Australian Association of Mathematics Teachers</td>
</tr>
<tr>
<td>AC</td>
<td>All options</td>
</tr>
<tr>
<td>ACARA</td>
<td>Australian Curriculum and Reporting Authority</td>
</tr>
<tr>
<td>ACSF</td>
<td>Australian Core Skills Framework</td>
</tr>
<tr>
<td>AMC</td>
<td>Australian Mathematics Competition</td>
</tr>
<tr>
<td>CAT</td>
<td>Computer and Algorithmic Thinking</td>
</tr>
<tr>
<td>CO</td>
<td>Correct response only</td>
</tr>
<tr>
<td>CPC</td>
<td>Category probability curve</td>
</tr>
<tr>
<td>CR</td>
<td>Constructed response</td>
</tr>
<tr>
<td>DIF</td>
<td>Differential item functioning</td>
</tr>
<tr>
<td>FS</td>
<td>Formula scoring</td>
</tr>
<tr>
<td>ICC</td>
<td>Item characteristic curve</td>
</tr>
<tr>
<td>ICCAMS</td>
<td>Increasing Competence and Confidence in Algebra and Multiplicative Structures</td>
</tr>
<tr>
<td>IM</td>
<td>Incorrect method</td>
</tr>
<tr>
<td>IS</td>
<td>Incomplete solution</td>
</tr>
<tr>
<td>MAWA</td>
<td>Mathematical Association of Western Australia</td>
</tr>
<tr>
<td>MC</td>
<td>Multiple-choice</td>
</tr>
<tr>
<td>MCQ</td>
<td>Multiple choice questionnaire</td>
</tr>
<tr>
<td>NAPLAN</td>
<td>National Assessment Programme - Literacy and Numeracy</td>
</tr>
<tr>
<td>NAEP</td>
<td>National Assessment of Educational Progress</td>
</tr>
<tr>
<td>NOTA</td>
<td>None of the above</td>
</tr>
<tr>
<td>NR</td>
<td>Number right</td>
</tr>
<tr>
<td>NRET</td>
<td>Number right elimination testing</td>
</tr>
<tr>
<td>OLNA</td>
<td>Online Literacy and Numeracy Assessment</td>
</tr>
<tr>
<td>PaCCs</td>
<td>Partially correct constructs</td>
</tr>
</tbody>
</table>
ACRONYMS AND ABBREVIATIONS (cont’d)

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>Proportion ignored</td>
</tr>
<tr>
<td>PISA</td>
<td>Programme for International Student Assessment</td>
</tr>
<tr>
<td>PK</td>
<td>Partial knowledge</td>
</tr>
<tr>
<td>PSI</td>
<td>Person separation index</td>
</tr>
<tr>
<td>RE</td>
<td>Reasonable estimate</td>
</tr>
<tr>
<td>RUMM</td>
<td>Rasch unidimensional measurement model</td>
</tr>
<tr>
<td>SCSA</td>
<td>School Curriculum and Standards Authority</td>
</tr>
<tr>
<td>SMART</td>
<td>Specific Mathematics Assessments that Reveal Thinking</td>
</tr>
<tr>
<td>SOLO</td>
<td>Structure of the Observed Learning Outcome</td>
</tr>
<tr>
<td>TIMSS</td>
<td>Trends in International Mathematics and Science Study</td>
</tr>
<tr>
<td>WACE</td>
<td>Western Australian Certificate of Education</td>
</tr>
</tbody>
</table>
CHAPTER 1: Introduction

The multiple-choice (MC) item is generally perceived to be a less than ideal question format to use in the assessment of achievement, and particularly for gathering information about student understanding in mathematics. It typically consists of an initial statement or question, followed by four or five expressions, only one of which is correct, that complete the statement or answer the question. There are four major concerns with the use of the MC format, and the first relates to the perception that MC items are inadequate for testing high-level thinking. The second concern is that students gain too much credit for guessing and in the opportunity to guess, the information gathered about what students know and understand is inaccurate. The perception that the MC format provides an advantage for males is the third concern, and the final concern relates to the limited information that is gathered about each student’s conceptual knowledge when they respond to a MC item.

The focus of this study is to provide information that can be used to improve the function of MC items in the assessment of student achievement and learning of mathematical skills and understandings. The assessment of achievement is an influential component of the education process and the results obtained by the students are often used to make important decisions about their future learning directions. It is imperative that the processes by which student learning is assessed are implemented in such a way as to provide measures of achievement that are as accurate and reliable as possible. While the context for which this investigation is designed and executed is specifically the mathematical achievement of students in early secondary school, the results can be applied to a variety of subject learning areas and across a wide range of ages.

1.1 Rationale

Considerable time is required to create good tasks which assess student learning, which are practical to conduct, and for which the marking process is accurate, fair, and manageable. Multiple-choice (MC) items have a prominent role in the suite of assessment tasks that are used to determine how well students have learned the concepts that they have been taught. As we continue to reflect on our approach to learning and assessment, we also look to improve our current practice. We should continue to work towards making our assessment process more efficient. This includes an attempt to improve the quality of the items which are integral to our assessment tasks, and an attempt to extract greater information about student achievement in a more efficient manner.
With the interest shown by governments, educators, and parents in the results of the national and international tests of mathematical literacy, which include significant numbers of MC items, it is likely that such tests will continue to be used. Generally, MC items are easy to score, and students find them simple to complete. With such efficiency and economy of use, it is likely that they will remain prominent in these tests. There are, however, many concerns with the use of MC items for the assessment of student achievement and there is a common perception that they are limited in their ability to assess student learning, particularly in mathematics. A study of the research literature on the assessment of student learning, the creation of MC items, and the analysis of responses to MC items, indicates that some of the concerns can be addressed. Addressing these concerns could result in the provision of more accurate and reliable measures of student achievement.

Compulsory national testing of numeracy was introduced into Australia in 2008, and the increased student exposure to MC items in numeracy tests prompted greater discussion of the limitations of MC items. There was also an increase in the use of MC items in textbooks of Mathematics and in classroom tests. At the time, there was little support offered to teachers for the creation of effective MC items, and to students about ways to respond to such items. Consequently, the MC items which appeared in student tests were often written by educators other than the teachers of the students, and the items were limited in how they could be used to assess student learning.

After several years of teaching mathematics to students in lower secondary, I had the opportunity to further my studies of the measurement of student achievement. During those studies I was introduced to the Rasch model for measurement and was surprised by the potential of the model, and the software that was used, to report on varying aspects of student achievement. Applying the model was an attractive aspect of the studies and I chose to collect data about the hierarchy of student development in understanding concepts associated with percentages. Some of the items in the test which was used to collect data were MC items and the prevailing attitude to the MC format was relatively negative. Many colleagues did not support the use of MC items to assess student understanding in mathematics. Such items were seen to be limited as the students were not required to show evidence of their thinking, nor of their solution processes. As a result of those studies, I became more interested in the measurement of student achievement and wanted to investigate ways by which it could be improved, and which involved the use of the Rasch model to analyse the data.
Further inspiration for the current study of MC items originated in my previous study of the hierarchy of learning percentages. While reading research reports of the hierarchies of learning fractions, decimals, and percentages, and the difficulty that students have in learning about these concepts, I realised the importance of having a good understanding of the more global concept of proportional reasoning. Additionally, while analysing the results of the study of percentages, I realised that for particular MC items, some students had selected incorrect options that were near to being correct. The selection of such options indicated that these students had better conceptual understanding of the item than the students who selected other incorrect options. It seemed logical that the students who were nearly correct deserved more credit than the students who selected the other incorrect options. The conclusion from Andrich and Styles (2011) that it is plausible to attempt to create MC items where extra credit could be given for this type of partial understanding, provided a challenge that could address this limitation of the use of MC items.

The main aim of the current study is to show that MC items can be improved: in their content and structure, and in the ways by which they are analysed. With the perceived dissatisfaction of MC items, and yet the importance of their use in student assessments, any improvement in the way the MC items could be used to assess student learning could alleviate concerns with their use. Thus, it is possible to aim at maximising their function, and allow a fairer and more accurate process of assessment to be implemented.

1.2 Assessment of mathematical understanding

MC items are prominent in national and international tests of mathematical understanding as well in many other assessments of student learning. Most of these tests have a high proportion of MC items so it is important that the MC item format is continually developed and improved for gathering information about student learning. This can increase the efficiency of the assessment process which is an integral component of the education program.

Assessment makes two major contributions to the education process. First, there is the provision of a measure, score, or ranking of student performance; and second, there is the information that is gathered about the quality of student understanding of the subject matter. The measure of the standard of performance may be used to indicate how well the content has been mastered, to determine if certain standards have been reached, and to compare performance with other students who have sat the same assessment.
Measures of performance are often interpreted as indications of success or failure, and they may be used to determine eligibility for awards or for entry to future learning opportunities. When the focus of the assessment process shifts to the quality of learning, the information which is gathered can be vital for planning subsequent learning tasks, for determining how to develop further conceptual understanding, or for addressing a misconception. With two significant contributions to the education process such as these, it is imperative that we continually seek to improve the process of assessment.

One important aspect of mathematical understanding is the ability to solve problems when the values and relationships between values are proportional in nature: more commonly referred to as proportional reasoning. The development of good proportional reasoning is expected of students before they leave secondary school but the teaching and learning of most of the skills necessary for this conceptual development are grounded in the primary years. Students in early secondary school should have developed the basic skills for manipulating fractions, decimals, ratios, and percentages for applying proportional reasoning in their later secondary years and in their adult life. However, the situation is such that many adults do not have a good grasp of the basic skills of proportional reasoning that are needed to reach the numeracy standard required in the workplace. Basic understanding of interest on investments, percentage change, and average rates are a few of the challenges for some adults. Proportional reasoning is a fascinating and important topic to investigate, and one for which there has already been considerable research but with much more needing to be done.

The demonstration of mathematical understanding is typically associated with the presentation of a task or problem to a student. After reading and interpreting the task, the student would generally create and communicate a series of steps to respond to the task or to solve the problem. This process of demonstrating the solution to a question is highly valued by teachers as it enables them to follow the student’s thinking, to gauge the student’s level of understanding, and to identify misconceptions when the solution is not reached. Such action is not feasible when MC items are used to assess student skills and understanding. However, by being creative with the structure of MC items, including an attempt to ascertain what the student knows about the concept, it is feasible to assess greater detail of student understanding than simply assess the ability to determine the correct answer to a question. Achieving this outcome would add to our insight of student understanding, improve the efficiency of the assessment process, and increase the precision of measures of student performance.
1.3 Concerns with the current use of MC items

One of the reasons why educators prefer to limit the use of MC items in assessments of mathematical understanding is the perception that they do not provide opportunities for students to demonstrate higher order thinking. Such thinking includes the ability to generalise results, justify decisions, and to create knowledge. While there are limitations to the potential for MC items to address a variety of student abilities, it is possible to design and create items that go beyond simple recall of knowledge and procedures. From research studies, it appears that it is this recall that many educators associate with MC items but there are reports of how MC items can be designed to test much higher levels of thinking.

A major concern with the use of MC items is the opportunity for students to select a correct response when they are not sure that it is correct. They have applied some level of guessing in their selection and obtained non-representative scores as a result. One issue which compounds the problem of guessing is the inability to determine the extent to which a student may have guessed their response, and hence the inability to determine how much the student knows about the concept in question. Currently, there is not one reliable process which is widely used, and which addresses this problem of the inflation of scores due to guessing.

There is a perception that males outperform females in most tests of mathematical ability and further to this, that males are more successful with the MC format. While in some assessment programs, the results of the males are always significantly better than those of females, there are other measures of mathematical achievement for which the reverse is true. The perception that males are more successful with the MC format has been linked to greater risk-taking behaviour by males who are thus believed to omit fewer MC items.

There is little information in the public domain about methods of scoring and analysing MC items. It is generally assumed that the test participant receives a score of one for a correct selection and zero otherwise, and there are no penalties for guessing. For some educators and assessment designers, the lack of scoring for any partial knowledge that a student might have is a concern. It seems illogical that a correct selection in a MC item should receive a score, and that all other selections indicate a total lack of understanding and knowledge about the item content. It appears that the measurement of student understanding is limited when any partial knowledge is not given credit and hence the measures achieved are less precise than they would be if such credit were given.
1.4 Planning for the improved function of MC items

To improve the way MC items function by attending to their construction and to the analysis of the responses, there are some relevant challenges to address. First, there is the challenge of determining the nature of higher order thinking in mathematics, how to assess such thinking within a MC item and hence the issue of what makes a MC item a good one. In addressing the second challenge, that of minimising the influence of guessing, the guessing needs to be identified before one can decide how to reduce the potential gains that the students receive. A third challenge is to confirm the existence of the perceived gender bias for MC items and to find ways to reduce this bias. The fourth challenge considered in this study is the identification of what constitutes partial knowledge and in particular, how does one recognise partial knowledge of concepts that relate to proportional reasoning.

These challenges may not be new but there is no indication that they have been addressed. In seeking to identify the features of good MC items, I tried to collect examples of items that had been used successfully in mathematics tests for students in the early secondary years. I was unable to obtain permission to study and use items from state and national testing; requests for access to item-writing frameworks and guidelines were also unsuccessful. Most item design and construction seemed to be either outsourced to private contractors or kept secure for future use. As such, there is little support with item construction for this study from current writers of MC items. It was therefore necessary to interrogate the research literature to identify the key features of good MC items and to see how the items should be structured to assess higher-order thinking, to minimise guessing, and to recognise partial knowledge.

Given the prominence of MC items in national and international tests of mathematics, one expects a considerable number of reports of earlier research studies involving the use of MC items; particularly with regard to guessing and gender bias. Perceptions by teachers that nothing is done to minimise the influence of guessing, and that males might be more capable of responding to MC items than females, imply that the results of research studies are inconclusive. The lack of information about the ability of MC items to test higher-order thinking and student partial knowledge suggests that research is this area is limited. These observations indicate that it is unlikely that the concerns with MC items, as outlined previously, can be addressed solely by the application of the results of the studies reported in the literature. However, it is expected that a study of the literature will provide some sound ideas and inspiration to guide the planning of any further investigation.
To investigate the construction of MC items which assess higher-order thinking and partial knowledge, a test of MC items was designed and conducted. For this test, proportional reasoning was chosen as the context for studying MC items because of its importance for students in early secondary. The MC items varied in difficulty and at least one of the incorrect options in each item was written to allow the students to demonstrate their partial knowledge. Other research studies were reviewed to determine how to write and score these incorrect options. This involved a review of tests that assessed skills relating to proportional reasoning and that consisted of a variety of item formats. The review also involved the analysis of student responses to these tests in an attempt to confirm the existence of partial knowledge and to identify aspects associated with the partial learning of a concept.

The construction of the test, which consisted solely of MC items, involved the formation of a framework to guide the number, content, difficulty, and organisation of the items. The test design provided alternate pathways to accommodate the different experiences and ability levels of the students by providing more difficult items for those who were capable, and easier items for those who were struggling to manage the standard curriculum. By including this differentiated form of testing it is possible to reduce the amount of guessing on the more difficult items. The analysis of the students’ responses to these items provided opportunities to reduce the amount of guessing, to compare the performances of males and females, and to examine the effect on achievement of giving credit for partial knowledge.

Some of the processes used in my earlier study of percentages are relevant for the current research. The software which was successfully implemented to create a test for the students and to gather student responses was used in this investigation. For the analysis of student responses from any external testing, as well as from my own test, Rasch Measurement Theory was applied: This theory is explained in detail in Chapter 6. This theory was chosen because possible outcomes include measures of the difficulty of the items used in the test, as well as measures of the achievement of all persons who sit the test. The type of analysis which was available can be applied when not all students respond to every item and thus it is suitable for the differentiated form of testing. When this theory was applied in my earlier study, it was possible to conduct a post-hoc analysis to reduce guessing in the MC items for the student test. This was also desirable for the current study. The software that has been written to implement Rasch Measurement Theory was also used in my previous study and it was found to be robust and reliable. The detailed output provided by the software in the previous study was also applicable and produced for the current research.
1.5 Significance of this study

As mentioned previously, the MC item is an important component of large-scale national and international tests of numeracy, the results of which are used by parents, schools, and educational authorities. Stakeholders use these results to make many decisions including those about future student learning opportunities, funding allocations for educational sectors and setting standards for student achievement: hence, the importance of the accuracy of these results. MC items are also used in many other subject areas and in tests for students of all ages. Improving the way these MC items can function to gather information about student achievement can be beneficial for many educators and students.

Addressing the issue of guessing might seem unnecessary to those who have not considered the different levels of guessing and the differential effect that guessing has for students of varying abilities. Not only do weak students benefit from guessing but the more capable students are disadvantaged when guessed responses are included in the test data (Andrich, Marais, & Humphrey, 2012). For the assessment process to be fair, this situation needs to be resolved, and this study provides a further opportunity to show how guessing can be reduced in a test consisting solely of MC items.

The ability to use MC items to detect and score student partial knowledge can increase the accuracy of the measures of achievement and also lead to fairer outcomes for the students. Currently it appears that partial knowledge is not scored, and this research provides justification for this process to be adopted in future. Identifying partial knowledge in MC items can add to the information obtained from other item formats to signal stages of learning to teachers and to provides ideas by which misconceptions can be addressed. For researchers, the identification of partial knowledge can be used to develop and confirm learning trajectories for different mathematical concepts and these in turn can be used to guide curriculum planning and development. While, in this study, the items for the student test were constructed purposefully to detect partial knowledge, the process which is used can be applied in other areas of mathematics and in other disciplines with the aim of identifying ideas to use for the construction of the options available for selection in a MC item.

Much still needs to be done to build on this work to improve MC items for the assessment of student learning and in this study, the identification of processes by which MC items can be constructed to be more effective, can enhance future development in this area.
1.6 Overview of the dissertation

Following this introduction and in the early chapters of this dissertation, the findings from a review of the literature are summarised. The focus of the review is on the improvement of the quality of assessments of mathematical understanding by attending to the creation of MC items and the analysis of responses to these items. In Chapter 2, an insight into the purpose, benefits, and importance of the assessment of student learning is presented. Details of various assessments undertaken by students in this state and country are provided and these include information about the prevalence of MC items in the various assessment programs. An outline is given of the assessment frameworks which have been designed for tests that are well known and widely used in mathematical testing programs. Consideration is also given to the different item types, including MC items, and of how they function to test mathematical understanding.

In Chapter 3 further detail of the nature of MC items, as presented in the research literature, is provided. Guidelines for creating MC items and information about the concerns with the use of MC items are included. These concerns relate to the inflation of scores due to guessing, the fact that MC items are believed to test only low levels of thinking, and the perception that males perform better than females in MC items. Other concerns include the scoring of MC items, in that marks are only allocated for correct responses and there are no opportunities to score part credit. Some studies, which report how such concerns can be addressed by attending to an improvement in the way MC items are created, or to improving the way they are scored, have been summarised in this chapter.

The next chapter in the dissertation provides background information from the literature about the various approaches to defining proportional reasoning, and about the skills and understandings that are necessary for students to develop if they are to have sound proportional reasoning. Following that section, there is a discussion in Chapter 5 of the various approaches to defining and recognising partial knowledge and this concludes with some insight as to what might constitute partial knowledge of proportional reasoning.

After the four chapters which outline the review of the literature, an explanation of the Rasch Measurement theory which has been applied in the analyses of the data in this investigation, is presented in Chapter 6. The explanation describes the mathematical basis for the model and the benefits of its application. There is also a discussion of the software used to apply this theory and of the support provided by the software for the analysis of the data and the interpretation and display of the results.
In Chapter 7, there is an overview of the analysis of secondary data sourced from national testing in Australia and from empirical studies in the United Kingdom. Rasch measurement theory was applied to the responses of students to both multiple-choice and constructed response items that tested skills relating to proportional reasoning. The purpose of this component of the study was to identify skills that could be described as partial knowledge. There were many records available in both sets of archived data and thus, any conclusions could be based on sufficient evidence. The data included both the correct and the incorrect responses from each of the students and this facilitated the identification of partial knowledge.

The construction of a test which consisted only of multiple-choice items, and which was designed and written to collect data from students, is described in Chapter 8. The discussion includes the identification of a framework for the development of the test and the construction of the MC items in order to test student partial knowledge of proportional reasoning. The processes by which the items were reviewed before the final data collection are described, along with the methods for collecting the data, and the characteristics of the students who responded to this online assessment.

The results of the analyses of the data collected from the Year 8 students who sat the online Year 8 test are presented in Chapter 9. The analyses included consideration of the accommodation for guessing, the addressing of higher-order thinking, and the allocation of part credit in MC items. The effects of these factors on the achievement of the students are outlined. The remaining section of this report discusses the findings of this investigation and their relevance for the improvement of the assessment process, particularly when MC items are used. The impact of the study for the teaching and learning of proportional reasoning in early secondary is also a consideration for the conclusion of this report.
CHAPTER 2: Assessment of mathematical achievement

Introduction

Discussions of assessment in the research literature and in the general community indicate a wide range of findings and opinions about the purposes and benefits of assessment. Assessment is variously described in the literature, and the term is used in this study to refer to the processes by which students provide evidence of their achievement and of the quality of their learning. Assessments are often identified as being one of two types, formative and summative, and it is the latter rather than the former that is addressed in this study. The focus of summative assessment is to determine levels of achievement and this is often associated with measuring or ranking performances. There are many different summative assessments of mathematical achievement undertaken by secondary school students in Australia, and these include classroom, state, national, and international tests.

To produce assessments for measuring mathematical achievement, individual items need to be created and these should be mapped to a test framework which has been designed to assess the knowledge, skills, and understandings of the relevant mathematical concepts. For school students throughout the country, the mathematics content and proficiencies, as summarised in the national curriculum documents, describe the knowledge, skills, and understandings that students are expected to learn and develop. When students are assessed they are generally responding to a prompt or a stimulus, and in mathematics their response is usually in one of two formats. First, there is the constructed response (CR) format where students create a short answer for their response. Second, there is the multiple-choice (MC) format for which the student is asked to select the correct response to the stimulus from a list of options provided by the author of the task. In many tests of mathematics for secondary students, a significant proportion of the test items are of the MC type.

2.1 Assessment

The creation of summative assessments was previously performed by teachers and a small number of experts in the field. However, the increased interest from the community, and the consequent influence on the assessment process, have resulted in the establishment of external authorities to conduct many of these assessments (Broadfoot, 2002). Increased accountability is required, standards are provided, performance targets are set, and comparisons are made. Broadfoot (2002, p. 38) stated that this influence of assessment on the education processes is
so pervasive that assessment activities now “shape the goals, the organisation, the delivery and the evaluation of education.” Lamprianou and Christie (2009), and Wiliam (2010) claimed that the effect of assessment on learning is powerful and positive, while Boud (2015) described assessment as the single most powerful influence on learning.

2.1.1 The nature of assessment

In the literature, assessment is often described as a *judgement*. Lambert and Lines (2013) refer to assessments being used to gather information to judge the effectiveness of teachers, schools, and systems. Assessment was described by Broadfoot (2012) as passing judgement on the student’s work and by Boud (2015) as judging the standard of student achievement of learning. Masters (2013) distinguished between assessment as a judgement of student success in what has been taught, and assessment as an understanding of problems associated with learning. For Masters, the current focus is on the former, but more effort needs to be afforded to assessment as understanding. While assessment might be described as a judgement of success or attainment by some researchers and stakeholders, many also would find such a summation limiting.

In an alternative approach to assessment, the processes involved, rather than the judgement of student achievement, are the focus of discussion and planning. Boud (2015) described assessment as a collection of activities from which information can be extracted and a more detailed description of each student’s progress can be determined. For Humphry and Heldsinger (2007, p. 24), assessment refers to “the process by which a teacher acquires qualitative information about a student’s progress in learning and development.” Similarly, Masters (2013) suggested that a growth perspective is needed for the assessment process and recommended that the emphasis should be on development rather than success.

The information gathered during the activities associated with the assessment process could indicate the learning achieved by the students and could also be used to compare achievement levels with international standards. Assessment data could then be used not only to see if standards are reached, but also to set the standards for minimum expected performance. Knowing about such standards can in turn provide valuable information for further research into curriculum planning, development, and delivery. The focus on development will also enable evidence about student progress in the growth of knowledge and skills to be collected. This could lower the emphasis on superficial learning or learning for the test.
For assessments to serve their purpose, they need to be reliable, valid, and authentic. Reliability describes the consistency with which identical performances on a task will always be awarded the same classification each time that task is used to assess performance (Malau-Aduli & Zimitat, 2012). Reliability is essential if an assessment is to be considered valid. Lambert and Lines (2013) noted that a task is valid if it is an accurate assessment of what it is designed to assess, and they suggested different types of validity should be considered. These types included construct validity, for which the intended attributes are appropriately tested; content validity, which is present when the assessment is a true reflection of the syllabus; and predictive validity, which exists when results are indicative of future performance.

Cumming and Maxwell (1999) reported that validity is a measure of the suitability of the task as an indicator of learning, and that this validity should be evaluated in terms of the expected educational values as well as the theories of teaching, learning, and assessment. They indicated that the authenticity of the tasks should be interpreted in terms of the tailoring of the assessment to the required learning outcomes. Cumming and Maxwell (1999) also suggested that students be offered realistic contexts and simulations of real-world events for their performances, but they acknowledged that the creation of artificial situations, and the location of assessments in context, can inadvertently contain information that will distract students from the focus of the required performance. Cumming and Maxwell recommended that the contexts for tasks be like those experienced in the school setting.

2.1.2 The purpose and outcomes of conducting assessments

According to Masters (2015, New Thinking, para. 1), there is only one fundamental purpose of assessment and that “is to establish where learners are in an aspect of their learning at the time of the assessment.” Establishing ‘where learners are’ would involve pursuing several outcomes. These outcomes could include processes to identify the nature of the learning achieved, the willingness of the learner to seek feedback on prior achievement, and the motivation and preparedness of the learner to extend their skills and understandings.

An alternative view is that the main purpose of assessment is to improve student learning (Australian Association of Mathematics Teachers [AAMT], 2017). Teachers should use the information from the assessment processes to identify those skills and concepts for which the students have good knowledge and understanding, and to identify challenging concepts and misconceptions. The information should be used by teachers to plan targeted intervention and future learning activities. Teachers should also use the information to reflect on their own
practice and to identify practices which work effectively to gather further information about student learning. Evaluating the success of current practice and support programs in schools, classrooms, and education systems is enhanced when assessment information is utilised.

There is an increasing expectation for teachers to use data from testing to improve student achievement according to Matters (2009) who recommends greater analysis even at the item level. Such an interrogation, according to Masters (2013), would support teachers in the evaluation of their learning programs and teaching strategies. Further inspection of the concepts that students have mastered, and of those which present challenges, can provide teachers with information that they can use to provide feedback to the students about their performances. Providing feedback on mathematical achievement to the students has a very powerful and positive influence on student achievement (Hattie, 2009).

In recent years, the assessment program for secondary students has evolved in intensity and quality and there has been a renewed interest in the less formal outcomes. Some of these have been identified by Broadfoot and Black (2004) as raising the level of achievement, bringing a renewed emphasis on curriculum development, setting standards for schools, and facilitating national and international comparisons of performance. Information on student learning can be used to report achievement over time and to compare achievement with international standards (Masters, 2013). Assessment data can also be used to set the standards for minimum expected performance as well as check that standards are reached.

For secondary students, measures or standards of achievement may be reported in bands, levels, standards, descriptors, proficiency scales, and measurement scales, as well as the traditional percentages and grades. The range of student achievement on an assessment task can be represented by a proficiency scale, described as a Learning Metric by Turner (2015). A Learning Metric has two main elements: a numeric measure of proficiency located on a scale and a proficiency description associated with locations on the same scale. The numeric measure is determined from marks generated in the traditional manner and the numeric scale formed is further divided into sections (levels or bands). Proficiencies are then described for each section. These descriptions provide details of the knowledge, skills, and understandings likely to have been demonstrated by students with scores in that section. Students would have a high probability of success when tested on the skills at their own level and an even greater probability of success on the skills at the lower levels of the scale.
The number of countries participating in national and international assessments has increased considerably since the 1990s, and in an extensive review of the literature on assessment, Baird, Hopfenbeck, Newton, Stobart, and Steen-Utheim (2014) reported increased competition between countries to produce large numbers of workers with high levels of knowledge. The aim of one project was to support nations to make international comparisons, to investigate their need for educational reform, and to develop their knowledge as a “key strategic resource” (Kellaghan, 2001, p. 95). National competitions and assessments provide data on student performance to students, parents, schools, and sometimes the whole community. Ideally, one use of such data should be for schools to reflect on their own individual performance. The availability of such data has the potential to fuel competition between schools, but it would be timely to note the finding that, in most countries, less than 10% of the variance in student scores was due to “between-school differences” (Wiliam 2010, p. 107). The variance is more likely to be due to other factors, especially prior achievement.

2.2 Types of assessment

There is general agreement within the education community that there are two main types of assessment, and several authors including Baird et al. (2014), Black and Wiliam (1998a, 1998b), Boud (2015), Broadfoot (1996), Lambert and Lines (2013), Masters (2014a), and Wiliam (2010) have reported on these types. The first type is referred to as summative assessment or assessment of learning. It typically occurs at the end of a course or period of learning and the student receives an overall score or grade. The second type is referred to as formative assessment or assessment for learning. Here, feedback is provided immediately and more often to the student about aspects of their performance on individual components of the assessment, thus allowing students to learn about their strengths and weaknesses and to improve their performances. A third type of assessment is called assessment as learning which refers to the processes by which students reflect on their own learning and assess their own performances (Suurtamm et al., 2016). In this type of assessment, the students are more actively involved in the assessment process, but they still need the guidance of the teacher to learn from the activity and to obtain the benefits of the self-reflection.

With an apparent emphasis on summative assessment, some have argued that formative assessment is underdeveloped, and it is difficult to find it working well in schools (Lambert & Lines, 2013). Black and Wiliam (1998a, 1998b) reported that in providing feedback there are substantial benefits for learning, but much can be done to improve the quality of the feedback.
In a later study, Black and Wiliam (2009) suggested that feedback is often ineffective because teachers have little time to plan it and may not be able to tailor it to individuals because they do not understand the students’ thinking processes. Another problem with feedback is that it is often misunderstood or poorly used by the students, and it can be demoralising and de-motivating (Broadfoot, 2002). For further improvement, Boud (2015) suggested that a more active involvement of students in the assessment and feedback process will help them to develop greater capacity to make judgements about their own performance.

All teaching involves a form of assessment for learning. The teacher sets the tasks, the student responds, and the teacher provides feedback on those responses. In this way, the teaching is tailored to the responses of the student and this facilitates reinforcement of learning and the clarification of misconceptions. This process of formative assessment has been described as “indistinguishable from teaching” when it is working well in the classroom (Lambert & Lines, 2013, p. 193). Most research studies report on giving feedback to students on errors and misconceptions. Few studies refer to the influence on learning of providing positive feedback on good results: such influence providing positive outcomes for future learning.

Assessments are defined as norm-referenced when comparisons between achievements are the focus of the output, and as criterion-referenced when outputs are described in terms of the concepts that have been developed. Commenting on the various types of assessment, Masters (2014a) concluded that the field of research into assessment is overcomplicated by the number of words being used to describe assessment and it will be beneficial to focus on the main aim of assessment which is to “establish what they (the students) know, understand and can do” (para. 3).

Several factors influence the type of assessment chosen and many types may be needed to determine progress and achievement. Assessment tasks should be designed, created, and implemented to be suitable for identifying student ability in the relevant learning domain. Masters (2014b, para. 5) claimed that a “significant proportion of what students are expected to learn can be assessed efficiently and reliably with a common externally developed test.” In comparison, Boud (2015) reminded educators that a single activity cannot capture all the information required. Baird et al. (2014, p. 9) claimed it is a challenge to extrapolate from “standardised assessments the rich, authentic, robust understandings they are supposed to represent,” but one could argue that we should continue our efforts to improve the processes of assessment because of the advantages this will have for improving student learning.
2.3 Assessments for secondary students

Many of the assessments, which are used to measure the mathematical achievement of secondary school students in Western Australia contain MC items. International assessments include the Programme for International Student Assessment (PISA) tests and the Trends in International Mathematics and Science Study (TIMSS) tests. For these tests, only a sample of students is chosen. At the national level, all Australian students in Years 3, 5, 7, and 9 participate in the annual National Assessment Program-Literacy and Numeracy (NAPLAN). To graduate from secondary school, students who do not reach Band 8 in the Year 9 NAPLAN numeracy test are required to demonstrate a satisfactory standard in the Online Literacy and Numeracy Assessment (OLNA) test (School Curriculum and Standards Authority [SCSA], 2017). Further voluntary competitions that are used to assess mathematical achievement include the Have Sum Fun competition (Mathematical Association of Western Australia [MAWA]), the Australian Mathematics Competition (AMC) and the Computational and Algorithmic Thinking (CAT) competition.

WACE

Students leaving secondary school are expected to achieve the Western Australian Certificate of Education (WACE) for which it is not compulsory to study mathematics. Of the different mathematics courses available in the final year, students can choose the most appropriate for their level of achievement and planned future study. There are currently no multiple-choice questions in any WACE examinations for the mathematics courses, but MC items exist in examinations for other courses and for mathematics in other states.

PISA

Every three years, a nationally representative sample of students is selected for PISA testing. The students are about 15 years old and nearing the end of the compulsory years of schooling. The PISA test includes an assessment of mathematical literacy which is described as “the individual’s capacity to formulate, employ and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain, and predict phenomena” (Organisation for Economic Cooperation and Development [OECD], 2017, p. 66). This is not a test of the intended curriculum but a test of student readiness to solve the types of problems that students are likely to encounter beyond school. Achievement is reported in six levels of proficiency and the data showing the proportion of students at each level are provided to participating countries.
TIMSS

For the TIMSS assessment, trends in mathematics achievement are monitored through testing a sample of Year 4 and Year 8 students every four years and the average age of the Year 8 students sampled is 13.5 years (Martin, Mullis, & Foy, 2013). For these students, the test is based on the intended curriculum for the year group, and the content or subject matter is described in terms of number, algebra, geometry, and data and chance (Sissel Grønmo, Lindquist, Arora, & Mullis, 2013). Participation in the two large-scale international tests, PISA and TIMSS, provides significant opportunities for the comparison of the achievement of Australian students with international students, and this means identifying areas where achievement is behind or ahead of students in other countries. Reports showing achievement based on gender, socio-economic status, and location are available; and the information given indicates where research might be conducted to investigate the factors which could lead to an individual country’s improvement (Thomson, 2010).

NAPLAN

Each year, Australian students in Years 3, 5, 7, and 9 participate in a national assessment program that includes tests of numeracy. For students in Years 7 and 9 there are two sections of the numeracy test: a calculator section, and a non-calculator one. Each student’s results are available to their parents, teachers, schools, and education authorities. The data can be used to determine the extent to which students are reaching minimum benchmarks and to identify gaps in student learning. Schools are informed of the percentages of students at the various levels (bands) for their own school, for similar schools, for the state, and the nation. NAPLAN testing for numeracy began in 2008 and longitudinal data is available to measure student and school performance over time. The skills and understandings assessed in the NAPLAN program “broadly reflect important aspects of literacy and numeracy in the Australian Curriculum adopted by each state or territory” (National Assessment Program, para. 3, 2016). Failure to demonstrate minimum standards in the NAPLAN testing program can signal where intervention and targeted support can be provided.
OLNA

Year 10 students needing to demonstrate the minimum level of numeracy, first sit the OLNA test in Semester One and then again at six-monthly intervals until they reach the required standard. The minimum level of numeracy that students need to demonstrate is described as those “skills regarded as essential to meet the demands of everyday life and work in a knowledge-based economy” (SCSA, 2017, para. 1). To demonstrate the minimum standard, students need to be competent with the skills described for Level 3 of the Australian Core Skills Framework (ACSF). For numeracy competence at Level 3 students are expected to work mostly in familiar contexts and occasionally with those less familiar. Students are required to select and use a variety of procedures to perform calculations and to solve problems, and to interpret and use formal mathematical language and symbols. Examples of skills that students should be able to demonstrate include: (a) calculate perimeter and area of rectangular shapes, (b) determine costs per kilogram, (c) determine the percentage of an amount using a simple fraction with a denominator which goes evenly into the amount, and (d) describe likelihood using fractions and familiar terms such as certain or impossible (Australian Government: Department of Education and Training, 2012).

OPTIONAL ASSESSMENTS

The Australian Mathematics Competition (AMC) and the Computational and Algorithmic Thinking (CAT) assessment are annual and voluntary competitions which are organised by the Australian Mathematics Trust (2018a, 2018b). The AMC competition has three papers for secondary students: Junior, Intermediate and Senior. The AMC items are designed to test mathematical thinking using content from the expected curriculum. The aims of the competition are to promote mathematics, to provide resources for further activities and to identify mathematical talent. Talent identification is also an aim of the CAT competition, which seeks to identify students who show ability with the type of logical thinking that would be expected of software developers, that is, the ability to identify and apply algorithms.

Specific Mathematics Assessments that Reveal Thinking (SMART) have been created by Stacey, Price, Gvozdenko, and Steinle (2013) and are available to schools throughout Australia. They are particularly suited to students in the late primary and early secondary years. These online tests are short, and many items allow the students to choose their answers. As well as the traditional MC questions, students can drag a graphic into a pre-defined position or move a slider to select a particular value. Descriptive feedback of individual
student performance is immediately available, and this provides diagnostic information to teachers for them to modify their teaching, and to plan learning experiences to help students overcome their difficulties. Suggestions for teaching the content are also provided and examples are given in Appendix 2.1. The items are designed to investigate the way students think and the data provided have been used for research into the types of errors and misconceptions that are common to students at different stages of development.

2.4 Developing an assessment framework

Test frameworks are used to guide the creation and organisation of items which comprise the test instruments for student assessment. They provide a mechanism for checking that the relevant content and skills are assessed in a balanced and unbiased manner. The identification of an appropriate guiding framework for an assessment assists in the conceptualisation of the learning domain and the operationalisation of the learning objectives; achieving these outcomes enhances the validity of the scores (Kind, 2013a).

Four key principles for the development of a framework to assess the attainment of the educational objectives were identified by Bloom, Engelhart, Furst, Hill, & Krathwohl (1956) during the creation of the original Bloom’s taxonomy. The first principle indicates that the framework design should reflect the distinctions that teachers make between student behaviours. The second principle indicates that the subdivisions of the framework should be logically developed, and the use of terminology should be consistent throughout. According to the third principle, the framework is designed to address only educational behaviours, and for the fourth principle the system should accommodate any objective that describes an educational goal.

Before a test framework can be designed and created, it is necessary to identify (a) the guidelines for the framework which would make it suitable for the nature of the assessment to be implemented, (b) the curriculum content to be assessed, (c) the students who would sit the test, and (d) the expected behaviours of the students who respond to the test items. Some of the frameworks described in the literature, and some of those in current use, were examined for the purpose of identifying guidelines for the development of a framework for the proposed data collection. This also provided further insight for the creation of the individual items as well as the construction of the actual test framework.
2.4.1 TIMSS framework

The TIMSS framework is described in two dimensions or domains: content and cognitive. Content refers to the subject matter, for example, fractions, decimals, ratio, proportion, and percentages. The cognitive domain describes the thinking processes which are further classified as knowing, applying, and reasoning (Sissel Grønmo et al., 2013).

Items described as knowing make up approximately 35% of the items and these items test student ability to recall facts, concepts, procedures, language, number types, and conventions. These items also test student ability to recognise symbols and equivalence, and the ability to use algorithms and read graphs. About 40% of the items are described as applying and, in these items, students are required to use familiar procedures, to create equivalent representations, or to apply their knowledge and understanding in a range of contexts. In these items, tasks should be familiar and routine. The remaining items are classified as reasoning, and students may be required to think systematically and logically and to use their knowledge to solve problems in new or unfamiliar situations.

2.4.2 PISA Framework

The framework for the PISA test of mathematical literacy, provided by the Organisation for Economic Cooperation and Development (OECD, 2017), specifies three dimensions: content, processes, and context. Mathematical content includes quantity, space and shape, change and relationships, and uncertainty. Processes refer to the mathematical skills and procedures necessary for solving problems, while context describes the situations to which the content, skills, and procedures are applied.

In applying the framework, items are allocated to one of three processes: formulate, employ, and interpret. About 25% of the items require students to formulate a situation or to recognise the mathematics required to determine the solution. The ability to use and apply mathematical knowledge and skill is described as employing and about half the items are allocated to this category. The remaining items focus on the students’ abilities to interpret and evaluate the results and conclusions in the context provided. Within each of the processes, there are seven mathematical capabilities that students might use to respond to an item, and any one item may require more than one of these capabilities. Consideration of both processes and capabilities for the items produces a two-dimensional framework that supports the planning of a balanced selection of items for the test of mathematical literacy.
The framework for the PISA test of mathematical literacy is given in Figure 2.1. Items are developed within one of four contexts which are identified as personal, societal, occupational, and scientific. Throughout the test, the items range in difficulty, and the proportion of items classified as belonging to each of the four contexts is approximately the same.

<table>
<thead>
<tr>
<th>Mathematical capabilities</th>
<th>Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>formulate</td>
</tr>
<tr>
<td>Communicating</td>
<td>25%</td>
</tr>
<tr>
<td>Mathematising (transformation of problem to mathematical form)</td>
<td></td>
</tr>
<tr>
<td>Representation</td>
<td></td>
</tr>
<tr>
<td>Reasoning and argument</td>
<td></td>
</tr>
<tr>
<td>Devising strategies for solving problems</td>
<td></td>
</tr>
<tr>
<td>Using symbolic, formal and technical language and operations</td>
<td></td>
</tr>
<tr>
<td>Using mathematical tools (calculators, computers)</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2.1* Framework design for PISA test of mathematical literacy

2.4.3 National testing in the United States of America

Every two years in the United States of America, the National Assessment of Educational Progress (NAEP) tests of mathematical knowledge and skills are given to a representative sample of students in grades 4, 8, and 12. The development of the test framework is overseen by the National Assessment Governing Board (2014) and one dimension of the framework is content, which covers topics including number. For Grade 8 students, the items test knowledge and skills in decimals, fractions, percentages, proportionality, and rates.

The framework lists three categories of complexity: low, medium, and high. Items of low complexity may require mechanical processing, recall or recognition of knowledge and procedures; they make up 25% of all items. For the 50% of items deemed to be of medium complexity, students may need to decide which calculations to use or they may be required to make connections between two or more mathematical ideas: typically, two-step operations. At the highest level of complexity, students may need to apply reason, analyse mathematical arguments, justify conclusions, or form generalisations. With three levels of complexity included in the design, mathematical thinking over a wide range of ability can be tested.
2.4.4 National Testing in the United Kingdom

The framework for the assessment of the national curriculum in the United Kingdom outlines the “purpose, format, content and cognitive domains” of the test (Standards and Testing Agency, 2015, p. 4). The detailed content, which is specified in the framework, aligns with the content of the national curriculum and, for students at the end of Key Stage 2 (Year 6), this includes fractions, decimals, percentages, proportions, and ratios. The skills associated with these content areas predominate in the section on number, ratio, and algebra to which the percentage of marks allocated is between 75% and 85%. Each test question is identified with one of the four strands of the cognitive domain and is allocated a level of difficulty. This alignment with the cognitive domain enables the skills and intellectual processes to be identified, and it shows where the national curriculum aims of problem-solving, fluency, and reasoning are addressed in the assessment. Items represent a sample of curriculum content and are placed in the test in perceived order of difficulty. Part of the test framework showing the classification of the cognitive domain and the levels of difficulty is provided in Figure 2.2.

<table>
<thead>
<tr>
<th>Cognitive domain</th>
<th>Rating scale for difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>low (1)</td>
</tr>
<tr>
<td>Depth of understanding</td>
<td>recall facts</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational complexity</td>
<td>no steps</td>
</tr>
<tr>
<td>Spatial reasoning, Data interpretation</td>
<td></td>
</tr>
<tr>
<td>Constructed response ...</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2.2* Framework (in part) for national testing in the United Kingdom
2.4.5 SOLO Taxonomy

In describing the quality of student learning, Biggs (1979) provided insight for the creation of assessment items without including the creation of a framework for writing assessments. By applying the Structure of the Observed Learning Outcome (SOLO) taxonomy, student responses were classified into one of five hierarchical levels which are, in order; pre-structural, uni-structural, multi-structural, relational, and extended abstract. A pre-structural response shows no logical relationship to the task, whereas in a uni-structural response, the student will respond with only one relevant action and ignore other relevant information which is provided. A response is classified as being multi-structural when the student identifies, but does not analyse or synthesise, several relevant aspects of the response. When the aspects are synthesised, and the student makes accurate conclusions, the response is deemed to be relational. At the highest level, a response is classified as extended abstract when student behaviour involves challenging the assumptions made, providing counter examples, and generating new data.

Applying the SOLO taxonomy to mathematical responses is difficult (Biggs & Collis, 1982). One reason for this difficulty is that in solving mathematical tasks there are often many ways by which a student can determine a correct answer, and these may even include incorrect thinking. Further probing would be necessary before identifying the category to which the response should be assigned. Biggs and Collis (1982) also indicated that it would not be easy to write a multiple-choice item where the different options represent the range of levels of the taxonomy. Using SOLO taxonomy to classify student behaviour would be challenging as the students would be selecting actions that others have provided, rather than performing the action themselves. In setting up a SOLO task, teachers need to know which behaviours they seek to assess and to have a clear understanding of how to classify the students’ responses.

2.4.6 Bloom’s Taxonomy

Although Bloom’s taxonomy was designed to “be a classification of the students’ behaviours which represent the intended outcomes of the educational process” (Bloom et al., 1956, p. 12), it can be adapted and used as a test framework. In the taxonomy two domains are used to classify behaviours: the cognitive domain and the affective domain. Within the cognitive domain, student responses to test questions could be identified as typical of one of six categories: knowledge, comprehension, application, analysis, synthesis, and evaluation. For a more detailed overview of the cognitive domain see Appendix 2.2. Using the taxonomy
enables test questions to be placed into one or more of the six categories. However, before the item is assigned to a particular aspect of mathematical behaviour, further information is needed. One needs either to know, or to make assumptions about, the teaching and learning conditions, the thinking processes needed to solve a problem, and the possibility that the student may select or determine an answer using incorrect reasoning.

In elaborating on the further use of the taxonomy, Bloom, Hastings, and Madaus (1971) outlined two ways by which mathematics test items can be described. First, items can be described in terms of the content or subject matter, for example whole numbers or integers. Second, items can be described in terms of the cognitive aspects of student behaviour required to provide a correct response to each item. Student behaviour is classified as belonging to one of four cognitive aspects: computation, comprehension, application, and analysis. Further classifications for two of these four aspects are shown in Figure 2.3 where computation is classified as either knowing facts, knowing terms, or using algorithms. There are more classifications for comprehension than the two that are shown in Figure 2.3, and the two other behaviours, application and analysis, are also not shown in this figure.

<table>
<thead>
<tr>
<th>Content</th>
<th>Cognitive aspects of student behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Computation</td>
</tr>
<tr>
<td></td>
<td>Comprehension</td>
</tr>
<tr>
<td></td>
<td>know facts</td>
</tr>
<tr>
<td></td>
<td>know terms</td>
</tr>
<tr>
<td></td>
<td>use algorithms</td>
</tr>
<tr>
<td></td>
<td>know concepts</td>
</tr>
<tr>
<td></td>
<td>know principles</td>
</tr>
<tr>
<td>Whole numbers</td>
<td></td>
</tr>
<tr>
<td>Integers</td>
<td></td>
</tr>
<tr>
<td>Rational numbers</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2.3* Part classification of content and behaviours in mathematics

The four types of cognitive aspects are considered hierarchical (Bloom, Hastings, & Madaus, 1971) with analysis considered of higher order than application and this in turn was deemed to be more demanding than comprehension and computation: the latter the least demanding of all. Computation involves the simple recall of facts, terminology and procedures with no decision making needed. Comprehension includes the recall of concepts, mathematical structure, and generalisations, as well as the transformation of mathematical elements, whereas for application, students select and use mathematical operations in familiar settings. The key feature of being successful in analysis is dealing competently with original and non-routine problems and relationships.
In the revised taxonomy for classifying test items, there are still six main classifications of behaviour and these are referred to as *remember, understand, apply, analyse, evaluate, and create* (Anderson et al, 2001). *Synthesis* in the original taxonomy has been replaced by *create* and is deemed to be at a higher level than evaluating because evaluation is needed for the creation of new material. Evaluation involves making “judgements based on criteria and standards” (Anderson et al, 2001, p. 31), whereas creating involves making a new and original product. The four aspects of the knowledge dimension were reconceptualised as facts, concepts, procedures, and metacognition. An overview of the two-dimensional matrix with the knowledge and cognitive aspects of the revised taxonomy is presented in Figure 2.4.

From left to right as represented in Figure 2.4, the intellectual demands should increase thus: to *create* should require greater ability than to *understand*. For the knowledge dimension, the lower the type of knowledge is located in the table, the more difficult it probably is to use: Remembering procedures is expected to be more challenging than remembering facts. A test consisting only of items requiring the recall of factual knowledge would tend to be a low-level assessment, and the addition of items requiring recall of conceptual knowledge would be expected to increase the intellectual demands of the assessment.

<table>
<thead>
<tr>
<th>Knowledge Dimension</th>
<th>Cognitive Process Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>remember</td>
</tr>
<tr>
<td>facts</td>
<td></td>
</tr>
<tr>
<td>concepts</td>
<td></td>
</tr>
<tr>
<td>procedures</td>
<td></td>
</tr>
<tr>
<td>metacognition</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2.4* Overview of Revised Bloom’s Taxonomy

More skilled individuals can understand and apply (transform) their knowledge of facts, concepts, and procedures: they are aware of the criteria used to determine how and when to apply this knowledge. They can also execute or implement these skills to solve routine problems, recognise patterns, or analyse data in situations that may not be exactly familiar, but will be like those previously experienced. In the revised taxonomy, understanding is considered a *lower* skill than applying; this indicates that the comprehension of a concept or procedure generally precedes the application. However, one can argue that some procedures can still be applied when comprehension of the procedure is lacking, for example, students will often learn that *to multiply by 10, add a zero.*
2.4.7 Designing a test framework

Having a rationale for a test framework is preferable to just listing the dimensions. A rationale is essential for maintaining construct validity when developing items and interpreting the results (Kind, 2013a). In an analysis of the development of the PISA, TIMSS, and NAEP frameworks for assessing learning in science, Kind (2013a) identified five dimensions used to create frameworks: attitudes, context, knowledge, conceptual knowledge, and behaviour. For the development of a framework these dimensions should be considered along with the content to be assessed and the current educational expectation of the target audience.

For this research into the function of MC items, it was not essential to collect information about students’ attitudes to Mathematics because it would not influence the analysis of the effect of changing the way MC items were written and scored. Hence, for this study, student attitude was not used as a dimension for the development of the test framework. The contexts used for writing the items were familiar for the students and not so overwhelming as to interfere with the students’ abilities to access the mathematics of the situations. The average age of the students was about 14 years so the contexts for the items included home life, finance, school environment, sport, and social situations. Knowledge and conceptual knowledge were combined to form a single dimension for this research and were identified as the content descriptions of the Year 8 curriculum that the participants have studied.

Of all the dimensions listed, the most difficult to define, and the one showing the most varied meaning between the various frameworks is behaviour. Kind (2013a) noted that knowledge and behaviour are inextricably linked and, while the subject area reported was Science, the same linkage is expected to apply to Mathematics. Framework design needs to reflect this fact by acknowledging that the testing of mathematical behaviour requires the use of mathematical knowledge, and similarly, testing mathematical knowledge requires the students to use some form of mathematical behaviour. Greater cognitive requirement, and hence more demanding cognitive behaviour, is associated with more complex knowledge, and for this reason Kind (2013b) recommends that the explicit definition of knowledge categories, for example, ratios and fractions, is an essential part of any framework. The identification of the learning progression within these categories, can assist with the recognition of levels of student behaviour. The mathematical behaviour described by Kind can be likened to the proficiencies that students are expected to develop in their learning of the mathematical content as described in the national curriculum and as summarised in the next section.
2.5 Mathematical achievement

One aim of conducting an assessment is to determine the students’ achievements. This involves identifying (a) what the students know, (b) what the students understand, and (c) what actions the students can perform. In summary, mathematical achievement is considered in terms of the knowledge, skills, and conceptual understandings that the students have developed from their learning, which is described as the “acquisition, transformation and evaluation of information” (Bruner, 1960, p. 48).

The acquisition of knowledge, or what students know, is described by Skemp (1971), Vergnaud (1990), and Piaget (1964) as an active process occurring over time. Skemp concluded that it is difficult to acquire knowledge without an existing schema, which can consist of complex conceptual structures as well as simple sensory-motor ones. The schema acts as a mental tool for supporting the integration of what is being learned with the existing knowledge. Vergnaud (2009) identified two forms of knowledge: the operational, which consists of physical and social actions; and the predicative, in which these actions are expressed in language and symbols. The acquisition of knowledge results in the ability to recall facts, terminology, procedures, concepts, principles, and mathematical structure from long-term memory (Bloom et al., 1956). Knowledge acquisition also enables students to recognise facts, terminology, concepts, and procedures because the information is familiar.

To determine what students can understand, it is necessary to see how they act on, or re-organise what they know. It is also necessary to see how they make connections with related knowledge, and how they integrate their knowledge and skills (Wong, 2009). What students understand has been the subject of much research, and a model of growth in understanding has been developed and explained by Pirie and Kieren (1994), Pirie and Martin (2000), and Martin and LaCroix (2008). These authors describe the process by which understanding develops and is demonstrated, as dynamic. Rather than being a product of associated actions, the development of understanding requires the learner to continually reflect on what he or she knows and to use this information to construct new images. It is a “process of constantly changing recursive growth involving shifting and interplay between specific actions and general and more abstract conceptualisations” (Martin & LaCroix, 2008, p. 121).
Curriculum overview

Mathematics for lower secondary students is generally organised in curriculum documents around two main components: content, and procedures. Content or mathematical knowledge refers to the topics to be taught, for example, Number and Algebra; these tend to be named along traditional lines. Procedures refer to the ways by which students might demonstrate their knowledge, skills, and understandings when they engage in the mathematical content. The three content areas in the Australian curriculum are number and algebra, measurement and geometry, and statistics and probability, while the procedures are called *proficiencies* (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2015a).

Kilpatrick, Swafford, and Findell (2001) conducted an extensive examination of research studies and reported the five aspects of mathematical proficiency which teachers and students should aspire to obtain if learning is to be successful. These are (a) conceptual understanding, (b) procedural fluency, (c) strategic competence, (d) adaptive reasoning, and (e) productive disposition. The first four of these aspects have been adopted as proficiencies for the Australian curriculum and renamed as *understanding, fluency, problem solving, and reasoning* (ACARA, 2015a). Students demonstrate understanding when they can connect related ideas and apply familiar concepts to generate new ones. They are fluent when they can choose and use appropriate procedures accurately and flexibly. Students solve problems when they investigate unfamiliar situations, decide which mathematics to use, and can generalise their results. When reasoning, students explain their thinking, justify their decisions, and make deductions: this may involve sophisticated and logical thinking.

To assess both student understanding of the content of the mathematics curriculum, and student proficiency in applying this mathematical knowledge, the creation of suitable items is important for the realisation of valid and reliable measures. The adoption of suitable items is also important in the identification of the quality of learning and of the processes necessary to improve this learning.
2.6 Item types

There are three structural types of items used to test mathematical competence. There are first the open-ended items, which usually consist of a single task requiring several processes and these may involve investigation, generalisation, and testing. Open-ended items are rarely used in large scale summative assessment of mathematical understanding because the students who misinterpret the question are unable to demonstrate their knowledge and skills and they do not get any credit for what they might know about the topic.

A second type of item, sometimes called short-answer questions and hereafter referred to as constructed-response (CR) items, requires the students to create and provide a short response to a stimulus or question. These items may contain several parts and they may require students to show correct mathematical behaviours in the process of determining the solution. These CR items are prominent in many assessments of mathematical competence including tests, examinations, and investigations.

The third type of item refers to the multiple-choice items which typically contain a stem (introductory text) and a list of options from which the students choose the best answer to accurately complete the stem or to answer the question posed in the stem. MC items are a significant feature of many of the assessments undertaken by secondary students, including the PISA and TIMSS tests. The framework for the 2015 PISA test indicated that there were three different types of items, one of which was MC, and that there were equal numbers of each type of item (OECD, 2017). Information on the exact numbers of each type is not provided for the TIMSS assessment but at least half the points in the total score come from MC items (Martin et al., 2013).

For the numeracy section of the OLNA test, each of the 45 items are of the MC type (SCSA, 2017). In the two national competitions (AMC and CAT) the proportion of MC items is constant each year; this proportion is given in Table 2.1. The proportion of MC items in the NAPLAN numeracy assessments is high in all tests but it is not constant each year or for each cohort. For the other NAPLAN items, the CR type, the students only need to enter a single numeric value. The extent to which multiple-choice items are used in assessments taken by lower secondary students in Western Australia is summarised in Table 2.1. The data show that multiple-choice items are prominent in these assessments.
Table 2.1  MC items in student assessments

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Cohort</th>
<th>Number of multiple-choice items</th>
<th>Number of items which are not multiple choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAPLAN</td>
<td>Year 9, 2013</td>
<td>46</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Year 9, 2014</td>
<td>50</td>
<td>14</td>
</tr>
<tr>
<td>OLNA</td>
<td>Semester One</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Year 10, 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIMSS</td>
<td>Year 8</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Sample questions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMC</td>
<td>Intermediate years</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>CAT</td>
<td>Every year</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>SMART</td>
<td>Year 8</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>% Type Quiz</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In many of the large-scale assessments of mathematical competence for secondary students throughout Australia, the proportion of the items which are in MC format is high. Statistics to describe the prevalence of MC items used in classroom tests are not available but it is generally known that CR items predominate in such assessments, and at times, a few MC items are also used in these tests to provide practice and preparation for the large-scale assessments. Given the dependence on MC items to assess student achievement in mathematics, as well as in other subject areas, it is essential that a continued effort is made to improve the quality of the items and to improve the standard of the information that they provide about student achievement and conceptual development.
CHAPTER 3: Multiple-choice items

Introduction

In most international, national, and state assessments of mathematical achievement the items are either of the constructed response (CR) or the multiple-choice (MC) type. Many of the assessments contain both types of items and, as described in the previous section, there are usually more MC items than CR items. With both types together, the students can demonstrate a wider range of skills and cognitive abilities than with one type alone. MC items contain a stimulus and a list of options, and the student is asked to select one or more options, though usually only one, to match the stimulus. There are typically three or four incorrect options, and these are often referred to in the literature as distractors. Selection of the correct option, the key, is typically awarded one mark and zero is given for the distractors and for missing responses. MC items are widely used in tests of mathematical understanding because they are economical to use, supposedly attractive to students, and easy to implement. The creation and development of such items requires considerable time and expertise as does the analysis of the students’ responses to these items. There are several concerns about the use of MC items in mathematics tests including the propensity of students to guess the response, the belief that only low levels of thinking are assessed with such items, the lack of partial credit, and a perceived gender bias. There are many references in the literature to suggest how such concerns can be addressed in order to improve the function of MC items and the quality of the information that they generate about levels of student achievement.

3.1 Types of items: MC and CR

The description given in the TIMSS 2015 framework provides a comprehensive overview and indication of the differences between the two types:

Multiple-choice items allow valid, reliable, and economical measurement of a wide range of content in a relatively short testing time. However, …, these items may be less suitable for assessing students’ ability to make more complex interpretations or evaluations.

… Because they (CR) allow students to provide explanations, support an answer with reasons or numerical evidence, draw diagrams, or display data, constructed-response items are particularly well-suited for assessing aspects of knowledge and skills that require students to explain phenomena or interpret data …

(Martin, Mullis, & Foy, 2013, pp. 92–93)
The effects on achievement of using these two types of items with similar stems (introductory text) were compared by Behuniak, Rogers, and Dirir (1996); Bonner (2013); Hohensinn and Kubinger (2011); O’Neil and Brown (1998); Rodriguez (2003), and Thawabieh (2016). In these research studies, CR items were generally found to be more difficult for students than the stem-equivalent MC items but the effects on achievement were inconsistent. Rodriguez (2003) found that the correlation in achievement was only high when the stems were equivalent, or when the both types of items tested similar content. Biria and Dehghan (2016) noted the lack of significant difference in the achievement of young adults when comparing item format but their results related to reading comprehension with a small group (n=44) where the response was more open-ended.

The influence of format on the variety of strategies used was also found to be inconsistent. O’Neil and Brown (1998) reported more cognitive strategies used, less self-checking done, and more students worrying about their test preparation with the CR items. Bonner (2013) also noted that more information about cognition is likely to be available when using CR items but found an increased variety of strategies, including guessing and working backwards, were used with the MC format. Bonner concluded that it was easier for low-ability students to identify the nature of the problem when using MC items and that the inclusion of these items in assessments is adequate for ranking students.

The common belief that MC items are much easier than items in which a student needs to construct a response is associated with the perception that these MC items only assess lower-order thinking. In a review of research into MC and CR items, Martinez (1999) concluded that MC items address only low-level processing and CR items tend to evoke more complex thinking. This occurred because the MC items were mostly written to test lower levels of cognition and was not due to the format of the item. Martinez suggested that the item types are more likely to measure equivalent traits at the higher levels of cognition where complex processing is required for both types of items. While there is evidence that CR items can assess a wider range of cognitive behaviours, and hence are more useful for diagnosing cognitive states, an analysis of current trends in item writing might show that the difference in the levels of thinking that both types of items can measure has diminished in recent years.

In a study of school leavers’ performance in Chemistry, Hudson (2010) found that the students performed better in the MC items than in the CR items regardless of the content covered. The greater success in the MC items occurred mostly in the recall type of item, and
for the more difficult items involving the application of ideas, there was little difference in student performance between the two item types. The students who were interviewed indicated a slight preference for having MC types of items because they were able to obtain clues from the items, work backwards from the options, or guess the answers. Students preferring CR indicated that they liked the idea of being able to gain some marks for partly correct answers. Hudson indicated further research was needed to confirm these findings.

In considering the amount of marking required for law students’ assignments, Huang (2016) conducted a study to identify a role for a test of MC questions in the assessment program. Student performance in a multiple-choice questionnaire (MCQ) was compared to their performance in a written assignment. The students indicated a strong preference for the assignment format, and Huang suggested that the students preferred to construct text because they believed that it was important to learn to write legal documents, a skill not assessable with MCQ. Although the MCQ tested different content areas and validity could not be established, the distribution of MCQ marks mirrored the assignment marks. Huang concluded that the MCQ can be used to assess a wider range of the course material, and it is as good as the written assignment for reporting performance, but it should not be the only type of assessment used in the course.

The conventional scoring for MC items is to give one credit point for the selection of the key and zero otherwise. To determine an alternative, and possibly more effective scoring method, Adebule and Awodele (2016) compared conventional scoring with liberal marking and confidence scoring in a chemistry test for secondary students. Liberal marking involved allocating a different score for each option according to the degree of correctness. For confidence scoring, points were awarded when students indicated their level of confidence in having selected the correct response. The results of comparing the three scoring methods indicated significant differences in the reliability though not the validity of the different methods. More research is needed to test the reliability and relevance of these findings across a wider range of populations and courses. These scoring methods are usually associated with guessing and partial credit, and these ideas are discussed later in this chapter.

For any assessment of mathematical skills and understandings, the format chosen for the test items will depend on many factors including the purpose of the assessment, the context in which it occurs, and the content and skills to be assessed. Both types of items provide benefits for assessment and should be included in the suite of assessment activities.
3.2 Attraction of MC items

There are several reasons why MC items are more popular than other types of items in some test situations. Students find them attractive because they know the answer is there and they do not need to find it. Furthermore, they know they can guess if the answer is not readily recognised and hence score points when the answer is unknown. Incorrect answers may be instantly recognised, clues given from the wording of the question, and student thinking directed towards a correct response. Not having to show working appeals to many students though others prefer to write out their solutions and many teachers prefer to see this evidence of the students’ mathematical thinking. A further attraction for some students is the objective scoring process because the marker cannot be prejudiced by poor handwriting, bad grammar, or personal bias. The marking of MC items is more accurate than for CR as the scoring follows a pre-determined plan that is independent of the marker. With the marking of CR items, the markers’ interpretations of the students’ responses are unlikely to be consistent in all situations and this will result in biased scoring.

There is evidence of the increasing use of MC items in classroom tests. Some of the more recently published textbooks that I have used not only contain MC items for students but are accompanied by a teacher version for the text that provides topic tests containing MC items. When MC items are used during a lesson, teachers can receive immediate feedback on student learning and make the appropriate adjustments to the activities conducted during that particular lesson. Well-written MC items can also provide diagnostic information about the students’ misconceptions and these can also be addressed at the point of recognition.

The attraction of using MC items in large-scale assessments continues and there is evidence to indicate that their use is becoming more extensive. Betts, Elder, Hartley, and Trueman (2009) suggested that the increase could be attributed to the fact that a broader range of the curriculum can be covered with fewer items than would be needed if only CR items were used. Furthermore, they suggested that MC items are attractive because they are easier to score and to administer to a large cohort of students resulting in a less costly marking process and a shorter turnaround time for reporting. Writing on the attraction of MC items, Madaus, Clarke, and O’Leary (1997, p. 433) reported that the “move back towards the traditional norm-referenced, multiple-choice mode in large-scale state programmes” of assessment in the United States was because the “select-the-answer” mode of testing was determined to be an efficient and cost-effective way of collecting evidence of student performance.
3.3 Creation of MC items

Writing effective MC items is necessary for scores to be valid and reliable. Scores are valid when the inferences made from the results are appropriate, meaningful, and useful (Messick, 1995). Reliability of the scores refers to the consistency with which they can be repeated over time (Malau-Aduli & Zimitat, 2012). A further desirable feature for MC items is the ability to discriminate, whereby they can be used to identify the range of abilities of the students sitting the test. If all the items are far too easy or far too difficult, then they will have failed to discriminate between the candidates. Several research studies have focussed on investigating the impact of different aspects of MC items on discrimination, validity, and reliability, and this has resulted in the identification of recommended guidelines for creating MC items.

Writing guidelines

Haladyna, Downing, and Rodriguez (2002) tested and validated a MC, item-writing guideline for classroom assessment and suggested that their taxonomy might have uses for large-scale testing. Their recommendations included (a) placing the main idea in the stem rather than in the options, (b) writing stems and options using positive language, (c) presenting choices vertically rather than horizontally, (d) keeping choices independent, (e) writing options in similar format, (f) ensuring all distractors are plausible, and (g) using typical errors of students to write these distractors.

Placing the main idea in the stem allows the focus of the item to be presented on first reading, and this reduces the need to remind those doing the test of the main idea in each of the options. The format of the stem should be presented as a question rather than as an open statement, according to Statman (1988) who claimed that, repeated reflection back onto the statement to be completed when considering each option, was more demanding for the respondents than holding a question in working memory. This recommendation was not supported by the research of Ascalon, Meyers, Davis, and Smits (2007) who found that there was no effect on item difficulty in a mock-driving licence test when the stem was a question rather than a statement.
Haladyna et al. (2002) also recommended that positive, rather than negative, language should be used in the stem. Karegar Maher, Barzegar, and Ghasempour (2016) reported that the findings from studies into the effect on item difficulty of negative language in the stem, was inconclusive. However, having negativity in the stem did influence the cognitive level at which the items were written. About 64% of the items with negative stems were described as testing the student ability to remember and this student behaviour was described as being at the lowest level of cognition. For the items with positive stems, the proportion which tested student ability to remember was only 30%. A further recommendation for writers was that they try to create options of approximately equal length and of similar grammatical structure (Haladyna et al. 2002), and this is to prevent the test participants from receiving clues that help them to identify the correct response.

Keeping choices independent allows the respondent to consider each option on its own merit. At times, item creators write one option that subsumes another. Consider the example provided in Figure 3.1.

<table>
<thead>
<tr>
<th>How many sides in a square?</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 1</td>
</tr>
<tr>
<td>b. 2</td>
</tr>
<tr>
<td>c. 3</td>
</tr>
<tr>
<td>d. 4</td>
</tr>
</tbody>
</table>

*Figure 3.1* Example of a poorly-written MC item

Theoretically all answers are correct, and the stem needs to be changed if the author wants the student to recognise that to be a square, the figure must have exactly four sides. The language of this item is not ideal as the sides of the square are not *inside* it, rather they make the square.

*All of the above and None of the above*

Another recommendation by Haladyna et al. (2002) is for writers of options to avoid *all of the above* and to take care when using *none of the above* (NOTA). If all of the above is used as the correct response, then the student who identifies two options as correct and who knows nothing about the other options, can immediately recognise that all of the above must be correct. If all of the above is an incorrect response, that is, a distractor, then it can be eliminated when the student recognises another option that is incorrect, and in this situation, the all of the above option has no purpose in the item.
It is possible to select NOTA as the answer by knowing all other options are incorrect, but this does not indicate that the student has good understanding of the concept being tested. Unlike Haladyna, Frary (1991) concluded that using NOTA in classroom tests can be compatible with good measurement. In the empirical studies conducted by Frary, the items with a NOTA option were found to be slightly more difficult than those without the option, and there was little difference between these two types of items in terms of the discrimination measures. The lower-ability examinees tended to avoid NOTA when they were not sure of the answer while the higher ability persons found NOTA more attractive when they were not sure of the answer. Frary suggested that NOTA forces examinees to consider all other distractors and discourages working backwards. In their research into MC items, Pachai, DiBattista, and Kim (2015) systematically substituted options with the NOTA option and they found that there was an increase in item difficulty when NOTA was the correct response but not otherwise. Their recommendation was that NOTA is best avoided.

Number of options

After their review of the research into the optimum number of options for MC items, Haladyna and Downing (1989) were unable to recommend any particular number and instead concluded that one should focus more on the quality of the distractors rather than the quantity. Since this review, the findings from further research about the ideal number of options have been inconsistent. After an analysis of 80 years of research, Rodriguez (2005) recommended three options, as did Bruno and Dirkzwager (1995) and Vegada, Shukla, Khilnani, Charan and Desai (2016). Four options were recommended by Deepak, Al-Umrani, Al-Sheikh, Adkoli and Al-Rubaish (2015) while Peck (personal communication, March 2, 2016) recommended five if more than one option is scored.

Items with fewer options should be quicker to develop, have fewer implausible distractors and facilitate having more items per test. Tarrant, Ware, and Mohammed (2009) also found that three options were ideal, but their studies were based on rescoring items with four options by redistributing scores relating to the least popular distractor. It is unlikely, that items developed with four options where one is removed, can be considered to be equivalent to the same item written with three options. While there seems to be more evidence supporting three options, most items in the national and international tests undertaken by students have four options: Tarrant et al. (2009) suggested that this practice might persist because the negative effect of guessing is perceived to be higher with only three options.
Position of key elements

The positioning of the correct options by creators of multiple-choice items was studied by Attali and Bar-Hillel (2003) who reported at least a 70% chance that the correct answers were in the middle positions of the list of options. They suggested that authors would write the question and answer, next determine distractors smaller and greater than the correct answer, and then order the options numerically; thus, the correct answer was unlikely to be first or last in the list. Attali and Bar-Hillel found mathematical evidence to support the hypothesis that, in such tests, these items were easier and less-discriminating than those where the key was first or last in the list. Attali and Bar-Hillel also investigated student selection of options in items where they did not know the answers and found that 80% of their selections were from the middle two of four options. While for both the creators and the respondents there was a bias towards the middle, it is possible that this is more of an edge-aversion.

Key randomisation was recommended by Bar-Hillel, Budescu, and Attali (2005) to reduce the bias introduced when test-writers try to balance the position of the correct options (key-balancing). They found that the positioning of the keys tended to be over balanced in key-balanced tests and they argued that key randomisation would make it difficult for students to work out any strategies that might have been used to place the correct options. The position of the item within the booklet for a test of fourth grade mathematics was studied by Hohensinn, Kubinger, Reif, Schleicher, and Khorramdel (2011). They found that neither the position of the item, nor the sequence order of the items, affected the difficulty of the item under test conditions when the students had plenty of time to complete the test.

Nature of the distractors

Distractors describe the incorrect statements that examinees can select to complete the stem or answer the question posed in the stem. They may be written to include some of the common errors made by students, or to reflect misconceptions widely held by the students (R. Peck, P. Rogers, & S. Zoumboulis, personal communication, March 7, 2016; Rogers & Zoumboulis, 2015). Using their teaching experience to predict answers likely to be given by the students is one of the principles used to guide the creation of distractors by authors of the SMART tests (V. Steinle, personal communication, March 7, 2016). Another guiding principle for the creation of distractors is to structure the incorrect responses to capture stages of learning (Briggs, Alonzo, Schwab, & Wilson, 2006; Smith, 1987) and this aspect of creating MC items will be discussed in much greater detail later in this dissertation.
For most MC items in a test, the distractors need to be plausible so that the correct answer, the key, is not so obvious to the students. The items can then be used to determine student achievement and to discriminate between students’ performances. I would suggest that having a few items, with an easily identifiable key, at the beginning of a test provides the weakest students with some success and an incentive to continue. For these items the distractors would not need to be highly plausible nor based on common misconceptions or stages of learning.

*Other factors for item writing*

Other factors to consider when writing multiple-choice items for assessments of mathematical achievement are the (a) content, (b) proficiency aspects of mathematical competence, (c) context, and (d) literacy demands (Connolly, 2011). Contexts need to be chosen so that they are sufficiently familiar for the students so as not to confuse or distract them from the mathematical demands of the task (Cumming & Maxwell, 1999). Greenlees (2010) summarised this imperative more succinctly by describing the attempt to make questions more realistic as confounding. The students can be affected by too much information which might be unrelated but which they might attempt to comprehend: They then use the information to solve the problem rather than be drawn to the relevant mathematics.

The common language of the items needs to be appropriate for the age of the students and to ensure that it is the mathematical, rather than the literacy, aspects that are being tested. In their study of language in mathematics tests, Abedi and Lord (2001) rewrote items from a large-scale assessment by modifying particular language features. The items were then checked by mathematicians to determine if there was a change in the mathematical content of the items. After the items were rewritten, there was a significant improvement in test performance for the Year 8 students of low ability, and for those for whom English was a second language.

To improve MC items, Abedi and Lord (2001) recommended that writers (a) use active tense rather than passive tense; (b) remove infrequently used words; (c) separate conditional sentences into two sentences; (d) rewrite relative clauses, for example, ask *determine the number of*, rather than *how many* components; (e) change complex expressions to simple ones; and (f) use concrete descriptions rather than abstract ones, for example, replace *the cost* with *the cost of the car*.
**Item review**

Before items are used in large-scale assessments they are usually piloted or subject to independent reviews. A sample of test participants may be selected to trial the items to identify any issues that might affect the larger collection of results. Items are generally reviewed by experts in the field of the assessed material, peer reviews, and sometimes by potential test participants. Malau-Abduli and Zimitat (2012) introduced a peer review system after asking item writers to create more MC items to test higher-order thinking in examinations for medical courses. The percentage of items which assessed simple knowledge recall fell over the three years that followed and the desired improvement in difficulty was achieved. There was also a decrease in the number of items with negative discrimination and an increase in the reliability of the items.

Item reviews by test participants can be implemented using cognitive interviews which have been described by Campanelli (1997) as the adoption of a cognitive approach during the planning and process of the interaction between an interviewer and their respondent. Such an approach is focussed on understanding the respondent’s interpretation of the language used, on appreciating the role of the respondent’s memory, and on gaining an insight into the perception and judgement shown when questions are answered. This manner of interviewing is not new and was used extensively by Piaget in his conversations with subjects while studying the development of children’s thinking (Piaget, 1952). The evidence gathered during these interviews can be used to check that the interpretation of the items is consistent between students, to edit and improve the items, and to confirm that the students are provided with all the information that they need to respond to the items (Collins, 2003).

Using a cognitive pre-testing process, Gehlbach and Brinkworth (2011) identified three potential problems for survey respondents and these are (a) ambiguity of meaning, (b) over-challenging vocabulary, and (c) ambiguity of context. Collins (2003) identified a further problem for respondents and that is, the complexity of the sentence structure. Editing items to remove such problems increases the likelihood that the items are interpreted as intended. The extent to which respondents interpret items as intended by their authors has been described as a measure of cognitive validity by Karabenick et al. (2007), Koskey, Karabenick, Woolley, Bonney, and Dever (2012), and Wildy and Clarke (2009). The editing of an item, following the identification via the cognitive interview of how it can be improved, can increase this validity and provide users with greater confidence in the accuracy of the measurement scale.
3.4 Concerns with the use of MC items

Even though there has been an increase in the quality and the popularity of MC items, there are several issues that generate considerable debate. The process by which a respondent selects a correct option is a major concern, as the selection may be for the wrong reason or because the student has guessed. Either way, the score will be inflated and not an accurate reflection of the student’s knowledge or achievement. There is a concern that only low levels of thinking can be assessed by MC items, and there is the perception that it is mostly learned facts and procedures that can be tested rather than the ability to think, reason, and solve problems. For some items the selection of particular distractors can indicate better understanding of the item content than the selection of other distractors. When such selection is not scored, there is some concern that the information gathered about student achievement is limited, and that the student’s partial knowledge is not recognised. The quality of the item and its contribution to measurement depend on the nature of the distractors and the relationships between them, so the quality of the distractors is a concern for creators and users of MC items. Another concern with the use of MC items to assess achievement is the perceived bias for males, and the perception that males are more able than females to respond to MC items. For all of these concerns, there are several reports in the research studies, and these provide insight into the impact of these concerns and how they might be addressed.

3.4.1 Guessing in MC items

In responding to a MC item, students may select the key for reasons other than knowing the correct option. It may be that they have no knowledge of the item and either guess randomly between all the options or, because they are able to eliminate some distractors, guess successfully between the remaining options (systematic guessing). The opportunity to guess is reported by Alnabhan (2002), Bar-Hillel et al. (2005), Betts et al. (2009), Stewart and White (2011), and Zimmerman and Williams (2003) as causing students’ scores to be inflated, thus decreasing test reliability and reducing the validity of any measures or scales determined from the scores. Alnabhan (2002) also suggested that the effects on reliability and validity will vary according to the nature of the guessing, whether it is random or systematic. Recommendations to minimise the effects of successful guessing in MC items relate to improving test and item construction, scoring student responses to MC items in different ways, and post-test manipulation of student scores.
Attending to test and item construction

Successful guessing can be minimised, and reliability and validity increased, by attending to test and item construction. The precision with which the score represents student knowledge can be enhanced by increasing the length of the test and this may be done by increasing the number of items or distractors (Alnabhan, 2002; Kubinger & Gottschall, 2007; Paek, 2014; Zimmerman & Williams, 2003). The nature of the distractors and their relationships to each other will influence the extent of guessing, and how these distractors affect guessing will depend on the test participants and their knowledge of the construct material. Bar-Hillel et al. (2005) suggested that cues to the correct response can be reduced by randomising the position of the key, by keeping the options similar in length, and by writing all options to be grammatically compatible with the stem. The position of the key should be randomised to reduce the advantage for students who are expecting a balanced key or who are likely to be favouring options from the middle of the list.

Another strategy designed to reduce the level of guessing is to provide a test that is not too difficult for the test-taker. An algorithm can be used to detect student success in a limited number of questions and according to those responses change the student’s pathway through the test (van der Linden & Glas, 2010). This process is referred to as adaptive testing: the test is adapted to more accurately match each student’s ability. In theory, students responding to items more commensurate with their own ability are less likely to guess. Compared to non-adaptive testing, adaptive testing has been found to generate lower rates of error and hence increased measurement precision (Martin & Lazendic, 2018). Weiss (1982), using both simulation and empirical studies, described other advantages of adaptive testing. One such advantage is a reduction in the number of items needed during adaptive testing to achieve the same levels of validity and reliability that are achieved in a non-adaptive test.

In a review of the literature on computer adaptive testing, Chuesathuchon (2008) described two types of strategies that were used to select the items to present to the students. In the first strategy, student ability was identified by the responses to a set of items, usually about 10, and then a suitable sub-test was provided according to the results. For the second strategy, student response to an individual item was used to determine the next item presented. Success in an initial item of moderate difficulty generated a more difficult item for the student to complete. The process continued until some pre-defined stopping criterion was met.
Alternative methods of scoring

The use of different scoring methods to reduce the amount and impact of guessing has been the focus of several research studies (Alnabhan, 2002; Bar-Hillel et al., 2005; Betts et al., 2009; Budescu & Bar-Hillel, 1993; Burton & Miller, 1999; De Laet, Vanderoost, Callens, & Vandewalle, 2015; Espinosa & Gardeazabal, 2010; Lau, Lau, Hong, & Usop, 2011; Lindquist & Hoover, 2015; Prihoda, Pinckard, McMahan, Littlefield, & Jones, 2008). Scoring methods may be based solely on awarding marks (positive marking) or can involve awarding marks plus applying penalties (negative marking). The conventional and most popular method of positive scoring in which no penalties are applied is often referred to as the Number Right (NR) method in which one mark is allocated to the selection of the key, and zero marks are allocated for any other selections. Where penalties are imposed, marks are usually deducted for incorrect or missing responses, and the most popular of these methods is known as Formula Scoring (FS). A commonly accepted algorithm for FS is summarised in the formula given below (Lindquist & Hoover, 2015, p. 16):

\[ S = R - \frac{W}{n-1} \]

\( S = \) corrected score, \( R = \) number of correct answers, \( W = \) number of incorrect options chosen and \( n = \) total number options.

In the study of scoring methods by Lau et al. (2011), Year 8 students identified each MC option in a test of TIMSS 2003 Mathematics items as correct, incorrect or unsure. Students scored 1 for each distractor eliminated, 1 if they identified the key (correct response), 0 if they were unsure, and -3 if they rated the key as incorrect. The authors referred to this scoring as Number Right Elimination Testing (NRET) and they reported minimal guessing and higher reliability when this scoring was applied. Further research is needed to gather more evidence to support their conclusions as the NR scores used for comparison had been generated from the NRET scores. It could be argued that the selection behaviour of the candidates would change if they were aware that only a correct selection warranted a score and different instructions would be necessary for each of these types of scoring. De Laet et al. (2015) compared a form of elimination scoring with NR scoring using work from first year engineering students. The students indicated that they found it easier, less stressful, and more time-consuming to use elimination marking, and overall, they preferred it to NR scoring.
Other marking schemes to reduce the negative impact of guessing on the validity of scores have been investigated. The method proposed by Bo, Lewis, and Budescu (2014) involved the possibility of having more than one correct option, and a count of the correct classifications of each option. The authors claimed that this reduced guessing. Awarding a fraction of the item score for all omissions, with no penalty for incorrect answers, was studied by Budescu and Bar-Hillel (1993). They reported that this method was preferred by participants who found that the bonus for the omissions was less threatening than a penalty for error.

Allowing the selection of more than one option and ranking the options in a perceived order of correctness, were reported by Alnabhan (2002) to reduce guessing and increase reliability. Adebule and Awodele (2016) investigated confidence scoring where the students indicated their level of certainty about their response when selecting their option to complete the stem. Students received credit for their level of confidence. While the method was found to reduce guessing, the accuracy of the scores may have been compromised by the added factor being measured, that of the students’ belief in their answers. The lack of further research into the effectiveness of these marking schemes for reducing the amount of guessing suggests that the implementation of these methods is problematic.

Different scoring techniques could result in undesirable consequences. There may be extra demands placed on test participants, including comprehension of the scoring system, increased time needed to respond to the items, greater knowledge of strategies for test completion, and the added difficulty of deciding whether to risk omitting or completing items. These demands can influence test reliability by introducing unrelated factors likely to increase the error in measuring the intended construct.

**Penalty scoring**

A likely outcome with penalty scoring is a change in the behaviour of the students who will respond to items according to their understanding of the penalty process. For younger students the concept of losing marks for guessing is likely to discourage a response when they think they know the answer but are not sure. Older students who might understand the impost of the application of penalties may behave otherwise. Betts et al. (2009) found that undergraduate students scored better results and left significantly fewer unanswered questions when told there was no guessing penalty. However, when penalties were applied, and awareness differed, the students who were told there was no penalty outperformed those who were told the penalty was being implemented.
In a comparison of two similar cohorts, also undergraduates, Prihoda et al. (2008) found that the mean increased, and the standard deviation decreased when the students knew guessing was to be penalised. They had been encouraged to guess regardless of the penalties. The significant improvement by the lower ability students was attributed to their concern about the potential failure at the prospect of losing marks for guessing and hence, their greater diligence in preparing for their examination.

Providing instructions about scoring techniques and explaining the impact of penalties for guessing is challenging. It is difficult to provide clear and accurate advice to students sitting a test with MC items and the students may not be able to make wise decisions about their strategy selection (Budescu & Bar-Hillel, 1993). Bar-Hillel at al. (2005) found that the instructions to candidates were often inaccurate, and candidates were usually told of the penalties but not of the statistical advantages of guessing. They were not usually told that guessing replaces a certain zero score with an expected positive though small score. Students may not understand the probabilities associated with different penalties, nor the significance of taking risks. Confusing instructions and decisions about guessing could influence student behaviour in responding to MC items and have a significant effect on the accuracy of the scores. Espinosa and Gardeazabal (2010) found that the effect of risk-taking behaviour on the measurement of ability is small compared to the error prevented. Confirmation of this finding could alleviate concerns about the impact of using penalty scoring to reduce guessing.

Problems with penalty scoring are associated with the uncertainty of the extent and nature of guessing. It is difficult to be certain that a student has guessed and to classify the guessing as random or systematic. In the application of penalties, the assumption is that all guessing is random and that all incorrect responses should be equally weighted. With the equal weighting of responses, some studies have found that while guessing was reduced, so was the validity of the scores (Lindquist & Hoover, 2015). The move away from negative marking may have been influenced by the fact that the correlation between scores in which guessing is penalised, and those where it is not penalised is still very high (Ben-Shakhar & Sinai, 1991). The use of penalties to discourage guessing was abandoned in The SAT Suite of Assessments in the United States of America in 2016 (The College Board, 2018). The theoretical conclusion by Budescu & Bar-Hillel (1993, p. 277) in which they “endorse the number-right scoring rule, which discourages omissions and is robust against variability in respondent motivations, limitations in judgments of uncertainty, and item vagaries” is still relevant.
Post-hoc remediation of guessing

Reducing or removing the effects of guessing from students’ scores can be done after the test and the students are freed from making decisions about appropriate strategies to maximise their scores. Stewart and White (2011) used probability theory to determine the proportion of guessing likely to occur in a vocabulary test and then worked backwards from the observed score to the likely number of words known. Using their method to identify the probability of guessing would be problematic in tests of mathematical understanding because of the greater variety of items in such tests. In the simulation study by Sočan (2015) empirically-derived weights were calculated from the theoretical responses to the items. These weights were then applied to the scores, but they were shown to increase the validity in only a limited number of situations. This research was limited to items with only three options and the qualitative differences between distractors were not considered. Neither of the processes studied by Stewart and White (2011) and Sočan (2015) is widely adopted in the scoring of MC items.

Guessing is not dependent only on the nature of the item but is more a reflection of the relationship between the proficiency of the person and the difficulty of the item (Andrich, Marais, & Humphrey, 2012). Scores that include guessed responses make items appear easier than they really are. It follows that the easiest items are those least affected by guessing, an assumption made by Andrich, Marais and Humphrey (2012) when they removed guessed responses from students’ scores during a post-hoc analysis. This analysis produced more valid estimates of item difficulty: a desirable outcome when items are to be used for item banks or for linking tests.

Andrich and Marais (2014) suggested that the removal of guessed responses gave greater scoring benefit to the more capable persons who are the least likely to guess and who more frequently answer the more difficult questions. They applied this theory to NAPLAN reading data from students in Years 3, 5, 7, and 9, and the results supported the proposal that the more capable students received greater benefit with the removal of the guessed responses (Andrich, Marais, & Humphrey, 2015). When the guessed responses were removed, the estimates of proficiency for the more capable students were higher, and there was evidence of greater growth in proficiency across the years. The process developed in their research can be used to produce estimates of item difficulty and student proficiency free from the effects of guessing. A detailed explanation of how responses are treated in their analysis is provided in Chapter 6.
3.4.2 Low level thinking in MC items

There is anecdotal evidence that many educators believe that MC items only assess low-level thinking and particularly the recall of knowledge. Karegar Maher et al. (2016) studied 2400 MC items from tests in medicine and they engaged experts to rate them according to the different levels of thinking required. The three levels of their taxonomy were (a) the ability to remember, (b) the ability to interpret data, and (c) the ability to solve a new problem. Of all the items studied, 904 belonged in the first level and 1496 belonged to the two higher levels. In another study of the levels of cognition, Masters et al. (2001) examined over 2900 MC items from 17 item banks associated with textbooks in nursing and allocated them to one of the four lower levels of Bloom’s original taxonomy, claiming that MC items cannot be used to test processes at the two higher levels of the taxonomy. According to their analysis, 47% of the items tested knowledge, 24.8% tested comprehension, 21.8% were written at the application level, and 6.5% at the analysis level.

The findings from these studies of levels of thinking in MC items support the claim of Martinez (1999) that MC items can be written to assess complex cognitive processes, and that of Zoumboulis (personal communication, July 8, 2015) who claimed that creators of MC items can be trained to write items that assess complex reasoning and higher-order thinking. Malau-Aduli and Zimitat (2012) reported that MC items often tested knowledge recall but after implementing a focus on writing items aimed at the assessment of higher order thinking, they reported that the items testing knowledge recall fell from 65% to 30% over three years.

One item from my previous investigation (Burfitt, 2014) required the students to interpret an unfamiliar situation and to understand the concept of 120% to work backwards to 100%. The item, shown in Figure 3.2, was deemed to test higher order thinking and only 18% of over 130 students from Years 9 and 10 chose the correct answer.

<table>
<thead>
<tr>
<th>The number of frogs in the creek in 2011 was 120% of what it was in 2010. If there were 60 frogs in 2011, then in 2010 the number of frogs must have been</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 40</td>
</tr>
<tr>
<td>b. 48</td>
</tr>
<tr>
<td>c. 50</td>
</tr>
<tr>
<td>d. 72</td>
</tr>
<tr>
<td>e. 80</td>
</tr>
</tbody>
</table>

*Figure 3.2* MC item deemed to test higher-order thinking

48
Creating items to test higher-order thinking

To create MC items which test higher-order thinking, it is desirable to have a framework for classifying the behaviours of the respondents to assist with the identification and classification of the thinking involved. If using the SOLO taxonomy in an assessment consisting of only multiple-choice items, the options would need to reflect the levels which describe student learning: pre-structural, uni-structural, multi-structural, relational, and extended abstract. According to Biggs and Collis (1982), it would be difficult to provide options at both the relational and extended abstract levels as well as provide options for each of the levels in every item. Using MC items to classifying student behaviour according to the SOLO model is challenging, and a further difficulty is that the students are selecting options written by the authors rather than providing evidence of their own particular behaviours.

Addressing high levels of thinking in MC items has been described by McCurry (2008) as encouraging conceptual thinking and testing understanding as well as knowledge. This may include following an argument, making judgements, interpreting unfamiliar stimulus material, discriminating between concepts, and analysing reasons supporting conjectures. It typically involves the three higher levels of the revised Bloom’s taxonomy, which are analysing, evaluating, and creating (Anderson et al., 2001). Higher-order thinking can be assessed by using visual material, such as a graph or table or by providing a context for the item and thus extend the thinking required to select the correct option (Haladyna et al., 2002).

Another suggestion for addressing higher-order thinking in MC items is to ask the students to identify or provide reason for their choice of options (Gbore, 2016). These two-tiered MC items provide opportunities for the students to think more deeply, to apply learned principles and to develop arguments to justify their choice of option. Other ways to add complexity to MC items is to ask the student to discriminate between highly plausible distractors or to ask for the best option (Bloom et al., 1971; Morrison & Free, 2001).

Techniques to address higher-order thinking make extra demands on item writers as well as on the examinees and are best implemented by specialists. Much of the research literature on addressing higher-order thinking in MC items refers to the use of the original and revised versions of Bloom’s taxonomy to classify the items. Using the taxonomy to write MC items for tests of mathematical understanding for secondary students was also recommended by Peck, Zoumboulis and Rogers (personal communication, March 7, 2016) who suggested using the same descriptors as used in the TIMSS framework: knowing, applying and reasoning.
3.4.3 Consideration of partial credit in MC items

A major criticism of using MC items to assess student understanding is that they are usually marked right or wrong and there is no opportunity to receive a score for knowing something, though not everything, which could assist with the selection of the correct response. While only the correct answer is scored, there may be some distractors (incorrect options) that are nearly correct, and if selected would indicate the student has some understanding of the construct being tested. These incorrect options may be better choices than other incorrect options and, when chosen there may not be any credit given to the student for their partial learning. In my earlier study (Burfitt, 2014) there were items in which some distractors were selected more often than others and these selections indicated that the students had partial knowledge of the item’s content. Part of one such item is provided in Figure 3.3.

<table>
<thead>
<tr>
<th>The price of a maths text book has risen 25% to a new cost of $100. The old price must have been:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. $75</td>
</tr>
<tr>
<td>b. $80</td>
</tr>
<tr>
<td>c. $100</td>
</tr>
<tr>
<td>d. $125</td>
</tr>
</tbody>
</table>

*Figure 3.3* MC item where selection of distractor may indicate partial knowledge

To select an appropriate response the student first needed to interpret the item. This should have immediately provoked the thinking that the answer had to be less than $100 and therefore options (c) and (d) could be eliminated. As expected, very few students chose either of the last two options and the most popular selection for students of low and average abilities was (a). In scoring the correct answer only, the students who interpreted the item but made an error in applying the percentage did not have receive any credit for what they were able to do.

*Elimination scoring*

The ability to eliminate some of the distractors has been described as having partial knowledge and many researchers have studied elimination scoring (Ben-Shakhar & Sinai, 1991; Bond et al., 2013; Burton & Miller, 1999; Frary, 1980). The NRET method described earlier is a form of elimination marking where students had to eliminate as many incorrect choices as possible before selecting the correct response. In that study, Lau et al. (2011)
claimed that they were able to detect and give credit for partial knowledge, but their method is more a reward for knowing what is incorrect rather than knowing what is partly correct. In the study of elimination marking by De Laet et al. (2015), the students were required to mark each option as possible or impossible (eliminated). The authors gave different levels of credit for partial knowledge according to the number of incorrect options eliminated. The students preferred this method of scoring MC items over the use of penalties, as they found it less stressful. Such a process has disadvantages, however, as the extra information collected would require an increase in the time needed to complete the test or a reduction in the number of items that could be tested. The content covered by the test is thus reduced.

Lesage, Valcke, and Sabbe (2013) conducted a review of scoring for MC items and identified three main types of scoring that accommodated partial credit: liberal marking, elimination testing, and confidence weighting. With liberal marking students can select more than one option and thus the MC item presents as a set of True or False statements. Along with elimination testing and confidence scoring (described earlier) each of these methods adds extra demands on the respondents and there is little evidence to indicate that any of these methods have been widely adopted for scoring partial credit in MC items.

**Partial credit scoring**

The results of the analysis of the responses of Year 3 students to CR items testing spatial and mathematical understanding indicated that it is possible to provide a more detailed achievement scale if items are scored polytomously, for example, by allocating 2 marks for a fully correct answer, 1 mark for a partially correct answer (partial credit), and 0 marks for missing or incorrect answers (Van Wyke, 2003). If an MC item worked with one option deserving partial credit, there could be more information statistically, that is, in terms of the precision of the estimates, than with simple correct/incorrect scoring. There could also be an increase in the qualitative information describing the conceptual understanding of the item. Having an MC item with more information can enhance the detail of the achievement scale and may even be equivalent to having two items rather than one. Thus, the measurement process can become more efficient without adding to the test demands on the participants.

Andrich and Styles (2011) used the data from a mathematics test for primary school students to identify distractors that contained information about the students’ partial knowledge. From a population of 20 000 students, they selected samples of 1000 students and examined their responses to 39 items in a state-wide test. The items were not written specifically to contain
distractors with information and Andrich and Styles found that only one of the nine items, hypothesised to contain a distractor with information, satisfied the necessary criteria when tested. For an item to function effectively to collect information about students’ partial knowledge, the distractor to be awarded partial credit has aspects of the correct response and, at the same time, is not an attraction for the most proficient students. Andrich and Styles (2011) concluded that it is challenging but plausible to create such items. Creating such items requires expertise of the relevant content knowledge, and an ability to construct and interpret distractors which facilitate the identification of partial knowledge.

3.4.4 Quality of the distractors in MC items

Writing high quality distractors is essential for the development of an effective MC item. The selection of the distractors can indicate the knowledge that the respondents are lacking as well as the knowledge that they have developed. Being able to detect the extent of the respondents’ knowledge can provide a valuable insight for planning further learning experiences. As described earlier, suggestions for creating good quality distractors include using common errors made by students, anticipating what students might think, and basing the options on known misunderstandings of the content. Distractors vary in quality and in what they contribute to the measurement of student ability. Some studies have provided guidance for improving the quality of the distractors, but more is still needed.

Distractors may be described as “functional” or “non-functional” but the definitions of these two terms vary. While Deepak et al. (2015) defined non-functioning distractors as those not chosen by any candidates, others (Malau-Aduli & Zimitat, 2012; Tarrant et al., 2009) use the 5% rule attributed to Wakefield (as cited in Rodriguez, 2005, p. 5): a distractor chosen by fewer than 5% of the candidates was classified as non-functional. For Tarrant et al. (2009), functional distractors also needed to show positive discrimination power: a measure of the difference in the proportion of responses between the upper and lower achievers. Tarrant et al. found in their study of teacher-developed tests for nurses, that only about half of the distractors could be classified as functional, and that very few items had three functioning distractors. They concluded that the distractors had been quite difficult to write.

Distractor dysfunctionality was described by Malau-Aduli and Zimitat (2012) as having a negative impact on the learner, not assisting with the measurement, and not contributing any psychometric (statistical) information. In their implementation of item peer reviews described earlier, they noted a decrease in the proportion of non-functional distractors over the three
years of the study, during which authors were aiming to increase the extent of higher-order thinking addressed in the MC items. While there are these disadvantages of retaining dysfunctional distractors in MC items, the presence of a few items where the answer is readily recognised because the distractors are easily eliminated, can provide confidence for the lower-ability students, particularly if such items are located at the beginning of the test.

Variation in the quantity and content of the distractors can affect item quality and important components of measurement, namely item fit, measurement error, item discrimination, and item difficulty (Ascalon et al., 2007; Haladyna, 2004; Sideridis, Tsaousis, & Al Harbi, 2016; Tarrant et al., 2009). Ascalon et al. (2007) structured the distractors to resemble the key in theme, word, and sentence length as well as content, and the effect was to increase the item difficulty and hence the validity of the scores. Haladyna (2004) suggested that giving more attention to the creation of distractors will lead to better items, and hence more effective scoring, better reliability, and greater precision for students of lower ability. Haladyna (2004) recommended the removal of distractors that were identified as not being useful, as did Tarrant et al. (2009) who also reported that the items became more difficult as the number of functional distractors increased.

Sideridis, Tsaousis, and Al Harbi (2016) studied the value of analysing the behaviour of the distractors in MC items and examined different approaches that could improve estimates of measurement. They applied Rasch analysis to the responses of over 2000 students to MC items in a national chemistry examination. The techniques they investigated included providing credit for some distractors, combining items to form super items, splitting items believed to contain informative distractors, and removing responses for individuals likely to have guessed (tailoring). All techniques, except for item splitting, were associated with lower measurement error and improved item fit.

When all distractors can be easily eliminated because they are not plausible, little information about the student’s knowledge is gathered. The quality of the distractors, and the relationships between them, need to be considered relative to the content as well as in terms of the statistics and metrics generated by their selection. Attending to the quality of the distractors provides opportunities for crediting partial knowledge as well as providing valuable information that can be used for planning future learning. The role of the distractors and the manipulation of techniques for scoring distractors in pre and post-hoc analyses are further discussed in later sections of this study.
3.4.5 Perception of gender bias

An ongoing concern in Mathematics is the achievement gap in performance for males and females. For each of the years from 2015 to 2017 in both Years 7 and 9, the proportion of males in the top two bands of achievement in the NAPLAN numeracy test was notably higher than the proportion of females (ACARA, 2015b, 2016a, 2017a). As expected, the relationship was reversed in the next two bands where the proportion of females was higher. These differences in performance between males and females are shown in Figure 3.4.

**Figure 3.4** Achievement in NAPLAN Numeracy in Years 7 and 9: 2015–2017
It is feasible to conclude that males are better than females at mathematics but given the dominance of MC items in these assessments, it is possible that the advantage for males is due to the item format.

There are studies that report significant differences between males and females in terms of their mathematical achievement. In a study of approximately 4000 Grade 11 students, Garner and Engelhard (1999) found that females significantly outperformed males on MC items in algebra, and males significantly outperformed females with MC items relating to proportional reasoning, geometry, number, and data analysis. The analysis of PISA 2000 and 2003 results by Liu and Wilson (2009) indicated a significant advantage for males in items relating to space and shape, but overall male performance was only slightly better. For students in Grades 4, 7, and 10, the results of Taylor and Lee (2012) showed that females outperformed males in data analysis, graphing data, reasoning, problem solving, multistep word problems, and mathematical representations. Meanwhile, males were more successful in conceptual understanding, procedural skill in geometry, measurement, algebra, and probability. Taylor and Lee (2012) reported the lack of consistency in findings from earlier research into gender differences in achievement of the various topics. The lack of definitive findings in this area indicates more research is needed before it possible to determine in which content areas of mathematics, if any, students of either gender are more successful.

While there is a community perception that males are more successful with MC items, the differences in achievement between males and females on MC items were reported as being insignificant in research studies by Behuniak et al. (1996), Bond et al. (2013), Bonner (2013), DeMars (1998), Hudson (2010), Krueger (1999), and O’Neil and Brown (1998). Simultaneously, studies reporting female or male advantages in the use of MC items were also located (DeMars, 2000; Liu & Wilson, 2009; Taylor & Lee, 2012). In the mathematics and science sections of a high school test, DeMars (2000) reported the superior performance of males on the MC items and the opposite effect for the CR items. Using data from PISA tests of mathematical literacy, Liu and Wilson (2009) reported a superior performance by males in the 2003 tests but the differences in the performances of the two groups in 2000 had not been significant. A third study in which males outperformed females involved MC items in reading and mathematics for students in Years 4, 7, and 10 (Taylor & Lee, 2012). In one study involving three MC examinations (Betts et al., 2009) females performed significantly better on one MC examination but not on the other two. From the research findings, it is difficult to conclude that either group is more successful with the MC format.
**Omission in MC items**

Of further concern is the detection of differences in patterns of omission by males and females and the influence that this may have on the accuracy of the measures of achievement. Studies reporting a lack of significant difference between males and females in their propensity to omit items include those of Betts et al. (2009), Funk and Perrone (2016), and De Laet et al. (2015). In comparison, Ben-Shakhar and Sinai (1991) investigated omission patterns in tests of 900 University and Year 9 students and found that in all types of tests where there were no penalties, males were less likely to omit and the difference, at times, was significant. Ben-Shakhar and Sinai also reported that the correction of scores for guessing reduced the scoring advantage that males had when there was no correction for guessing, and it increased the advantage for females who had achieved well at the higher levels.

In any study of the relationship between gender and omission patterns, factors which influence student behaviour in MC tests should be considered. Baldiga (2014) found that, when there was no difference in knowledge or confidence, females answered significantly fewer MC items than males if a guessing penalty was imposed. Risk-taking behaviour only accounted for about half of the difference between the omission rates of the two groups: confidence was considered as another possible influence. In the study of approximately 2000 students with an average age of 11 years, girls were found to skip significantly more of the mathematics questions when the questions were difficult (Riener & Wagner, 2017). This difference was only reported for the students doing the academic track courses and not for those on track to do the vocational courses. Furthermore, the gender differences disappeared when the students could win a reward (e.g., a certificate) for their performance.

Studies of omission patterns from approximately 430 000 students over 20 years showed that, except in Grade 3, the tendencies to omit did not influence the overall differences in the mean scores of males and females (von Schrader & Ansley, 2006). In Grade 3, girls tended to omit more MC items than boys in mathematics tests, and boys tended to omit more in reading and vocabulary. Von Schrader and Ansley had considered omission rates for Grades 3, 7, and 11 and they noted a decrease in the proportions of omissions over the 20 years, as well as a decrease from Grades 3 to 11 with less than 1% of omissions in Grade 11. They reported higher rates of omission for the low achievers in mathematics, and also higher rates of omission in the speeded computation test by students in the high ability group. The group with the highest achievement for any test did not necessarily have the lower tendency to omit.
As with mathematical achievement and performance on MC items, the differences between males and females in terms of their omission patterns are not due to gender alone. The nature of gender differences in omission patterns depends on many other factors including age, incentive to succeed, the use of penalties for guessing and the difficulty level of the items.

3.5 Improving the function of MC items

There is anecdotal evidence that many educators would prefer to ask students to show their solution processes rather than use MC items to assess student competence in mathematics. Regardless of the perception that these items do not provide accurate measures, such items can provide valuable insights into student learning. At times, assessment using only MC items provides valid scores that can be used for determining measures of student achievement, and on other occasions CR items alone or in conjunction with MC items provide the required information. As well as contributing to the measurement of student achievement, MC items are popular, easy to score, practical to implement, and economical to use. It is highly likely that they will continue to be used in a wide variety of national and international assessments, and thus it is worthwhile to continue to improve their creation, delivery, and analysis.

Research findings can provide valuable information for the creation of items and support for writing items that address the higher levels of mathematical thinking. It is expected that the growth in the proportion of items that address higher order thinking will stimulate further expectation of such items in assessments and hence more research into providing such items. Guidelines for creating distractors will continue to be produced as more is learned about how student thinking and understanding of mathematics develop. Improvement in this area should lead to better distractors and more suitable relationships between them. This in turn will improve the quality of items in terms of their levels of difficulty and discrimination.

Concerns with the inflated scores due to guessing can be addressed by improving the quality of the items and by providing tests that are not too difficult for the students. This could also be achieved by using an adapted testing regime or by conducting post-hoc analyses of student scores. Possible differences in achievement between males and females can be reduced by ensuring the context of items is not biased towards either group. It is expected that the findings presented from this study of the function of MC items will contribute to the growing research in the creation, implementation, and analysis of MC items for use in assessments of mathematical achievement, and of achievement in other learning areas.
CHAPTER 4: Proportional reasoning

Introduction

Proportional reasoning has been chosen as the mathematical context for the investigation of MC items in this study because it is an important understanding for students to develop as they study Mathematics throughout their school years. It involves the acquisition of skills that are useful well beyond secondary schooling. Proportional reasoning has various connotations in the literature but there is general agreement that it involves the ability to understand and solve problems when the relationship between values is proportional. It is an important skill for people to acquire for them to appreciate the mathematics relevant to their social, educational, and employment situations. For sound proportional reasoning, students should adopt mathematical behaviours indicative of an understanding of proportionality. Students should also develop their knowledge and skills for working with fractions, decimals, percentages, rates, and ratios.

The Australian Curriculum: Mathematics indicates the relevant knowledge and skills that school students need to acquire for the development of sound proportional reasoning. The growth of expertise in these skills, and hence in proportional reasoning, occurs over many years and several developmental pathways have been identified. Many students find the concepts associated with proportional reasoning difficult to understand and to apply. Their learning of a concept is at times only partial and may occur in parallel with the development of identifiable misconceptions. Knowing how students think about proportions can provide valuable information for the selection and timing of appropriate learning experiences.

4.1 Importance

Supporting students to develop sound proportional reasoning should be a major goal of educating secondary students. Proportional reasoning pervades all areas of mathematics for lower secondary students and underpins many aspects of the upper school curriculum including similarity, spatial understanding, finance, percentage change, trigonometry, functional relationships, and algebraic formulation. Steinthorsdottir and Sriraman (2009) described the acquisition of sound proportional reasoning as critical for the development of higher mathematical reasoning while Jitendra, Star, Rodriguez, Lindell, and Someki (2011) referred to it as a bridge to higher mathematics.
Sound proportional reasoning is an essential understanding for applying laws in Physics, for balancing equations in Chemistry, for interpreting scale diagrams and for demonstrating many other higher-order skills. It is also a basic understanding that students need to develop to become functionally numerate adults. For adults to appreciate and to make wise decisions about what constitutes an appropriate pay rise, a good return on an investment, a better buy, or a healthier food item, it is essential that they have good proportional reasoning skills.

Proportional reasoning is described by Long (2011, p. 65) as a critical threshold concept, one for which students develop “radical new perspectives in relation to both the reconceptualised mathematics and the associated operations.” For Watson, Jones, and Pratt (2013), ratio and proportional reasoning make up one of the seven key concepts which spread across, and connect, the many aspects of mathematics curricula; these key concepts are “not taught explicitly” but “valued implicitly” (p. 14). Proportional reasoning is classified by Siemon, Bleckly, and Neal (2012) as one of the six big ideas of number and as the key big idea for students in early secondary, one that students should have mastered by the end of Year 8. The concept of a big idea is described by these authors (p. 22) as “an idea, strategy, or way of thinking about some key aspect of mathematics without which, students’ progress in mathematics will be seriously impacted.” This way of thinking is not easily defined as it includes many other skills, concepts, ideas, and strategies, but it does provide a point of reference and is recognised when seen in student activity.

Reasoning in proportions is a thread which connects many mathematics topics and is of such importance that every effort should be made to support students in their development of the associated ideas (Lamon, 2007). The development of proportional reasoning is “critical to mathematical and scientific thinking and (there should be) a sense of urgency about the consistent failure of students and adults to reason proportionally” (Lamon, 2007, p. 637). For this domain of reasoning students need to become proficient in several topics including fractions, ratios, and proportions. The importance is highlighted by Lamon (2007, p. 629):

Of all the topics in the school curriculum, fractions, ratios, and proportions arguably hold the distinction of being the most protracted in terms of development, the most difficult to teach, the most mathematically complex, the most cognitively challenging, the most essential to success in higher mathematics and science, and one of the most compelling research sites.
4.2 Defining proportional reasoning

Within the many research studies of proportional reasoning and its associated concepts, there is an extensive range of terminology used to define the associated behaviours. There are descriptions where, according to Lamon (1993), they include the non-proportional thinking used when solving problems about proportions. In a review of the literature, Tourniaire and Pulos (1985, p. 181) described the increasing sophistication of the research as “changing from a view of proportional reasoning as a global ability, or a manifestation of a general cognitive structure, to a more differentiated view focusing on describing the procedures used in proportional reasoning and how they are influenced by task and person parameters.”

De la Torre, Tjo, Rhoads, and Lam (2013) analysed studies published after 1985 and noted that definitions of proportional reasoning related to the researchers’ concepts of proportion, and these in turn depended on the authors’ definitions of ratios and rates. Beckmann and Izsak (2015) linked two alternative perspectives on proportional relationships to two different approaches to the research. They reported that the research was mainly about visualising proportions as variable numbers of fixed quantities, but more research on the fixed numbers of the variable parts perspective would provide better insight into student learning. While the definitions of what constitutes proportional reasoning are quite different, and this impacts on the interpretation of studies in this field, an analysis of the definitions used in research studies shows that they have much in common.

In some research studies the focus for the definition of proportional reasoning is on making multiplicative comparisons. Such behaviour is also identified in the literature as multiplicative thinking, which is described by Siemon (2015) as an ability to recognise and solve problems involving multiplication and division with a range of different numbers including fractions, decimals, and percentages. Parish (2010) and Fernandez, Llinares, Van Dooren, de Bock, and Verschaffel (2011) referred to the concept of proportional reasoning as that of making multiplicative comparisons between quantities, and they highlighted the necessity to include an ability to distinguish between proportional and non-proportional situations.

Dole, Clark, Wright, and Hilton (2012) included in their description of proportional reasoning, the need to understand as well as to make the multiplicative comparisons, and they described proportional reasoning as “knowing the multiplicative relationship between the base ratio and the proportional situation to which it is applied” (p. 195). The nature of the comparisons is crucial for proportional reasoning to be evident and, describing it simply as
multiplicative thinking to make comparisons provides insufficient insight into the true nature of proportional reasoning: Further detail is warranted.

In some studies of student development and achievement, proportional reasoning has been defined by implication with the term used to describe the behaviours of students while they solved problems with ratios or proportions (Howe et al., 2015; Jitendra, Star, Rodriguez, Lindell, & Someki, 2011; Lawton, 1993; Misailidou & Williams, 2003, 2008; Noelting, 1980a, 1980b; Steinthorsdottir & Sriraman, 2009). Definitions of ratios vary in studies of proportional reasoning where they are described either as comparisons of quantities of the same unit, or of quantities of different units, or not described at all. Examples of ratios with different units include “3 lollies for 10c” and “80 km per hour” (Parish, 2010, p. 469) and a comparison of the number of cats with the number of cans of food eaten by the cats (Steinthorsdottir & Sriraman, 2009).

Ratios are mostly defined as comparisons of quantities of the same unit, for example, red and yellow roses. Jitendra et al. (2011) used this definition of ratios and extended their definition to include percentages as special types of ratios, a comparison not seen elsewhere in the literature. Some references to ratios in the literature imply that they are either included in proportions, which are not always explicitly defined, or treated as a separate topic to proportions. When comparing studies of achievement in proportional reasoning, the terms used in connection with the related behaviours, for example, ratios, need to be clearly defined.

In some studies, the focus when describing the meaning of proportional reasoning is on the equality of two ratios. According to Kilpatrick et al. (2001, p. 8), “Proportions are statements that two ratios are equal. Understanding and working with the relationships in a situation involving proportions is called proportional reasoning.” Inhelder, Piaget, Parsons, and Milgram (1958, p. 314) also stated that mathematical proportions can be considered as an equality of two ratios and they expressed this relationship as \( \frac{x}{y} = \frac{\bar{x}}{\bar{y}} \). In other studies (Fisher, 1988; Tourniaire & Pulos, 1985), the equivalent algebraic representation \( \frac{a}{b} = \frac{c}{d} \) was used as an essential component of any definition of proportional reasoning. Tjoe and de la Torre (2014a) also included the ability to use this formula as a pre-requisite for proportional reasoning, and they claimed that students also need to identify situations as being proportional or non-proportional in nature.
Being able to set up the algebraic representation of the equality of two ratios, and to use the formula to solve a problem with a missing value, for example to determine a value for \(x\) in the equation \(\frac{3}{x} = \frac{7}{12}\), is considered to be using proportional reasoning by some authors (Fisher, 1988; Kilpatrick & al., 2001) but not by others (Lamon, 2007; Lesh, Post, & Behr, 1988; Tjoe & de la Torre, 2014a). Students with sound proportional reasoning should have a deeper understanding of the multiplicative relationship between the two quantities which are themselves relationships. This understanding (Lamon, 2005, p. 4) would include “detecting, expressing, analysing, explaining and providing evidence in support of assertions about proportional relationships.” For Lamon (2007, p. 637), proportional reasoning means:

- supplying reasons in support of claims made about the structural relationships among four quantities, (say \(a, b, c, d\)) in a context simultaneously involving covariance of quantities and invariance of ratios or products; this would consist of the ability to discern a multiplicative relationship between two quantities as well as the ability to extend the same relationship to other pairs of quantities.

While Karplus, Pulos, and Stage (1983) reported that the application of a linear relationship between two variables is evidence of proportional reasoning, others (Lamon, 2005, 2007; Long, 2011; Nabors, 2003) state that understanding and manipulating the formula \(y = mx\) that models this proportional relationship is also essential. Students should be able to use the formula to establish a statement of the equivalence of two ratios as described previously, that is, \(\frac{x}{y} = \frac{\dot{x}}{\dot{y}}\) (Briggs & Peck, 2015). This formula expresses the relationship for direct proportion: as one variable increases, so does the other.

An example of direct proportion can occur with the relationship between mass and cost, for example, the greater the number of kilograms of apples purchased, the greater the cost. For indirect or inverse proportion, there is a reciprocal relationship between two variables and as one variable increases the other decreases; for example, the greater the number of persons removing litter, the less time it should take to remove all the litter. Lamon (2007) states that the ability to interpret, establish, and manipulate the formula, \(y = mx\), exemplifies the larger construct of proportionality. This is a more complex and more challenging concept than proportional reasoning and it requires understanding symbolic notation, the rate of change, the constant of proportionality, and linear and reciprocal relationships.
The types of behaviours described by Vergnaud (1983) as belonging to the *multiplicative conceptual field* include all the aspects of proportional reasoning described above. Vergnaud describes the simple direct relationship between two measures, which is represented by the linear equation \( y = kx \), as the *isomorphism of measures*. An example of a problem belonging to this category is: *If 3.5 kg sugar is needed for 5 kg fruit in a recipe, how much sugar would be required for 8 kg fruit?* Another type of problem of reasoning with proportions is identified as a *multiple proportion* in which one measure space is proportional to two different and independent measure spaces. Examples include, *If it costs $45 per day to hire a bicycle, how much would it cost 17 people for 13 days?* and *How long will a 1.5 kg tin of coffee last if 3 people each use 125 g per person per day?* While Vergnaud did not define the term proportional reasoning, these are examples of what is now considered as proportional reasoning. Long (2011) goes further by describing proportional reasoning as a cognition within the general field of multiplicative thinking and reasoning, which is described by Vergnaud as the *multiplicative conceptual field*.

There is considerable overlap in the given descriptions of proportional reasoning and one can argue that none of them is mutually exclusive. For this investigation into the role of MC items, proportional reasoning will be taken to include all of these considerations because each one contains some element of reasoning with proportions. However, to provide a common framework for discussing proportional reasoning, and for designing related activities, the idea that proportional reasoning “involves recognising and working with relationships within relationships (i.e., ratios) in different contexts,” as suggested by Siemon et al. (2012, p. 32), will be adapted and adopted for this study.
4.3 Behaviours associated with proportional reasoning

Examining the behaviours associated with sound proportional reasoning enables the identification of the skills and understandings that students should have developed if they are deemed to be proficient in this area. Several behaviours which students are expected to adopt have been described in some of the major studies and these are summarised in Table 4.1. The expected skills are quite complex, and their development will occur over considerable time.

Table 4.1 Behaviours associated with proportional reasoning

<table>
<thead>
<tr>
<th>Characteristics of proportional thinkers according to Lamon (2005)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students:</td>
</tr>
<tr>
<td>• distinguish proportional from non-proportional situations</td>
</tr>
<tr>
<td>• develop and efficiently use strategies for solving problems</td>
</tr>
<tr>
<td>• think of multiples as units for scaling, e.g. per 100 g, per 250 g</td>
</tr>
<tr>
<td>• understand the terminology associated with proportions</td>
</tr>
<tr>
<td>• represent numbers as a variety of products</td>
</tr>
<tr>
<td>• use decimals, fractions and rates flexibly</td>
</tr>
<tr>
<td>• use all four operations mentally and efficiently</td>
</tr>
<tr>
<td>• scale up and down using appropriate strategies</td>
</tr>
<tr>
<td>• can formulate and use statements of equivalent ratios.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indicators of proportional reasoning according to Siemon et al. (2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students:</td>
</tr>
<tr>
<td>• use multiplicative thinking (cf. absolute thinking) to analyse change</td>
</tr>
<tr>
<td>• identify and describe relationships between quantities</td>
</tr>
<tr>
<td>• work flexibly with rates, ratios, rational numbers, percentages and integers</td>
</tr>
<tr>
<td>• dilate 2-D images using a scale factor.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attributes of proportional reasoning according to Tjo &amp; de la Torre (2014b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students:</td>
</tr>
<tr>
<td>• have pre-requisite skills, e.g. use all four operations flexibly, know factors and multiples, use fractions to represent quantities, understand equivalent fractions</td>
</tr>
<tr>
<td>• compare and order proper fractions</td>
</tr>
<tr>
<td>• express situations as proportions or ratios, including equivalent ratios</td>
</tr>
<tr>
<td>• identify multiplicative relationship between quantities</td>
</tr>
<tr>
<td>• distinguish between proportional and non-proportional relationships, and direct and indirect proportions</td>
</tr>
<tr>
<td>• apply algorithms, for example, the cross-multiplication process.</td>
</tr>
</tbody>
</table>
4.4  Skills and understandings needed for sound proportional reasoning

Proportional reasoning is recognised as one of the sub-elements of the numeracy continuum in the Australian curriculum, and while there is no further reference to proportional reasoning in the content descriptions of the mathematics curriculum documents, there is reference to both reasoning and to proportion (ACARA, 2015a). Reasoning is one of the proficiencies identified as a mathematical behaviour to be addressed in every year level. There are very few mentions of proportion in the mathematical content sections of the documents and these only refer to students in Years 8 and 9. The scant reference to proportion is provided in Table 4.2, and while the lack of reference to proportional reasoning in the curriculum documents is a concern, it does not imply that it is not essential for teachers to teach, and for students to learn, the relevant skills. The curriculum documents specify the skills and understandings necessary for the development of sound proportional reasoning and it is the role of the teachers to interpret these documents and provide learning opportunities for students to apply reasoning to their knowledge of proportion.

Table 4.2  Proportion in the curriculum documents

<table>
<thead>
<tr>
<th>Location</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 8:</td>
<td>Reasoning includes justifying the result of a calculation or estimation as reasonable, deriving probability from its complement, using congruence to deduce properties of triangles, finding estimates of means and proportions of populations.</td>
</tr>
<tr>
<td>Year 8 content:</td>
<td>Explore the variation of means and proportions of random samples drawn from the same population.</td>
</tr>
<tr>
<td>Year 9 content:</td>
<td>Solve problems involving direct proportion. Explore the relationship between graphs and equations corresponding to simple rate problems.</td>
</tr>
<tr>
<td></td>
<td>Elaboration: identifying direct proportion in real-life contexts</td>
</tr>
</tbody>
</table>

As described in Chapter 2, the Australian curriculum: Mathematics is organised as two strands: content and proficiencies (ACARA, 2015a). The content strands are number and algebra, measurement and geometry, and statistics and probability. Content sub-strands include fractions and decimals, real numbers, money and financial mathematics, chance, and linear and non-linear relationships. It is these descriptions of content that indicate the skills and understandings that students need to develop for sound proportional reasoning (see Appendix 4.1): A summary of the content relevant for Years 6 to 8 is given in Table 4.3.
<table>
<thead>
<tr>
<th>Table 4.3</th>
<th>Curriculum content of skills for proportional reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic</strong></td>
<td><strong>Skills relating to aspects of content</strong></td>
</tr>
<tr>
<td><strong>Number</strong></td>
<td>Know common multiplication and division facts</td>
</tr>
<tr>
<td></td>
<td>Know fraction, decimal, percentage equivalents</td>
</tr>
<tr>
<td></td>
<td>Be able to convert between number formats</td>
</tr>
<tr>
<td><strong>Fractions</strong></td>
<td>Use fractions to represent given descriptions</td>
</tr>
<tr>
<td></td>
<td>Compare and order fractions</td>
</tr>
<tr>
<td></td>
<td>Create and interpret equivalent fractions</td>
</tr>
<tr>
<td></td>
<td>Determine the fraction of an amount</td>
</tr>
<tr>
<td></td>
<td>Use all four operations on fractions</td>
</tr>
<tr>
<td></td>
<td>Convert between proper and improper fractions</td>
</tr>
<tr>
<td></td>
<td>Use fractions as numbers, measures and operators</td>
</tr>
<tr>
<td></td>
<td>Determine and interpret fractional inverses</td>
</tr>
<tr>
<td><strong>Decimals</strong></td>
<td>Use all four operations on decimals</td>
</tr>
<tr>
<td></td>
<td>Compare and order decimals</td>
</tr>
<tr>
<td></td>
<td>Multiply and divide decimals by powers of 10</td>
</tr>
<tr>
<td><strong>Percentages</strong></td>
<td>Determine a percentage of an amount</td>
</tr>
<tr>
<td></td>
<td>Express a relationship as a percentage</td>
</tr>
<tr>
<td></td>
<td>Change an amount by a percentage</td>
</tr>
<tr>
<td></td>
<td>Calculate percentage change</td>
</tr>
<tr>
<td><strong>Ratios</strong></td>
<td>Create and interpret ratios</td>
</tr>
<tr>
<td></td>
<td>Create equivalent ratios</td>
</tr>
<tr>
<td></td>
<td>Express ratios as fractions</td>
</tr>
<tr>
<td></td>
<td>Share an amount according to a given ratio</td>
</tr>
<tr>
<td><strong>Rates</strong></td>
<td>Calculate a unit price</td>
</tr>
<tr>
<td></td>
<td>Use units flexibly</td>
</tr>
<tr>
<td></td>
<td>Recognize rates</td>
</tr>
<tr>
<td><strong>Proportion</strong></td>
<td>Identify and operate flexibly with a range fractions,</td>
</tr>
<tr>
<td></td>
<td>percentages and decimals as proportions.</td>
</tr>
<tr>
<td></td>
<td>Distinguish two types of change - non-proportional</td>
</tr>
<tr>
<td></td>
<td>and proportional change</td>
</tr>
<tr>
<td></td>
<td>Apply a scale to dilate a 2-dimensional figure</td>
</tr>
<tr>
<td><strong>Linear relationships</strong></td>
<td>Represent relationships on the Cartesian plane</td>
</tr>
<tr>
<td></td>
<td>Interpret and translate linear relationships</td>
</tr>
<tr>
<td></td>
<td>Express relationships as tables, words and symbols</td>
</tr>
<tr>
<td></td>
<td>Create and solve linear equations</td>
</tr>
</tbody>
</table>
The four proficiencies, which identify the behaviours of students as they learn the content of the Australian Curriculum: Mathematics, are described in general terms:

   The proficiency strands are understanding, fluency, problem-solving and reasoning. They describe how content is explored or developed; that is, the thinking and doing of mathematics. The strands provide a meaningful basis for the development of concepts in the learning of mathematics and have been incorporated into the content descriptions of the three content strands. (ACARA, 2015a)

For each year group, the expected mathematical behaviours are described in the Australian curriculum in terms of the four proficiencies: understanding, fluency, problem-solving and reasoning (ACARA, 2015a). Students should develop fluency in calculations and multiple representations, and they should understand the connections between various forms of numbers and select appropriate methods. Students should develop the skills needed to solve problems and they should be able to explain and justify their results. A summary of the proficiencies with a focus on the skills and understandings necessary for the development of sound proportional reasoning for each of these year groups, is provided in Table 4.4. Greater detail for each of the four proficiencies is provided in Appendix 4.2.

For this investigation into the function of MC items, the definitions used for key words, ratios, rates, and proportions, are those which are in common use throughout Australia, and which are also found in the literature. Ratios refer to comparisons of like quantities which may have different units, but which have some relationship. Such comparisons include the numbers of boys to the number of girls in a class, or the heights of two people with one stated in metres and the other in centimetres. Rates refer to the comparisons of unlike measures, for example, speed, where there is a relationship between distance and time. Another example would be density, for which there is a relationship between mass and volume.

The term proportion can be used in two ways. It can refer to part of a whole amount, for example, the proportion of the total given to any one person may be a quarter. It can also be used as a ratio, for example, the allocation of money to two people could be in the proportion of 1 : 3. Even though the use of some terminology is inconsistent between teaching and research jurisdictions, and some terms have different meanings in different contexts, it is possible to create a test instrument which is independent of any difficulties due to the various interpretations of these three terms.
<table>
<thead>
<tr>
<th>Year</th>
<th>Proficiencies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Understanding</strong> includes:</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>using fractions and decimals to describe probabilities, representing fractions and decimals in various ways and describing connections between them</td>
</tr>
<tr>
<td>7</td>
<td>recognising equivalences between fractions, decimals, percentages and ratios</td>
</tr>
<tr>
<td>8</td>
<td>connecting rules for linear relations with their graphs</td>
</tr>
<tr>
<td><strong>Fluency</strong> includes:</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>calculating simple percentages, converting between fractions and decimals, using operations with fractions, decimals and percentages</td>
</tr>
<tr>
<td>7</td>
<td>representing fractions and decimals in various ways, investigating best buys</td>
</tr>
<tr>
<td>8</td>
<td>calculating accurately with simple decimals: recognising equivalence of common decimals and fractions</td>
</tr>
<tr>
<td><strong>Problem solving</strong> includes formulating and:</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>solving authentic problems using fractions, decimals and percentages</td>
</tr>
<tr>
<td>7</td>
<td>solving authentic problems using numbers</td>
</tr>
<tr>
<td>8</td>
<td>modelling practical situations involving ratios, profit and loss</td>
</tr>
<tr>
<td><strong>Reasoning</strong> includes:</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>explaining mental strategies for performing calculations</td>
</tr>
<tr>
<td>7</td>
<td>applying an understanding of ratio</td>
</tr>
<tr>
<td>8</td>
<td>justifying the result of a calculation or estimation as reasonable, finding proportions of populations</td>
</tr>
</tbody>
</table>
4.5 Development in proportional reasoning

From a very early age, children have a natural capacity to recognise similarity and they can visualise proportional change of a qualitative nature (Lamon, 2005). They know about enlarging and reducing and can identify distortion when the dilation is not uniform and dissimilar drawings are provided. A child may not be able to quantitatively explain such dilation but “before the subject arrives at the calculation of numerical relations, he isolates an anticipatory schema for qualitative proportionality” and, “at first, the compensation as well as the proportion are exclusively qualitative” (Inhelder, Piaget, Parsons, & Milgram, 1958, p. 315–316).

Research studies with young children contain reports of observations and measurements of their proportional reasoning skills. Children aged six and seven years could identify the reducing sizes of portions as the number of people sharing the whole amount increased (Nunes, Desli, & Bell, 2003). Jeong, Levine, and Huttenlocher (2007) reported that children aged 6–10 years were successful when playing a game using proportions in spinners with continuous regions, though not with spinners divided into separate and countable (discrete) sections. In studying the link between spatial and proportional reasoning skills in children with an average age of seven and a half years, Frick and Mohring (2016) were able to measure the children’s proportional reasoning skills. They created a scale of achievement by asking the children to mark on a line, a level of cherry flavour to match a picture representing the proportion of cherry juice to water.

A significant amount of research has been conducted on the development of proportional thinking in older children and young adolescents, and some of the findings relating to their development in these skills are reported later in the next section. According to Heller, Post, Behr, and Lesh (1990), there are many adults who do not develop sound proportional reasoning and who have difficulty with the basic concepts associated with fractions, rates, and proportions. In a study by Fisher (1988) only six of twenty practising secondary mathematics teachers managed to get all four proportionality problems correct. For one of the four questions, less than 50% of the teachers provided a correct response. The teachers tried to use formal strategies rather than intuitive ones by setting up equations that they did not understand. The teachers also had difficulty with inverse proportion, for example, when they were asked to determine the time six workers would need to mow the lawn given that it would take nine workers five hours, many of their answers were unreasonable.
In another study involving teachers, Cramer, Post, and Currier (as cited in Van Dooren, De Bock, Evers, & Verschaffel, 2009, p. 188) found that 32 out of 33 pre-service elementary teachers set up an equation for proportion, rather than apply sound reasoning, to solve the following problem: ‘Sue and Julie are running equally fast around a track. Sue started first. When she had run 9 laps Julie had run 3 laps. When Julie completed 15 laps, how many had Sue run?’ Lamon (2007) suggested that the small number of adults who are good proportional thinkers indicated that proportional reasoning does not develop automatically but needs to be taught specifically. Lamon’s claim that the development of proportional reasoning is complex, consists of many aspects, and occurs over many years is supported by other researchers (Dole, 2000; Kilpatrick et al., 2001; Tourniaire & Pulos, 1985; Watson et al., 2013). If teachers and other adults are experiencing difficulties in solving proportional reasoning problems, then extra care is essential when such topics are taught to school students.

4.5.1 Aspects of the development of proportional reasoning

For the development of sound proportional reasoning, students need to become proficient in the knowledge, skills, and understandings necessary for calculating with fractions, decimals, percentages, ratios, rates, and linear relationships. To be successful with proportional reasoning, students also need to interpret the nature of the proportion, identify the required operations, and apply their skills with different number types and representations in a variety of situations. These various aspects associated with proportional reasoning provide challenges for students in their learning of these topics, their application of the skills, and hence in their development of sound thinking and reasoning with proportions. The ability to deal with both discrete and continuous quantities, an appreciation of the nature of the unit of proportion, and the recognition of the presence and type of proportion are some of the student challenges that are inherent to the development of sound proportional reasoning.

Discrete and continuous quantities

Findings from research indicate that children’s success with proportioning tasks varies according to the type of quantity. The types of quantities studied are discrete, or countable quantities, and continuous quantities, for example, area and volume, which are quantities with an infinite number of possible values. Young primary children were found to be more successful with continuous quantities than with discrete, and it was suggested that their inability to count accurately at an early age may have contributed to their lower performance with the latter (de la Torre, Tjoe, Rhoads, & Lam, 2013).
Similar findings were reported in other studies where children were presented with coloured bars representing proportions of juice and water (Boyer & Levine, 2015; Boyer, Levine, & Huttenlocher, 2008). The bars were either undivided except for the two colours (continuous representation) or divided into sections (discrete representation). Children were successful in recognising bars with similar proportions for continuous quantities from about six years of age, but it was not until about the age of 10 years that they developed efficiency with discrete quantities. In the study using these coloured bars, an experiment to improve student problem-solving with discrete quantities resulted in improved performance by the Grade 4 students but not by the students in kindergarten and Grade 2.

Nunes and Bryant (2009) found that the type of quantity did not affect performance when young primary students were asked to decide if the proportions received were equal, but it did influence the outcome when the students were asked to quantify their answers. Their results indicated that students were more successful when quantifying discrete rather than continuous proportions, and this was attributed to the possibility of being able to count out the shares rather than having to identify them as ratios or fractions. Determining three quarters of a bar of chocolate could be achieved by breaking the bar into four pieces and selecting three of them; sharing three cakes between four students was more difficult.

Watson et al. (2013) attributed student success with discrete quantities to their experiences of using repeated addition for multiplication. By comparison, Fernandez et al. (2011) found that student performance was unaffected by the type of quantity. The students were using variables which represented discrete and continuous quantities and the authors concluded that there was little difference in the mathematical thinking and the computation needed to complete the tasks.

The research findings of the performance of students on tasks using discrete and continuous quantities are inconsistent. It is possible that the ages of the children, their prior experience, the difficulty of the numbers used, and their ability to count, all influence their success with comparing proportions of discrete and continuous quantities.
**Unitising**

An important aspect of developing sound proportional reasoning is the ability to unitise. There is scant reference in the literature to the influence that the ability to unitise has on student success. Lamon (1996, p. 170) described unitising as the “cognitive assignment of a unit of measurement to a given quantity,” and in a later study, Lamon (2005) claimed we have a natural tendency to unitise. As students grow in experience and understanding, they can form and apply increasingly complex units by combining single units into multiples, or by thinking of measures in more flexible ways. Students can translate costs per kilogram to costs per 100 grams and they can convert, possibly with the aid of technology, costs per 440 grams to costs per 100 grams. The ability to perform these conversions and to recognise when to use such conversions is an asset for the development of proportional reasoning.

*Proportional and non-proportional situations*

The earlier reference to the inability of pre-service teachers to identify the track-running item as a non-proportional task suggests that students would experience the same difficulty. Research in the recognition of proportionality, and the provision of proportional responses to items that were non-proportional in nature has been described by Fernandez et al. (2011), Tjoe & de la Torre (2014a), and Van Dooren et al. (2009). In the study by Tjoe and de la Torre the responses of over 400 Grade 8 students to four MC items were analysed. In the two items testing the application of proportions, over 90% of the more proficient group were successful, but in the two items which required students to distinguish between proportional and non-proportional situations, these capable students were much less successful: with only 32% and 44% correct for the two items. This latter performance did not differ much from that of the less proficient students who were far less successful in the application items.

The extent to which the students distinguished between proportional and non-proportional situations was influenced by the numbers used and the nature of the ratios in the items. The presence of non-integer ratios appeared to trigger a proportional approach even when the task was non-proportional. The lack of success suggested that the students did not fully understand the task and they had not considered the context for which it was written. It is possible that they had not been exposed previously to questions of this nature. Lamon (2005) suggested that, rather than immediately try to set up a statement of equivalent ratios, students should consciously analyse problems, start by deciding if the change is proportional or not, and then decide if both amounts increase together or if one increases while the other decreases.
**Intensive and extensive quantities**

When proportional thinking is presented in terms of relationships of relationships, it may not be easy to conceptualise the various quantities and relationships. Piaget (1952) described the differences between extensive and intensive quantities, which can both be individually represented by a single number, in terms of the ability to add such quantities. Extensive quantities include length and height: measures of length or height can be added. By comparison, the sum of intensive quantities is not defined: 15°C + 25°C is not equal to 40°C. Intensive quantities are often composed of two extensive quantities, for example, speed, which has two variables, time and distance (Kaput & West, 1984). It often does not make sense to add the numbers representing different speeds. A car going 10 km per hour then 20 km per hour does not travel at a total of 30 km per hour.

Nunes, Desli, and Bell (2003) distinguished two types of intensive quantities according to the wholeness of the variables when considered together. First, the components can be considered as separate items, and as a complete unit. An example of this occurs when sugar is combined with lemon juice, and the resulting taste (the whole) depends on the relationship between the quantities of each component (lemon and sugar). Second, the components do not really combine, and the outcome of being put together is described in terms of both components; for example, a speed of 60 kilometres per hour involves a reference to both the distance covered and the time taken. Similarly, the ratio of males to females, where two-thirds of a class are female, includes the proportional statement (two-thirds) plus the unit (males and females).

Lamon (2005, 2007) described intensive quantities, for example, slope and density, as neither directly measurable nor observable. Howe, Nunes, and Bryant (2011) suggested that all intensive quantities can be represented as ratios, and a subset of these can be represented as fractions. For sound proportional reasoning, the development of the skills and knowledge for using and understanding intensive quantities is essential. Only when students can identify the relationship within the component parts, such as in the ratios or rates, can they appreciate the need to think multiplicatively when using intensive quantities. For many students, this understanding of the relationship and the ability to represent and calculate with the numbers develops over considerable time. Identifying calculations involving intensive quantities can be quite complex, but it becomes much easier when understanding of the relationship between the component parts develops.
Within and between ratios

Within a statement of equivalent ratios further relationships can be described. For the ratio of \( m : k = w : b \) there is a relationship between \( m \) and \( k \) which is the same as the relationship between \( w \) and \( b \) and both of these relationships are referred to as being within-state. There is also a relationship between the matching parts of each ratio, that is, the relationship between \( m \) and \( w \) which is the same as between \( k \) and \( b \), and these relationships are referred to as being between-state (Noelting, 1980b). To establish the equivalence of the ratio \( m : k = w : b \), a scaling operation is applied to a within-state relationship using a between-state strategy. There is a scalar relationship between the within ratio components, and a functional relationship between the between ratio components. (Küchemann, Hogden, & Brown, 2014).

According to Küchemann et al. (2014), Year 8 students were more likely to recognise and apply proportional reasoning for numeric problems when the within-state ratio was more easily recognised than the between-state ratio. Two versions of each problem were written for their study and for one problem these versions are summarised as follows:

For 11 people, the recipe required 33 ml sauce. How much for 25 people?
For 11 people, the recipe required 25 ml sauce. How much for 33 people?

Students found it much easier to use the second form (within-state ratio) where there was scaling up from 11 people to 33 people (91% correct) than the first form where there were 3 ml per person and the relationship between people and sauce was functional (51% correct). It was the reverse outcome for two versions of an item on scaling a geometric figure where both versions also had the same numbers. When between-ratio scaling was used, and the scale was easily recognised, 75% of students were correct, but only 36% were correct for the within-ratio version of the same item.

Direct and inverse proportion

There are two types of proportional relationships, direct and inverse. In direct proportion, as one variable increases so does the other, for example, as the number of kilograms of apples purchased increases, so does the cost. This direct proportion relationship is linear, and this type of relationship is generally easier to represent using worded descriptions or symbols than an inverse relationship. Inverse proportional relationships involve one of the variables decreasing as the other increases; for example, when the time taken to do a shared task decreases as the number of people available to do the task increases.
While working with 105 children aged 6–8 years, Nunes et al. (2003) found that the students had difficulty with questions about inverse relationships with both discrete and continuous quantities, even though they were only asked for qualitative answers. In a further aspect of their study which involved 113 children with an average age of 8.5 years, Nunes et al. (2003) found that the children’s problems with inverse proportions were influenced by the nature of the quantities used. They concluded that the tasks involving intensive quantities were more difficult for the students than those where the quantities were extensive.

Similar confusion with direct and inverse proportion was reported by Arican (2018), de la Torre et al. (2013), Fisher (1988), and Tjo & de la Torre (2014b). In Arican’s study, eight pre-service mathematics teachers were interviewed about their strategies for solving nine problems with proportions where there was a missing value. The most common error of the teachers was to use a direct proportion strategy when there was an inverse proportional relationship. Two of the teachers also used an inverse proportion strategy when there was a direct relationship. Recognising the nature of the proportional change, even if unable to identify the type, is a logical precursor to identifying a solution process. Being able to determine if an answer should be greater or less than one of the numbers given provides insight for recognising a reasonable answer and for checking that the answer is correct.

4.5.2 Development pathways

Throughout the literature there are varying descriptions of the order in which students develop proficiency in the skills and understandings necessary for the demonstration of sound reasoning and thinking about proportions. Most descriptions located in research reports are confined to a subset of the necessary skills associated with interpreting and calculating with proportions; for example, fractions, decimals and percentages. Descriptions of developmental (or learning) pathways for most of these aspects are likely to provide detailed insight into student growth in just one skill or understanding but are unlikely to provide detailed insight into student growth in the broader concept of proportional reasoning.

Learning pathways for proportional reasoning which are focussed on interpreting problems involving ratios are different and can indicate to some extent the development that occurs. The study of these, and other pathways of learning can provide valuable insight into the nature and timing of student development of mathematical behaviours. Such insight can be used to identify a hierarchy of concept difficulty and hence to guide item development.
**Lamon’s suggestion for the development of proportional reasoning**

According to Lamon (2005, p. 9), the concepts integral to proportional reasoning develop with experience and over a long period of time “in a web-like fashion, rather than in a linear order.” Aspects of the development of proportional reasoning are interdependent and progress in one aspect affects the development of other aspects. As each student’s knowledge of rational number increases, so does their ability to reason, measure, unitise, and recognise covariation. The growth in these aspects then supports further development of the student’s rational number skills and understandings. Lamon’s (2005, p. 9) representation of these aspects and their interconnections is provided in Figure 4.1.

![Diagram of proportional reasoning](image)

**Figure 4.1** Key aspects of proportional reasoning (Lamon, 2005, p. 9)

Parish (2010) supports Lamon’s suggestion of the non-linear nature of development in proportional reasoning. From interviews with a small number of children from Grades 5 to 9, Parish was able to provide general descriptions of seven possible stages of learning about fractions as ratios. These stages included the use of additive thinking (using addition when multiplication was required), a pre-proportional approach, scalar and functional reasoning, and the use of algebraic methods to represent and solve complex problems. Further research to confirm this order of skill acquisition, and to provide evidence that the progression applies to a greater number of students of varying ages, would provide better understanding of student development in this content area.
Contributions from Piaget and Inhelder

Experiments conducted by Piaget and Inhelder have provided considerable insight into stages of development in some of the aspects of proportional reasoning, including proportions in measurement and number, as well as probability (Hart, 2004; Lovell, 1971; Inhelder et al., 1958; Piaget & Inhelder, 1975). Investigations of children’s attempts to identify rectangles which were the same (similar) indicated that there were four stages of development. Initially the effort was deemed not to be serious; second, the child enlarged without measuring; third, the child could recognise rectangles that were similar but not reproduce them; and finally, the child could scale the rectangle by a fractional amount.

Children showed evidence of similar stages of development when they were asked to indicate the number of sprats (discrete quantities) or fishfingers (presented as continuous quantities) which were needed for an eel that was proportional in size to another eel. The eels were 5, 10 and 15 centimetres in length. Initially the children were able to identify the direction of the change, and later they made some attempt to quantify the number of sprats needed before developing the ability to determine the true proportions.

Piaget and Inhelder (1975) reported that children aged 7–12 years could not envisage the equality of two ratios, that they had a qualitative intuition of proportionality, and only later could they see true proportional relationships. Not until the ages of 12–15 years could the children generalise to the abstract and establish the equation representing the equivalence of two ratios (Inhelder et al., 1958).

The conclusions of Piaget and Inhelder (1975) were based on conversations with a small number of children, and the stages of development that they formulated are not described in later research as learning pathways. These findings of Piaget and Inhelder have provided researchers with sound ideas for further investigation into learning pathways and with experiments on which to model tasks for the collection of information about student ability in proportional reasoning. In many of the studies of learning pathways described in the following section, the work of Piaget and Inhelder has been cited. However, the nature of the questions asked, the format for collecting students’ responses, and the extent of interviewing individual children in the more recent studies are notably different to the methods used by Piaget and Inhelder.
**Noelting’s experiments**

In an early study to determine if the development in the understanding of ratios and reasoning with proportions is hierarchical, Noelting (1980a, 1980b) conducted a series of experiments with 321 students aged 6–16 years using the perceived taste of orange juice of different concentrations. Noelting’s description of the stages of development in the conceptual understanding of ratios was ordered on the student’s age. This was identified as the time that 50% of the students were correct on the item. The features of each stage were described in terms of the student’s ability to identify, describe, and compare different ratios.

The proposed stages of development identified by Noelting (1980a, p. 231) are summarised in Table 4.5, where the final columns show an example of the two ratios (presented as squares and circles) that students at the corresponding stage could successfully compare.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Age</th>
<th>Characteristics of Stage</th>
<th>Typical comparison of A and B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>Identification of elements</td>
<td>( \bigcirc ) ( (1,0) ) ( \square ) ( (0,1) )</td>
</tr>
<tr>
<td>IA</td>
<td>3</td>
<td>Comparison of first terms only</td>
<td>( \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc ) ( (4,1) ) ( \square \square \square \square \square ) ( (1,4) )</td>
</tr>
<tr>
<td>IB</td>
<td>6</td>
<td>Same first terms, compare second terms</td>
<td>( \bigcirc \bigcirc ) ( (1,2) ) ( \square \square \square \square ) ( (1,5) )</td>
</tr>
<tr>
<td>IC</td>
<td>7</td>
<td>Inverse relation between terms in the ordered pair</td>
<td>( \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc ) ( (3,4) ) ( \square ) ( (2,1) )</td>
</tr>
<tr>
<td>IIA</td>
<td>8</td>
<td>Equivalence class of (1,1) ratio</td>
<td>( \bigcirc ) ( (1,1) ) ( \square \square \square \square ) ( (2,2) )</td>
</tr>
<tr>
<td>IIB</td>
<td>10</td>
<td>Equivalence class of any ratio</td>
<td>( \bigcirc \bigcirc ) ( (2,3) ) ( \square \square ) ( (4,6) )</td>
</tr>
<tr>
<td>IIIA</td>
<td>12</td>
<td>Ratios with two corresponding terms multiple of one another</td>
<td>( \bigcirc \bigcirc \bigcirc ) ( (1,3) ) ( \square \square \square \square ) ( (2,5) )</td>
</tr>
<tr>
<td>IIIB</td>
<td>15</td>
<td>Any ratio</td>
<td>( \bigcirc \bigcirc \bigcirc ) ( (3,5) ) ( \square \square ) ( (5,8) )</td>
</tr>
</tbody>
</table>
In the early stages of development, according to Noelting, students could accurately compare ratios where one of the components is the same in both ratios. They could also compare ratios when the first ratio has more of the first component than the second, and the second ratio has more of the second component than the first. At a later stage, students were successful with ratios that are multiples of each other and then later they could compare ratios with different numbers of components and with non-integer scaling. Karplus et al. (1983) proposed a progression in the application of ratios which was similar to the findings of both Noelting and Piaget but suggested that there should be separate progressions for the within-state and the between-state comparisons. Karplus et al. (1983) also proposed a hierarchy of the strategies which the students selected and applied after recognising the type of proportionality.

Hierarchies of learning about ratio and proportion

The strategies used by two classes of Grade 5 students to solve for $x$, equations like \( \frac{8}{12} = \frac{42}{x} \), were investigated by Steinthorsdottir and Sriraman (2009) during their study of the four-level developmental trajectory for ratios which had been proposed by Carpenter et al. (as cited in Steinthorsdottir & Sriraman, 2009). At the lowest level of the learning trajectory, if the students provided any purposeful calculations, the focus was on the difference between numbers. At Level 2, the students treated ratios as a single unit, and they were able to scale them up and down using whole numbers. Only at Level 3 could students scale the ratio by fractions or mixed numbers, and the ratio was still treated as a single unit. At Level 4 students could see the within-state and between-state relationships within the two ratios in the equation, determine the easiest computations, and conceive of the scale in abstract terms.

Hart (2004) used the responses of over 2000 students aged 13–15, to 27 items on ratio and proportion, to determine the four levels of a hierarchy of student understanding of rates and ratios. At Level 1, the students could use simple integer multiples such as 2 or 3, and they were able to determine halves of quantities. At Level 2, the rates were deemed easy to find, or could be determined by finding half of the original amount and adding it on to the starting value. By Level 3, the students could solve problems with non-integer rates such as \( 1 \frac{2}{3} \) if set in contexts that prompted a unitary approach (recognising the quantity for one unit), or that allowed building up one part at a time.
At the highest level, students were able to use non-integer ratios like 5:3 and solve scaling problems where the unitary method was not meaningful. Commenting on this hierarchy of development, Hart (2004, p. 100) also noted the “very slow progression of attainment from year to year.”

Applying Rasch Measurement Theory to the responses of 212 students to items on proportional reasoning, Misailidou and Williams (2003) investigated the levels reported by Hart (2004). They were able to extend Hart’s descriptions to include typical errors at each of the levels. Five levels were identified and the first and last of these were defined by the absence or extension of the skills at Levels 1 and 3.

At the lowest level of the proposed hierarchy, students could halve and double on easy items, but they provided incomplete reasons or used addition methods to solve problems. Additive strategies were still prominent at Level 2, where students also tried to halve and double in the wrong situations, and where the numbers formed quite difficult ratios. Even at the highest level, students could be prompted to use additive strategies when they should have been using multiplication. A summary of these errors as they relate to the different levels of difficulty is provided in Table 4.6.

Table 4.6 Developmental hierarchy (Misailidou & Williams, 2003)

<table>
<thead>
<tr>
<th>Level</th>
<th>Typical performance</th>
<th>Typical errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Answers found by multiplying by 2 or 3 or by halving Use scalar ratio of single digits</td>
<td>Incomplete reasoning Additive strategy on easy items</td>
</tr>
<tr>
<td>2</td>
<td>Use scalar and functional ratio Simple multiplication on scalar ratios Can scale with 1.5 by adding half the amount</td>
<td>Halving and doubling when it is inappropriate Constant sum Frequent use of additive error High ability students at times use the incorrect build up method</td>
</tr>
<tr>
<td>3</td>
<td>Can use large numbers and fraction operators Can work with unfamiliar context</td>
<td>Predominant use of additive strategy with items provoking it Higher ability pupils only err with the incorrect application of the build-up method</td>
</tr>
</tbody>
</table>
The descriptions of children’s development in their understanding of proportional reasoning, which have been provided by Steinthorsdottir and Sriraman (2009), Misailidou and Williams (2003), and Hart (2004) are very similar at the lower levels of the hierarchies. Even though the descriptions are slightly different at the higher levels, the skills that are typical of those levels are similar. The general trend is that the student progresses from using doubling and halving to scaling with small whole multiples before managing non-integral scaling. Simultaneously, the student starts with additive strategies before developing a sense of, and an ability to deal with, a multiplicative approach to solving problems where the relationship between the quantities is proportional.

Learning pathways for concepts associated with proportional reasoning were described by Long (2011) in terms of the problems that students were able to solve. Seven levels of problems were identified when Rasch Measurement Theory was applied to the students’ responses. The mathematical structure, notation, type of sharing, and range of numbers used, all influenced the level to which an item was allocated. Items at the lowest level required students to determine $x$ in a ratio statement similar to $1:3=9:x$ where the numbers were less than 30. An item at level 5 was a multi-step problem containing a proportional relationship involving distance, time, and speed. For that calculation, students needed to determine $x$ in a ratio statement similar to $120:x=1.5:1$ where the numbers were less than 160. Greater detail of the behaviours of students at each level would provide useful insights for teaching and learning, and further investigation of this hierarchy is warranted.

*Learning trajectories and learning progressions*

Descriptions of student development in conceptual understanding are sometimes referred to as *learning trajectories* (Confrey, Maloney, & Corley, 2014). These trajectories are developed independently of curriculum and achievement standards and are influenced by the instruction received by the students. The identification of a learning trajectory starts with descriptions of the students’ behaviours, based on observation and the research literature. The behaviours are then confirmed using empirical evidence from the students’ responses to graded tasks. The stages of the learning trajectories do not represent levels of achievement that students will master in a specific order, but rather are descriptions of behaviours that students are probably able to demonstrate at a time in their learning.
In the equipartitioning learning trajectory (producing equal-sized groups of parts when sharing) described by Confrey et al. (2014), there are 16 levels of proficiency mapped to 13 different types of behaviours. Equipartitioning is seen as fundamental for learning about fractions and ratios, and the trajectory describes student learning from Kindergarten to Grade 8. Seventeen other learning trajectories for mathematics have been proposed; such learning trajectories can be used to inform standards and to identify appropriate tasks to move students along the developmental continuum.

For the formulation of learning progressions, which are like learning trajectories in that they describe growth in specific knowledge and skills, Briggs and Peck (2015) identified a specific process. for the considered domain, for example, proportional reasoning, first, student learning is interrogated and recorded, and the relevant literature is investigated. A learning progression is then proposed, and evidence is gathered using tests, class studies, or student interviews, to refine and confirm the suggested pathway. Linking theories of learning with empirical studies allows descriptions of student growth to be well founded.

Briggs and Peck (2015) analysed theories of the development of student skill in dealing with proportions and proposed a theoretical progression of learning for proportional reasoning. They described the progression in terms of student attributes, and they provided examples of typical activities performed well at each level. The developmental pathways described previously refer to the component skills for developing proportional reasoning, but the learning progression from Briggs and Peck has a greater focus on proportional reasoning as a single concept. Their description of the skills and understandings needed for development in proportional reasoning are more general than those in the progressions described previously.

At the lower levels of the progression, the focus is on the student’s ability to share, and then at the higher levels the focus shifts to equivalence and relationships. At the highest levels of the progression, students can theoretically solve a variety of problems with different rates and ratios. In comparison with the pathways described earlier, the progression proposed by Briggs and Peck (2015), as shown in Table 4.7, extends over a greater age range and includes consideration of spatial reasoning and linear relationships. A more detailed indication of the types of questions to which students could successfully respond at each level of this progression is provided in Appendix 4.3. Briggs and Peck suggested that the development of learning progressions is necessary for the design of vertical scale scores to support criterion-referenced and norm-referenced interpretations of achievement.
Table 4.7    Student attributes in a learning progression (Briggs & Peck, 2015)

<table>
<thead>
<tr>
<th>Level</th>
<th>Grades</th>
<th>What students know and can do</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>K-2</td>
<td>Justify equivalence of shares by counting Name shares from a collection using intensive or extensive units</td>
</tr>
<tr>
<td>2</td>
<td>1-3</td>
<td>Recognize naming shares needs number and size of pieces Use geometric / measurement ideas to justify equivalence of shares</td>
</tr>
<tr>
<td>3</td>
<td>1-5</td>
<td>Understand $n$ times as much and $n$ times single share gives all Justify equivalence by reconstructing whole from given part</td>
</tr>
<tr>
<td>4</td>
<td>3-5</td>
<td>Recognize qualitative compensation as inverse relationship Distinguish multiplicative and additive relationships Use transitivity (e.g. if $a &gt; b$ and $b &gt; c$, then $a &gt; c$) in explanation</td>
</tr>
<tr>
<td>5</td>
<td>3-5</td>
<td>Name fair shares in multiple ways and explain their equivalence Explain equivalence of different methods of creating fair shares Use and justify $p$ by $n = p/n$</td>
</tr>
<tr>
<td>6</td>
<td>4-7</td>
<td>Build one ratio from another Scale up or down a ratio in a single step Use ratio as multiplicative comparison Name unit rates</td>
</tr>
<tr>
<td>7</td>
<td>6-8</td>
<td>Predict for linear relationships $y = mx + b$ Calculate unit rate Interpret slope as rate of change</td>
</tr>
</tbody>
</table>

Learning pathways, progressions, hierarchies, and trajectories are all terms used to describe the progress of conceptual development. These descriptions usually refer to the mathematical behaviours shown by students at times in their learning, and they allow the hypothesised growth from grade to grade to be easily visualised. Descriptions of key activities and student behaviour at the various levels provide support for the creation and ordering of items that could be used to measure student proficiency. The descriptions also provide information for the design of items, which could be presented to students across two or more levels, or which could be used as linking items.

After establishing learning progressions, assessment items can be purposefully written at each identified level of achievement and, by varying the context, or types and values of numbers used in the items, at different levels of difficulty. The ability to identify stages of conceptual development is beneficial for teachers who could use the information to plan relevant learning activities for the students: activities to cater for the students’ current stages of development; and to provide pre-requisite learning for the stages of development to follow.
4.6 Challenges for learners

Developing sound proportional reasoning is challenging for many students. It not only involves concepts which are difficult to master, but it is also associated with difficulties which occur when the student has problems with mathematical literacy and mathematical language conventions. Students need to know terms and notations specific to mathematics, and they must be able to read and interpret the learning activities and assessment tasks which are used to further their development. It is more challenging for students to use a three-part ratio, for example, 2 : 4 : 5, than it is to use 2 : 4 and 4 : 5 as single ratios. Students need to use the mathematical language which is appropriate to the domain of proportions and relevant to the skills associated with understanding proportions. Students need to distinguish between a fraction of and as a fraction, a quarter as much and a quarter of, between an increase of 10% and an increase of $10.

It is necessary for students to know the relevant symbolic language and the appropriate use of such language. They need to understand and use appropriately: the % symbol, the colon separator for ratios, fraction representation, and variables to represent changing values. Dealing with the intended use of fractions can be difficult. In some situations, the ratio 2 : 3 is represented as the fraction \(\frac{2}{3}\) without any context, and in other situations the same ratio, 2 : 3, can indicate an amount is divided into two unequal parts with relative sizes of \(\frac{2}{5}\) and \(\frac{3}{5}\). Neither interpretation makes sense without some description of the situation to which the ratio applies. With fractions, students may see the two numbers used for the numerator and denominator as two separate numbers without appreciating that these numbers form a relationship with a very different value to either of the actual numbers.

Challenges such as these are just some of the many difficulties faced by students as they learn about proportional reasoning. This is not surprising when they understand only part of the instruction about a concept, and as a result learn only part of the knowledge and skills presented. A further explanation of what it might involve having only part understanding of a concept is discussed in the following chapter.
CHAPTER 5: Partial knowledge

From the research literature, descriptions of partial knowledge, also referred to as partial information, appear to relate to one of three focus areas: first, the student selection of options; second, the quality of the distractors; and third, the acquisition of knowledge. For the first focus area the demonstration of partial knowledge is classified according to student behaviour in the selection of options to complete MC items. The award of partial credit is then based on the scoring methods associated with the different behaviours. The second focus for describing partial knowledge is on the creation and quality of the distractors (incorrect responses), and the allocation of credit for the selection of options with information. These informative options, although incorrect, contain some, but not all, aspects of the item’s correct response and their selection may indicate that the student has demonstrated partial knowledge.

The third focus for describing partial knowledge is on the development of conceptual knowledge and understanding. Awareness of this development can assist MC item writers to design items which contain distractors with information, and to construct items with particular levels of difficulty. Insight into student development of understanding the key concepts in proportional reasoning would facilitate the identification of what might constitute partial knowledge in proportional reasoning. This can support the creation of items that allow credit to be awarded for this partial knowledge. The allocation of credit for partial knowledge should generate more accurate measures of student achievement, provide more reliable rankings of performance, and give greater insight into student learning.

5.1 Selection of options

Some of the scoring techniques which have been designed to discourage guessing assume that all guessing is totally random: that the student does not know the answer; has no partial knowledge of the item’s content; and is selecting options at random (Budescu & Bar-Hillel, 1993; Burton & Miller, 1999). If the student does have some knowledge of the content of the test and of the individual items, but does not know the correct response, they could be using some system to guide their selection of options. As their knowledge increases, students are expected to guess less often and with greater accuracy (Stewart & White, 2011). Being able to distinguish between systematic and random guessing, and adjust the scores accordingly, would allow credit to be given for partial knowledge, thus increasing the reliability and validity of the scores (Alnabhan, 2002).
The answer-until-correct format for MC items is an effective way of determining a student’s partial knowledge according to Slepkov, Vreugdenhil, and Shiell (2016). The student is allowed to select one option at a time, receive immediate feedback, and continue in this manner until all possibilities are exhausted. The more attempts the student needs to identify the correct response, the lower their score. According to these authors, this method does not require an assessment of the relative merits of the incorrect options, nor an interpretation of the student’s level of knowledge. The authors claimed that their use of this technique with chemistry undergraduates was an effective way of identifying where partial credit can be allocated for partial knowledge when measuring the level of student understanding.

In some studies, participants could select more than one option when they were not sure of the correct answer (Alnabhan, 2002; Bush, 2001; Bo, Lewis, & Budescu, 2014; De Laet et al., 2015). If just the correct response was selected, their knowledge was classified as complete. If they selected a few options, including the correct response, then their knowledge was classified as partial and they were awarded part marks.

The elimination of some, but not all incorrect options in a MC item is considered to be a demonstration of partial knowledge (Bar-Hillel, Budescu, & Attali, 2005; Bond et al., 2013; De Laet et al., 2015; Espinosa & Gardeazabal, 2010; Lau et al., 2011). However, the elimination of an option may only be classified as true partial knowledge if the test participant is certain that the option is incorrect (Ben-Shakhar & Sinai, 1991; Burton & Miller, 1999). In this case, the participant who has not identified the correct response and who is certain that one or more distractors are incorrect, might have guessed between the remaining options. Participants using this systematic guessing may not be deemed to have demonstrated true partial knowledge. Bond et al. (2013) considered the elimination of all incorrect answers as full knowledge, and a removal of a subset of incorrect answers as partial knowledge. They defined the elimination of the correct answer as misinformation which was called complete or partial according to the number of distractors eliminated.

Correct answers can be learned, recognised, or rationalised, and they can be selected with varying levels of confidence (Lindquist & Hoover, 2015). Being able to score all degrees of full and partial knowledge is a challenge, which if even only partly successful, would enhance the validity of the required measures. Determination of partial knowledge is limited when only student behaviour in the selection or elimination of options in MC items is considered. The quality of the distractors also needs to be considered.
5.2 Informative distractors

The distractors are an important part of MC items, and their quality influences the difficulty of the item. For items to prompt accurate recall and thinking, and thus reveal different levels of proficiency, the distractors need to be plausible by addressing common misconceptions, by being logically or aesthetically appealing to the examinees, or by containing some aspects of the correct response. Distractors which contain some, but not all aspects of the correct response are referred to as informative distractors or distractors with information (Andrich & Styles, 2011; Sideridis et al., 2016). The selection of informative distractors, which reflect partially correct responses, indicates that those who choose such options have less proficiency than those who select the correct response, but greater proficiency than those who select non-informative distractors as the correct response.

Scoring existing distractors

Awarding credit for partial knowledge is seen as an improvement on dichotomous scoring (1 for a correct response and 0 otherwise). One way to grant such credit is through the application of empirical option weights, which involves the allocation of a range of scores for a single item (Diedenhofen & Musch, 2015). To calculate the weight for each option, the correlation between the total score and the frequency of that option, is estimated from the current test or a previous one, and then weights are applied to the scores for those responses. Diedenhofen and Musch reported increased validity and reliability with the use of empirical weights for MC items where the options reflected different levels of knowledge. They recommended that analysts should continue to use these empirical option weights, and that creators of MC items should write options which test blatant misconceptions, or which address different levels of knowledge. Such recommendations may not be suitable for easy items where there is little discrimination between test participants. This is one reason which may explain the lack of recent research in this area.

Another scoring method in which partial knowledge can be credited was proposed by Bo et al. (2014). The participants’ proficiencies would be determined from the information provided by the responses to the distractors as well as to the correct options. This method involves the calculation of Hamming distances and could also be used when there are multiple correct options within the item. Hamming distances relate to the count of the classifications of the options when the participants are instructed to mark two options.
While the simulation study by Bo et al. (2014) indicated that the process appeared to provide improved estimates of participant proficiency, extra time and expertise would be needed to create items with two equally correct options for an empirical study. It could be challenging for the students to respond to items with two correct responses, and in other items to select two options, particularly when they recognise one of them is incorrect. It would be a difficult task to create items which would suit this type of analysis and it has not been possible to locate sample items showing that this has been achieved in mathematics.

*Deliberate construction of distractors with information*

Smith (1987) created 50 items containing distractors with information. The distractors were written to reflect the theoretically different stages of vocabulary acquisition. For each item the students needed to select the synonym for the stem word, and the scores available ranged from 0 (selecting a word that sounded or looked like the stem) to 4 (correct response). Other scores included 2 for selecting the antonym and 3 for being nearly correct. Using responses from 1400 college students and adults, Smith compared this method of awarding partial credit for distractors with information, to dichotomous scoring. The analysis supported Smith’s hypothesis that there was information about the participants’ learning in the incorrect responses. It is possible that this partial credit method of scoring worked well because all the items had a similar format, and the different options reflected the different numbers of steps that the students had completed to identify that option as the correct response.

In a study of student responses to MC items, Briggs et al. (2006) designed each option to correspond with developmental levels of understanding. First, they created a construct map which displayed levels on the learning hierarchy and descriptions of the expected learning at each level. Their construct map for the Earth and Solar System was created using the research findings of student understanding of the content. This map was then used to inform the construction and editing of the ordered multiple choice (OMC) items with each option linked to a particular level on the map. The report showed how examining the selection of options could be used to identify the level of student understanding. The success of using OMC items to gauge the different levels of student achievement relies on a sound understanding of the construct being measured, its hierarchical nature, and the stages by which the student progresses along this hierarchy. Not all constructs are suitable for the identification of levels of student understanding, and it would seem that for proportional reasoning, such fine grained recognition of the levels of development is yet to be achieved.
Another way of constructing informative distractors is through the use of cognitive diagnostic models. These models are “developed specifically for the purpose of identifying the presence of multiple fine-grained skills,” (de la Torre, 2009, p. 163) and hence to provide more diagnostic information from individual MC items. Different numbers and qualities of these skills or attributes are addressed in each option. For the MC item provided by de la Torre, students needed to recognise the solution to the subtraction of a fraction from a mixed number. Each option addresses different attributes, thus allowing different diagnostic information to be identified in the students’ responses. The correct response is Option D and students have borrowed from a whole number, subtracted the fraction, and simplified the fractional answer. Only one of these attributes is addressed if the students select Option A (subtract the fraction), but two attributes are addressed with the selection of Option B.

This item, as shown in Figure 5.1, appears to provide opportunities to detect partial knowledge, but this would depend on how the item is used and scored. The stem does not include an instruction to provide a simplified fraction, and one could argue that Options C and D are both correct and deserve equal credit. Options A and B are likewise equal in value, but the model suggests that the selection of Option B would attract a higher score than the selection of Option A. Options B and C have the same number of attributes but the quality of these is quite different and it would not be appropriate to award the same score for their selection. This discussion of just one item which has been designed to capture a “subset of the knowledge state that corresponds to the key” (de la Torre, 2009, p. 168) indicates that is difficult to construct items where each option reflects a different level of student learning.

<table>
<thead>
<tr>
<th>Item with options</th>
<th>$\frac{4}{12} - \frac{7}{12}$ =</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. $\frac{3}{12}$</td>
<td>B. $\frac{1}{4}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attributes or fine-grained skills</th>
<th>Borrow from a whole number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subtract the fraction</td>
</tr>
<tr>
<td></td>
<td>Simplify the fraction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Options</th>
<th>A. addresses only 1 attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B. and C. address just 2 attributes</td>
</tr>
<tr>
<td></td>
<td>D. addresses all 3 attributes</td>
</tr>
</tbody>
</table>

*Figure 5.1* MC item addressing different fine-grained skills
In a later study, Tjoe and de la Torre (2014b) identified six measurable attributes of proportional reasoning and then used these to create 10 MC items which addressed one or more of these attributes. The attributes included comparing fractions, constructing ratios, and distinguishing proportional from non-proportional relationships. Middle-school students’ responses to these items were analysed and the students were interviewed about their solution processes. The results indicated that the proposed attributes were demonstrated and validated. For this 2014 research the attributes were not at the same fine-grained level as in de la Torre’s (2009) study and any partial knowledge demonstrated with the selection of different options was not considered. The processes used by Tjoe and de la Torre (2014b) could be adopted to identify the many fine-grained skills needed to write MC items for proportional reasoning, but this would require considerable expertise in the construct and in the knowledge of student development in these skills.

5.3 Development of knowledge

The determination of estimates of student proficiency based on awarding credit for partially correct or even incorrect answers seems intuitively unacceptable, but many research studies provide evidence to justify such action. The acquisition of knowledge is described earlier as resulting in the ability to recall facts, terminology, procedures, concepts, principles, and mathematical structure, or to recognise them because the information is familiar (Bloom et al., 1956). As student knowledge develops, errors, misconceptions, or alternative conceptions at a higher level of conceptual understanding may also develop. In a review of the literature on student conceptions in mathematics and science, Confrey (1990, p. 42) reported that the development of misconceptions, which were widespread and difficult to remediate, was “unavoidable and necessary in the development of knowledge.”

In a later evaluation of the relevant research, Smith, diSessa, and Roschelle (1993) reviewed the development of student misconceptions, also in mathematics and science, and noted that intermediate states of understanding existed. They suggested that misconceptions form when existing knowledge is overgeneralised or is applied in inappropriate situations. An example of this occurs when students believe that 0.12 is less than 0.109 because they are applying their knowledge of whole number order to decimals. Smith et al. concluded that novice and expert knowledge have much in common, and in fact, well-established misconceptions have “their roots in productive and effective knowledge” (p. 24).
More than just an error

Ben-Zeev and Sternberg (1996) described the production of errors by overgeneralisation and overspecialisation. When learners encounter new situations which they cannot resolve, they resort to implement familiar rules that have only some of the features relevant for the new situations. Children learning to count may initially persist with twenty-eight, twenty-nine, twenty-ten; they are accurately applying their knowledge of counting without knowing the need to transition to thirty. In the example provided by Smith et al. (1993), the students have also overgeneralised their understanding of number.

Ben-Zeev and Sternberg (1996) and Ben-Zeev (1998) provided other examples of rational errors made by students. The errors are described as rational because they have their origins in situations where they make sense. Students are inspired to create their own rules from something they have learned, or they may apply an algorithm from elsewhere. Their rule is correct but not for the situation to which it is applied. A common error of this type occurs when students are learning about fractions using pizzas. One piece out of two, and one piece out of three, makes two pieces out of five so they conclude that \( \frac{1}{2} + \frac{1}{3} = \frac{2}{5} \).

A rational error common to students in early secondary is the syntactic error that is made when students translate a worded description to symbolic notation. If told there are 6 times as many dogs as cats, these students will write \( 6d = c \) (where \( d \) represents the number of dogs and \( c \) the number of cats). Errors may be triggered by words that remind students of a process for which the word frequently applies. The word left often prompts students to subtract, so when asked “How much money did a person start with if they had $3 left after paying $20 for an item?” young students would subtract the smaller number from the larger number and suggest that the answer is $17.

The application of an algorithm in a situation where it does not make sense is another type of rational error described by Ben-Zeev and Sternberg (1996). An example of this is provided in the previous chapter when proportion was applied in the non-proportional situation of running around a track. During my investigation of the hierarchy of learning about percentages (Burfitt, 2014), the students made a similar error. Most students in the medium to high levels of ability selected the response where the proportion had been used in the calculation rather than recognise that no calculation was needed. For the MC item shown in Figure 5.2, these students identified 12% as the correct response.
Ben-Zeev (1998) noted that errors are rarely random in nature, that they are often quite logical, and that they can be constructive for student learning because they provide further support for the development of understanding. If left unchallenged, students may accept the errors as being correct and adopt them as new understandings, which then become difficult to remediate. Knowledge of the common errors that students make can provide insight into the way students learn, and this in turn helps teachers to identify appropriate learning experiences which allow students to develop more accurate conceptual understanding.

**Errors as indicators of progress**

Evidence that misconceptions, alternative conceptions, or errors were considered as indicators of progress was provided by Sadler (1998) in his study of the MC items used to test scientific understanding. The distractors for the 47 items were written to contain alternative conceptions and to be as plausible as possible. Sadler reported that all items had at least one distractor with information. The student support for the distractors with information increased as student achievement increased and then declined at the very highest levels of achievement. Sadler concluded that, if students of medium ability find the alternative conceptions increasingly attractive, then this is evidence that the development of these misconceptions is an intermediate step in the development of full knowledge. Kind (personal communication, May 27, 2016) cautioned against using misconceptions to create the distractors because it has been known to result in low test reliability and item discrimination. Kind suggested that using misconceptions to write distractors is more distracting for the higher ability students and recommended the use of ordered options in MC items as an alternative.
In a large-scale study of number and algebra skills, Hodgen, Brown, Coe, and Küchemann (2012) investigated the extent to which certain errors or misconceptions developed as student understanding increased. Their study involved an analysis of incorrect responses in a constructed-response test which was administered to students aged 11 to 14 years. Their conclusion that certain errors increased then declined as student proficiency increased were similar to Sadler’s (1998) results as described earlier. One such misconception was the belief that multiplication makes bigger. Students who persist with this thinking will expect that calculations such as 5.6 x 0.5 will result in a product greater than either given number. These students are applying their knowledge of whole-number multiplication. Coe (personal communication, May 19, 2016) proposed that a misconception may be on the way to learning a concept for some students, and as such it is better than zero knowledge and should be considered as worthy of being awarded some credit. So, while selecting a misconception is selecting an incorrect answer, it could be considered as demonstrating partial knowledge.

The description by Gila, Dreyfus, and Hershkowitz (2010) of the development of knowledge about understanding probability gives further insight into how partial knowledge, errors, misconceptions, and alternative conceptions may develop and persist in a person’s thinking. Gila et al. described partially correct constructs (PaCCs) as incorrect answers based on accurate knowledge, or as correct answers based on inaccurate knowledge. In the development of complete knowledge, a person needs to recognise the relevance of what they already know to the new situation, combine these known elements, and integrate them to form new knowledge. Knowledge may be incorrectly formed when elements are missing, when essential elements are not recognised, or when chosen elements are unrelated to the new situation. The notion of PaCCs, according to Gila et al. (2010) suggests the students have only partial knowledge of the considered content.

**Credit for partial knowledge**

The studies described above provide support for the premise that partial knowledge deserves partial credit. Distractors are informative, that is they have information, when they contain aspects of the correct response and are selected by the students with only partial knowledge of the concept tested by the item. The students may have developed misconceptions that are grounded in other accurate knowledge and understanding, and which are related to the concept that they are studying. Such misconceptions may develop as understanding increases and thus be deemed to be partial knowledge.
For this study of the function of MC items, the interpretation of what is meant by *partial* knowledge includes both of these considerations. More specifically, students may be said to have partial knowledge of a concept when they know some but not all details of the concept, or when they have developed a well-known misconception grounded in some correct thinking. To create MC items which can assess this partial knowledge, writers of items need to have sound understanding of the test content, good insight into student conceptual development, and knowledge of the misconceptions associated with that development.

In my earlier study (Burfitt, 2014) the MC items had not been created to investigate partial knowledge, and in a few items, some incorrect options were more frequently selected than others. On examination of these items, it appeared that students had partial understanding of the concepts being tested. As seen in Table 5.1, the students at the lower levels of achievement (Levels 1 and 2 out of 5), were selecting options that were close to being correct, or which indicated that the students held a common misconception. Rather than select 25% when asked to identify the percentage equal to 25 out of 100, lower-ability students chose a quarter. In another item, students incorrectly applied their knowledge of fractions and selected 60% when asked to identify the percentage for 6 out of 8. The relevant items and the popular incorrect selections for these and other items are listed in order of item difficulty in Table 5.1.

<table>
<thead>
<tr>
<th>Item</th>
<th>Levels</th>
<th>Correct answer</th>
<th>Incorrect student choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>37%</td>
<td>37 out of 100</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>25%</td>
<td>a quarter</td>
</tr>
<tr>
<td>16</td>
<td>1, 2</td>
<td>20%</td>
<td>10 out of 50 as 10%</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>75%</td>
<td>6 out of 8 as 60%</td>
</tr>
<tr>
<td>32</td>
<td>1, 2</td>
<td>Less than 11</td>
<td>81% of 11 as greater than 11</td>
</tr>
<tr>
<td>35</td>
<td>1, 2</td>
<td>25%</td>
<td>8 out of 32 as 24%</td>
</tr>
<tr>
<td>36</td>
<td>All</td>
<td>10%</td>
<td>90 to 99 as an increase of 9%</td>
</tr>
<tr>
<td>43</td>
<td>1, 2, 3</td>
<td>20</td>
<td>29% of 70 as closer to 30</td>
</tr>
<tr>
<td>60</td>
<td>All</td>
<td>$\frac{1}{3}$%</td>
<td>$75 to 100 as a 25% rise</td>
</tr>
<tr>
<td>65</td>
<td>2</td>
<td>50</td>
<td>If 120% = 60, then 100% = 72</td>
</tr>
</tbody>
</table>

**Table 5.1 Popular incorrect responses in MC items**
Further investigation is warranted before confirming that the students have demonstrated partial knowledge in these items from the earlier study. If partial understanding is confirmed *number right* scoring for the items would not capture all the information available about student proficiency.

From a review of the research into student conceptual understanding in mathematics, science, and programming, Confrey (1990, p. 19) concluded that “misconceptions were documented to be surprising, pervasive, and resilient.” While some researchers see misconceptions as errors, it seems that errors appear naturally as knowledge develops, and it is reasonable to describe this as having partial knowledge. Rewarding partial knowledge can be achieved by scoring the elimination of distractors, by asking students to rate the options in order of preference, by asking students how confident they feel about the answer they have chosen, or by scoring the distractors. Having partial knowledge of a concept can be considered to be on the way to developing a full understanding of the idea, and it is possible that MC items can be written to allow students to demonstrate their partial knowledge. It is also possible that awarding credit for such partial knowledge can provide improved estimates of student ability.

5.4 Partial knowledge in proportional reasoning

Students may be said to have partial knowledge of a concept when they know some but not all details of the concept, or when they have developed alternative conceptions or misconceptions grounded in some aspects of correct thinking. Identification of what constitutes partial knowledge in a concept involving proportional reasoning is challenging. It requires the ability to identify the fine-grained aspects of knowledge which are collectively needed to develop full knowledge or understanding of the concept. Furthermore, identifying partial knowledge requires an awareness of the common misconceptions that students develop as they learn new knowledge, and an appreciation for the processes by which such misconceptions develop. The proposed pathways for the development of students’ skills in proportional reasoning present potential sequences for the major concepts that students need to master. The researchers who provided these pathways have not claimed that the stages of development indicate that students have partial knowledge of the concepts described, but rather that the stages indicate the probable orders in which these concepts develop.
From my reading of the research findings reported in the literature, it appeared that partial knowledge in proportional reasoning could be allocated to one of four possible types: additive thinking; proportion ignored; incomplete solution; and reasonable estimate. In additive thinking, problems on proportion are solved using the total difference rather than the proportional difference. When ignoring the proportion students are only considering amounts as fixed values rather than as proportions, and when they are only able to complete part of a calculation the solution is deemed to be incomplete. Students may not know an exact answer to a question but may know whether the solution lies within a particular range of values, and for this study such solutions are considered as reasonable estimates.

**Additive Thinking**

Using addition instead of multiplication to solve problems involving proportions has been widely reported in the literature, and it has been described as the use of additive strategies or as additive thinking (Hart, 2004; Karplus, Pulos, & Stage, 1983; Misailidou & Williams, 2003; Noelting, 1980a; Piaget & Inhelder, 1975; Tourniaire & Pulos, 1985). When there is a proportional relationship between two variables, students might use the difference rather than the scale factor to add to one of the measures and hence determine a missing value. In the equation \( \frac{3}{x} = \frac{7}{12} \), responding with 8 for the value of \( x \) is an additive error in which the student concludes that the relationship between 3 and 7 shows a difference of 4 and then expects the relationship between the \( x \) and the 12 to also show a difference of 4. Alternately 12 is 5 more than 7 and \( x \) should be 5 more than 3.

An example of an MC item which attracts additive thinking is shown in Figure 5.3.

<table>
<thead>
<tr>
<th>Thabang can run 4 times around the track in the same time that Tshepo can run 3 laps. When Thabang has run 12 laps, how many laps has Tshepo run?</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 9</td>
</tr>
<tr>
<td>b. 11</td>
</tr>
<tr>
<td>c. 13</td>
</tr>
<tr>
<td>d. 16</td>
</tr>
</tbody>
</table>

*Figure 5.3  MC item for proportional reasoning (2) (Long, 2011)*
For this item which tests proportional reasoning, the selection of the option which was indicative of additive thinking, was much greater than expected by chance for three of the five ability levels of the students. The percentages of students at each ability level who selected the option for additive thinking were, in order from lowest to highest ability, 45%, 47%, 39%, 20% and 8% (Long, 2011, p. 239). The proportion of students selecting Option (b) for the item shown in Figure 5.3, was much higher than for the other distractors at every ability level.

In a study of the responses to items on ratio and proportion from over 2000 students, Hart (2004) found a similar result. Thirty percent of the students were using the additive strategy on four of the most challenging items in the test. The items did not require algebraic methods for determining the correct answers, for example, solving an equation of equal ratios, but they did require reasoning of proportions. The same type of error occurs when students use absolute change rather than relative change while scaling two-dimensional diagrams.

Lovell (1971) suggested that the use of additive thinking was a stage in the development of accurate proportional reasoning but Lamon (2007) claimed that it was unclear that additive thinking is a stage of development. Nunes and Bryant (2009) suggested that additive thinking is not a step along the way to thinking multiplicatively because the two types of thinking have different origins. Additive thinking involves combining quantities and identifying absolute differences with usually one variable, while multiplicative thinking includes sharing by subdivision and two variables are involved. Nunes and Bryant argue that young children are capable of both types of thinking.

In the study by Misailidou and Williams (2003) additive thinking was the most common type of error made by students aged 10 to 13 years, and 26% of the variability of success on proportional reasoning tasks was due to inappropriate additive thinking. The students with the higher ability ratings who were incorrect on the items were more likely to be using additive thinking than making any other type of error. Using Rasch Measurement Theory, the authors conducted a further analysis in which they only scored the use of the additive strategy. The resulting scale for the use of the additive strategy was at least as stable as the scale for proportional reasoning. The error was made in items where the scalar or functional relationships (within and between members of the ratio) consisted of multiplication or division by non-integers. Misailidou and Williams concluded that students who make additive errors are more capable than those who make other types of errors and suggested that the use of additive thinking is a step along the way to the development of multiplicative thinking.
Misailidou and Williams (2003) described other errors that were made by students when reasoning with proportions and they reported that these errors were not as pervasive as additive thinking. One error that they described involves doubling or halving when it is inappropriate, and students with such tendencies appear to be less able than those resorting to additive strategies. Another error they described related to the interpretation of *sameness.* When asked to share out objects so that each person receives the same, or to scale an object so that the figures are the same, students may think that means the total for each person or the length for each object has to be equal rather than each share being in proportion.

*Proportion ignored*

Another type of error occurs when the proportional nature of the item is not considered and only part of the given information is used to solve the problem. Behr, Wachsmuth, Post, and Lesh (1984) reported that 30% of the students added the numerators and denominators to find \( \frac{1}{2} + \frac{1}{3} \), answering with \( \frac{2}{5} \) instead of \( \frac{5}{6} \). For a similar item in the Year 7, 2013 NAPLAN numeracy assessment, 16% of the students chose \( \frac{12}{14} \) for the answer to \( \frac{7}{8} + \frac{5}{6} \). In another example (Behr et al., 1984) the students had to select an estimate for \( \frac{12}{13} + \frac{7}{8} \) from 1, 2, 19, 21, and don’t know. The correct answer was identified by 24% of the students but there were 28% who selected 19 and 27% who selected 21. It appears that students selecting 19 or 21 were considering the numerators to be independent of the denominators. They were not seeing the relationship between the two parts of the fractions, nor were they aware of the sizes of these fractions. Hart (2004) reported a similar error where 20% of students aged 12–13 indicated that \( \frac{4}{8} \) was greater than \( \frac{2}{4} \). In another example provided by Hart, the student is told that for every 1 part of mercury there are 5 parts of copper and for every 3 parts of tin there are 10 parts of copper. When the proportion of copper is ignored, the student incorrectly concludes that for every 1 part of mercury there are 3 parts of tin.

Lamon (2005) reported that speed and other similar intensive quantities can be problematic for students. They commonly average speeds over different distances to determine the overall mean speed. This too, points to a lack of understanding of the proportional relationship between the two components of the rate.
Incomplete solution

Other than the errors described earlier, Misailidou and Williams (2003, p. 352–354) named two further significant errors as “incorrect build up” and “incomplete reasoning”. Incorrect building up can occur when a student is given a situation in which 3 kilograms cost $5 and they are asked for the cost of 7 kilograms. With incomplete build up, the student thinks that 6 kilograms will cost $10 so just one more is needed and thus gives $11 as the answer. Students may indicate incomplete reasoning when they start a problem recognising some of the requirements of the task but then proceed to complete the task with inaccurate reasoning. A student not completing the previous item would give $10 as the answer.

In an item examined by Hodgen et al. (2012) students were asked to write eleven tenths and many wrote 0.11. The authors suggested that the students showed partially correct understanding because they knew that the tenths started in the first column after the decimal. It is possible that the students did not think about what they had written, nor did not realise their error, and hence their reasoning and their solutions were incomplete. This error was not prominent among the students of lower ability, but its frequency increased among the students with medium ability and then declined for students with the highest levels of ability.

A student may provide an incomplete solution when only part of the proportionality is used to solve the problem because the student does not recognise the scale of the proportion and their thinking is limited to concepts relevant to linear relationships. This may occur when solving problems of scaling area and volume, when linearity is assumed and applied. An example of such thinking is seen when the students are asked to identify the change in the area of a square if the sides are doubled, and they conclude that the area is also doubled. Modestou and Gagatsis (2007) used three different types of tasks to test for this error and they discovered that the error was not occurring by chance. They concluded that this error “resists, persists, and reappears regardless of students’ grade and of the tests’ settings” (p. 75).

Other types of partial knowledge might exist with the recognition of the operations needed to solve the problem but the inability to complete the computation. Given the area of a rectangle and asked to determine the length, students may recognise the need to divide the area by the length but if the numbers are unfamiliar or fractional, the students may identify the computation process but not the answer. Scaling similar triangles can result in the same type of error if students know that one triangle is a scaled version of the other, but they are unable to perform the calculations to determine the lengths of the sides.
Interpreting a situation then representing the associated proportion as a statement of equivalent ratios, but showing an inability to solve the equation, is another example of an incomplete solution. If students can identify the direction of change and recognise if their response should be greater than or less than an expected value, then this shows they are able to interpret the task but are unable to complete the solution process.

*Reasonable estimates*

At times, students know an approximate but not an exact value for the answer. They may know that the answer is within a certain range of values, or that it is less than or greater than one of the values given. This can lead to the recognition of a reasonable estimate. A question from Piaget (as cited in Hart, 2004) about the number of sprats fed to eels is one example where this thinking occurred. Children were told that an eel 15 cm in length is fed with 9 sprats and then they were asked to determine the number of sprats needed for an eel that is 10 cm in length. At a basic level of understanding, the children were able to respond by comparing numbers and they could decide if the answer was more, or less than a number without necessarily being able to quantify an answer. Responses of 5, 7, and 8 sprats would indicate that the child has the idea that the answer is more than half of nine and less than 9 but that they are unable to determine the correct response. For such items these types of responses indicate that the child can determine or recognise reasonable estimates.

5.5 *Rewarding partial knowledge in MC items*

The demonstration of partial knowledge in MC items has been described in the previous section in terms of the selection of options, the quality of the distractors, and the development of knowledge. The selection and scoring of distractors regardless of their content may not indicate what the students know about the item and for this study it is not considered to be a valid display of the partial knowledge of a concept. Providing a score for partial knowledge can only be justified when the quality of the distractors chosen by the students indicates that they have an awareness of some aspect of the correct response. To score distractors simply because they are frequently chosen by the candidates, or because a mathematical model indicates that there is a correlation with the correct response is difficult to justify. While it may seem unfair to score an incorrect response, it seems logical to do so if the incorrect response can be linked to the progress of student development of conceptual understanding.
In the research described in this chapter, there were studies which indicated that all the distractors in an item could be constructed to contain partial knowledge of the concept, and in some situations, it seemed possible that the distractors could be ordered to attract different scores. To achieve this in the current study, it would be necessary to have a detailed knowledge of the many fine-grained skills which would combine to form each concept and the relative value of each of these skills; such information is not available in the literature. However, from the literature it has been possible to gain sufficient insight into student development in the various skills associated with proportional reasoning to identify some aspects of partial knowledge of individual concepts. Thus, it is possible to construct MC items with distractors which allow the students to demonstrate their partial knowledge in proportional reasoning.

As indicated previously, this partial knowledge may be a misconception about the item content or in the inappropriate application of accurate skills and knowledge. It may be that students have applied an idea from one problem and used it to solve another where it is not relevant. Student knowledge may also be incomplete, or the student is unable to complete the required process to determine the correct response. Wherever it appears that the selection of the option is based on an error or a misconception that is linked to having some knowledge of the correct response, then the recognition of partial knowledge is justified.

To reward students for their partial knowledge in a MC item one could consider constructing the item with four options: a correct response; one distractor marginally relevant to the concept; a second distractor containing some information; and a third distractor with even more information. The application of such a framework is challenging in mathematics, and particularly for concepts associated with proportional reasoning. The conceptual field is complicated and contains many interdependent behaviours.

From the research literature, as described previously, it seems reasonable to classify partial knowledge of proportional reasoning as one of the four types: additive thinking; proportion ignored; incomplete solution; and reasonable estimate. While it seems possible to identify these different types of partial knowledge, more information about student development of the skills and understandings necessary for proportional reasoning is needed to support the creation of items in which all distractors address different levels of knowledge. Hence, in this study, the focus will be on creating just the one informative distractor which will provide an opportunity for the student to demonstrate their partial knowledge.
Students can demonstrate their partial knowledge when they select options which indicate they have some, but not all, the information needed to identify the correct response. Such options are deemed to be informative because they provide extra information about student learning. MC items are not generally constructed to contain informative distractors, nor are they analysed to measure any contribution from the distractors to the measurement of proficiency. However, test feedback to participants often contains an indication of the frequency of the selection of the distractors and any connection that this may have with common misconceptions held by the students. In an effort to locate research using post-hoc analysis of responses to MC items in order to develop a greater understanding of what might constitute partial knowledge of proportional reasoning, two internet searches were established: one for multiple-choice items, and one for proportional reasoning. Another purpose of these searches was to identify any effect that the recognition of partial knowledge might have on student behaviour and on the measurement of proficiency. Research articles on these two topics have been regularly provided for several months but little relevant information has been identified.

Current scoring practices for MC items involve the dichotomous method where the correct response attracts a score of 1 and there is no score for any of the distractors. In this system, any partial knowledge of the concept receives the same score as a total lack of knowledge. A more detailed scale of achievement in mathematical understanding was produced when students’ answers to constructed response questions were scored polytomously where there were part marks allocated for some partially correct responses (Van Wyke, 2003). Polytomous scoring has also been used to confirm the identification of distractors worth partial credit (Andrich & Styles, 2011), and it can be considered as a method of scoring items designed to reward partial knowledge. Given the current research knowledge and scoring methods used in large-scale assessments, a logical way to proceed would be to allocate zero for the lack of knowledge and understanding of the concept, award a higher score for the correct response, and give a score between these two for an answer that indicates partial knowledge. Further discussion of the scoring of responses with partial credit will be provided in the following chapters.
5.6 Research questions

With the widespread use of MC items in assessments of mathematical achievement, and the need to provide valid measures of proficiency, it is worthwhile to continue to improve the use and quality of MC items as well as to address the issues of concern. More accurate measurement scales can be developed if credit is given to the students for their partial knowledge and if the impact of guessing is reduced. Achieving these outcomes requires that greater attention be given to the creation of MC items. Only a few studies have reported the benefits of the deliberate construction of multiple-choice items containing distractors with information about achievement, and the findings are restricted to a limited content area. Insufficient advice to guide the construction of MC items to measure partial knowledge has been located, particularly in the important field of proportional reasoning. Furthermore, the impact of awarding credit for such information has not been addressed in these studies.

Studies indicating how the proficiency scale for proportional reasoning would be impacted by the award of partial credit have not been found in the research literature, although there are a small number of reports of the impact of awarding partial credit on other proficiency scales. Two important considerations when scoring partial knowledge have not been the subject of earlier research. First, there is the consideration of gender and the possibility that scoring partial knowledge might provide a benefit for either males or females. Second, there is the issue of the accuracy of measurement; there is little indication in the research of a significant improvement in proficiency measures when partial credit is given.

Descriptions of research into the use of adaptive testing and tailored analyses to create more accurate measures of mathematical proficiency in lower secondary have not been prominent in the literature. There has, however, been a significant amount of research into different scoring methods to counter the effect of guessing, but few studies have investigated the extent to which the scale for item difficulty and the measures of person proficiency are affected by guessing. Comparison of the performance of males and females in areas of mathematics has also been the subject of considerable research but few studies compare the guessing behaviour of these two groups and their abilities in demonstrating their partial knowledge.
Research questions

To provide a greater insight into these areas where the research is limited and to address the problems outlined previously, the following research questions are posed:

1. For distractors to contain information about student achievement in mathematics, how should multiple-choice items be designed and created?

2. To what extent is the proficiency scale for achievement in proportional reasoning affected when distractors with information about student learning are awarded partial credit?

3. To what extent is the proficiency scale for achievement in proportional reasoning affected when guessing is reduced in MC items.

Learning to create MC items with distractors containing information about student learning has significant benefits for test creators, not only in mathematics and not only for secondary students. It would enable writers to develop MC items which could generate scores with greater validity and authenticity, and which could be interpreted with greater confidence. Creating such distractors would help educators gain further insight into student misconceptions as well as develop a deeper understanding of the student’s conceptual development. To date there have been many attempts to reduce the effects of guessing in MC items, but none of them are seen to be fair and effective for large-scale assessments. The method suggested in this study (see Chapter 6) is more recently proposed and could be successful in increasing the attraction of MC items to educators as well as provide more accurate measures of student achievement.

Improving the structure and function of MC items, and improving the analysis of these items, could provide greater detail about student learning in mathematics without increasing test demands. With the recognition of partial knowledge and the reduction of the effects of guessing, MC items could be more effectively and economically used and provide improved measures of student performance. Furthermore, the use of more effective MC items could provide greater insight into the knowledge and skills necessary for conceptual development: in this study, the development of proportional reasoning. Such information would be of great value to teachers for the planning of their students’ learning experiences. The achievement of these outcomes would provide significant benefits to those who construct, analyse, and use MC items: both educators and students.
CHAPTER 6: Rasch measurement theory

Introduction

Rasch measurement theory provides an ideal framework for the analysis and reporting of student achievement in tests of mathematical understanding. The theory has been applied in a variety of studies, including those relating to the assessment of skills and understandings necessary for the development of sound proportional reasoning. It has been used to inform the analysis of student responses to both multiple choice and constructed response items. The application of Rasch measurement theory produces estimates of student proficiency and item difficulty, as well as measurement scales which describe a continuum of achievement, and which indicate the relationship between these estimates. Other desirable features of assessment can be facilitated, and these include having different scoring methods for items, allowing students to respond to alternate sets of items, and reducing the effect of guessing in the estimates of student proficiency. These features can assist with improving the way MC items function by allowing partial credit to be given, adaptive testing to be implemented, and scores to be adjusted to counter the inflation due to guessing. No other theory for the analysis of responses is seen to provide these valuable insights into item characteristics and student achievement in the same coherent manner.

The application of Rasch measurement theory, which is also referred to in the literature as Rasch modelling or Rasch analysis, can be implemented using RUMM2030 software (Andrich, Sheridan, & Luo, 2016). The software provides the estimates for item difficulty and person proficiency as well as the common scale along which these are located. It is possible to see how well the items are targeted to student proficiency, and to study student behaviour in the selection of the various options in the MC items. In studying this behaviour, evidence is available to confirm the hypothesised existence of partial knowledge of the concepts addressed in the individual items. The software can be used when students do not respond to every item: either when the test is adapted to individual student proficiency; or when the student deliberately chooses to skip an item. The software provides processes to support the identification of responses which indicate the extent of guessing, and it provides mechanisms by which the effects of guessing can be reduced in the students’ responses (Andrich et al., 2016). With these and other features, RUMM2030 software provides all the functions necessary for the analysis of responses to both MC and CR items for the investigation into the processes deemed likely to improve the function of MC items as previously described.
6.1  Research using Rasch measurement theory

With the considerable benefits of using Rasch measurement theory there are many research studies reporting the application of the theory in the educational field. The identification of pathways of learning has involved the use of Rasch modelling to determine scales of student achievement and item difficulty. To see how student performance is affected when different types of items are used, success when responding to multiple-choice items has been compared with success when responding to short answer questions for which students are required to create their own responses. Rasch modelling has been applied to identify differences in the proficiencies of males and females, and to locate item types that might favour subgroups of test participants. Some research has focussed on the effects of guessing and on ways by which the scales are affected when guessing is removed from the item estimates. Different scoring techniques have been studied to determine the most fitting way to score individual items, and this has included the allocation of credit for partly correct responses. More recently, studies of student behaviour in the selection of options in multiple-choice items and in the provision of partly correct responses have been reported in the research literature. Findings from the studies of student behaviour have given greater insight into the students’ conceptual development, and this has provided valuable information for teachers and curriculum planners. Research using Rasch modelling is extensive and only a limited number of relevant and significant studies are provided in this report.

6.1.1  Developmental scales in mathematics

Research involving the identification of student conceptual development is enhanced with the application of Rasch modelling to the students’ responses to test items. The resulting scales with descriptors of performance at various levels have been described as developmental scales, developmental pathways, learning trajectories, or hierarchies of learning. In my earlier research (Burfitt, 2014), Rasch modelling was used to identify a hierarchy of learning about percentages for students in Years 7 to 10. Students responded to both MC and CR items, and scales of student proficiency and item difficulty were produced. The effects of guessing were removed from the student and item estimates, and the scales were used to identify the concepts that were both mastered and challenging for the different year groups, and to generate advice for teaching about percentages. There was evidence of common errors and misconceptions in the students’ responses and, while these were not addressed in that study, I considered that they would be a relevant focus for future research.
A Numeracy Achievement Scale was produced when the Rasch model was applied to the responses of students from Kindergarten to Year 6 (Looveer & Mulligan, 2009). The scale reflected an ordering of concepts and strategies consistent with the expected curriculum, and it indicated the differential growth in numeracy between adjacent year groups. The findings provided valuable information that teachers could use to plan learning tasks. Another common scale of achievement was produced when partial credit was given and Rasch modelling was applied to items which tested statistical concepts (Watson, Kelly, & Izard, 2006). Over a period of 10 years, data were collected from more than five thousand students from several grades. Even though items evolved over the period of the study, it was possible to identify trends in performance for the different grades and to relate these trends to changes in the nature and implementation of the curriculum.

Rasch modelling was used to produce a developmental scale of mental computation with part-whole numbers by Callingham and Watson (2004). The 122 items, which were provided on a professional recording, and for which no written calculations were allowed, tested the students in Grades 3 to 10 on their skills with common fractions, decimals, and percentages. There were 122 CR items and they were organised into sets containing overlapping items. Evidence of discontinuities on the resulting scale of item difficulty indicated that there were six levels of performance across the eight grades. While one would expect most Grade 10 students to be located predominantly in the top one or two levels, Levels E and F, there were still 20% of these students at Levels C and D. There was evidence to suggest that student performance was affected by partially understood rules: Callingham and Watson (2004, p. 82) suggested that more detailed developmental pathways could be described in future if the “common errors are recognised and scored as partial understanding”.

In an earlier study, Callingham and McIntosh (2001) had used similar items and analysis to develop an eight-level scale of mental computation. Over 1400 students responded to 238 items which tested skills associated with number operations. Students operating at Level 8 could multiply and divide by 0.5 and add simple unit fractions, whereas at Level 1, student skills were restricted to the addition of single digit numbers. Callingham and McIntosh (2001) used their results to identify an order in which to teach the skills associated with multiplication by single-digit numbers, and they suggested that multiplication and division could be considered at the same time when teaching these concepts. Such information is important for curriculum writers in their ordering of content and for teachers to identify appropriate tasks for learning.
For the creation of a hierarchy of development in proportional reasoning with students aged 10–14 years, Misailidou and Williams (2003, 2008) used Rasch modelling to scale items for the diagnosis of student strategies and the measurement of student proficiency and progress. They compared the difficulty estimates of items presented using mathematical models, for example, diagrams and pictures, with the estimates of items with no models, only text. They found that the best models to support teaching students to reason about ratios were pictorial designs and double number lines.

It was also when using the Rasch model that Misailidou and Williams described some items in terms of their capacity to promote additive thinking which they suggested was a “stepping stone to multiplicative understanding” (Misailidou & Williams, 2003, p. 359). After applying the Rasch model and interviewing students aged 10–13 years, they developed a measurement scale for the additive strategy and their results showed that the students who made the additive errors were more capable than the students who made other types of errors.

Rasch measurement theory has been applied in other studies of development in the skills associated with the conceptual field of proportional reasoning (Briggs & Peck, 2015; Long, 2011; Siemon, Izard, Breed, & Virgona, 2006; Wong, 2010). Within each of these studies, as well as the studies described in this section, the use of link items enabled the students of various ages and grades to be placed on the same scale. Some of these hierarchies of development have been described in Chapter 4.

6.1.2 Factors affecting performance

Some of the key factors which affect student performance, and which are relevant for this study, have been the subject of research where the Rasch model has been applied. These factors include item type, gender, guessing, scoring, and partial knowledge. The effect of these factors on student achievement have been described in greater detail in previous sections and only a brief overview is provided here. The potential for researchers to apply the Rasch model in the investigation of these factors is worth noting.

*Item type*

Research of student success with various item types was described in Chapter 3, and in two of these studies the Rasch model had been applied to the students’ responses. Hudson (2010) found that student performance in items with knowledge and recall questions in Chemistry was significantly better with MC items than for CR items, but there was little difference when
the items tested the application of knowledge. Similarly, Hohensinn and Kubinger (2011) found that CR items in tests of reading skills were more difficult for students than MC items, and for the latter, selecting two out of five options was more difficult than selecting one out of six options. Hudson (2010) also established a trial testing regime which confirmed the superior performance of males which had been indicated in the external state examinations but, in applying the Rasch model, the gender differences became more apparent. It revealed an advantage for males, but this was only at the higher levels of proficiency; there was a distinct disadvantage for females in the stoichiometry items.

**Gender**

Taylor and Lee (2012) combined the study of item type and gender when they used the Rasch model to analyse students’ responses in reading and mathematics with both MC and CR items. As reported in Chapter 3, males performed at a higher level in some areas of mathematics, including conceptual understanding and probability, while females outperformed males in problem solving and logical reasoning. Where differences existed for the students’ success in different item types, it was noted that males were generally more successful in MC items and, females were more successful in CR items.

**Guessing**

Scores for MC items can be inflated due to guessing which may be random or systematic. When Rasch modelling is applied, the items where guessing has occurred will appear easier than they would do without guessing (Andrich, Marais, & Humphrey, 2012, 2015), and further examination is necessary to identify for which items the amount of guessing has a significant effect on the measures of item difficulty. The extent of guessing in a MC item will depend on the amount of knowledge that the student has about the item’s content. The guessing is identified as being random when the student has no knowledge of the content and selects an option at random. When the student has a system, which they apply in their selection of an option when they do not know the correct response, the guessing is described as systematic. A process has been developed to remove the inflation from estimates of item difficulty, due to random guessing, with the development and application of the Rasch model (Andrich & Marais, 2014; Andrich et al., 2012, 2015). This process will be outlined later in this chapter. Further findings from these studies to reduce the impact of guessing have shown improvement in the resulting scales of person proficiency and of item difficulty.
Scoring

Van Wyke (2003) used the Rasch model to compare dichotomous and polytomous scoring for mathematics items which had been used in assessments with students in primary school. Three types of polytomous scoring were identified: hierarchical, incremental, and decremental. Hierarchical scores are aligned to levels of understanding, and the greater the understanding the higher the score allocated. Incremental scores are allocated when each part of an item is credited with a score, even when the parts are at the same difficulty level. With decremental scoring, marks are deducted for errors and omissions. The use of Rasch modelling in Van Wyke’s study of CR items indicated that these types of polytomous scoring were appropriate for only some items. Writing MC items with distractors that could fit any of these types of scoring would be challenging.

Partial knowledge

In Smith’s (1987) study of scoring procedures to assess partial knowledge in vocabulary, Rasch modelling was applied to the students’ responses. An analysis in which items were scored dichotomously, with 1 for a correct answer and 0 otherwise, was compared with an analysis in which the items were scored polytomously with scores ranging from 0 to 4. The items had been designed to capture partial knowledge and to be scored polytomously, and there was some evidence to indicate that this scoring method revealed the presence of information in the distractors as well as produce more valid and reliable results.

Studying student behaviour in the choice of options in MC items is enhanced when Rasch modelling is used in the analysis of the responses. When the student selects an incorrect option that shows they know something about the item’s content but not enough to select the correct response, then the option is described as an informative distractor or a distractor with information. It follows that the selection of such distractors should contribute to the measure of student performance on the proficiency scale. By providing an opportunity for the student to gain partial credit, the scale will be a more accurate reflection of student achievement (Andrich & Styles, 2011). The application of the Rasch model using polytomous scoring can accommodate such partial credit.

This description provides an overview of the factors, which have been studied with the application of Rasch measurement theory. Greater elaboration of the factors believed to have the potential to improve the function of MC items is located later in this chapter.
6.2 The Rasch model

The formulation and application of the probabilistic model known as the Rasch model began with Georg Rasch in the 1960s (Rasch, 1960). In using the model, the “role of the populations can be abolished” (p. 3) and as a result, processes such as regression analyses and analyses of variance are redundant. Rasch describes the model’s two parameters as “a ‘difficulty’ for each test (or item) and an ‘ability’ (the term proficiency is more commonly used) for each person.” The results from the estimation of each parameter are independent of the tests or items used in the calculations. Wright and Stone (1979) explain that Rasch and others have shown that the only formulation allowing the estimation of item difficulty and person proficiency to achieve these independent results is the Rasch model. This seemingly simple model, which is easy to use, “makes both linearity of scale and generality of measure possible” (Wright & Stone, p. 15). Hence, it can be applied in a great variety of research studies in the social sciences.

To measure student achievement, it is necessary to develop a test instrument that provides opportunities for students to demonstrate their proficiency. It is not possible to observe this proficiency directly, but it is feasible to observe manifestations of the latent trait (the variable) and to convert the observations to measurements (Andrich, 1988). For the Rasch model the variable needs to be unidimensional, that is, mapped to a single real number line. Variables which test individual achievement address many dimensions including the ability to read the item, the capacity of working memory, and the processing power to determine the result. It is possible for one variable to dominate the construct to the extent that the measures obtained can be attributed mainly to that variable and hence the variable identified as unidimensional.

The independence of the results of parameter estimations is described as the invariance of comparisons (Andrich, 1988). The variation in student proficiency can be quantified and compared without reference to the difficulty of the items that constituted the common assessment. Similarly, the variation in item difficulty can be described without referral to the proficiency of the students who responded to the items. This invariance is an essential feature of true measurement. Other essential features are the maintenance of a unit, and additivity (Andrich, 1988). The units in the Rasch model, known as logits, are constant across the scale and their sum can be interpreted. Logits are not units in the same sense as centimetres, as each analysis has its own an arbitrary unit in terms of logits, short for the log-odds of success given the proficiency and the difficulty. The Rasch model accommodates these essential features: invariance of comparisons, maintenance of a unit, and additivity, and thus satisfies measurement demands as expected in the physical sciences.
6.2.1 Determining the estimates

Items may be scored dichotomously where a score of 1 is given for a correct response and 0 otherwise. In the processes which transform these scores into probabilistic estimates of proficiency, the Rasch model is applied. The formula used describes the probability that a person with a proficiency level symbolised as \( \beta_n \) will score 0 or 1 on an item with a difficulty level symbolised as \( \delta_i \). The equations for determining the probability of obtaining 1 (Equation 6.1) and of obtaining 0 (Equation 6.2) are given below (Van Wyke, 2003).

\[
p_{ni} = P(X_{ni} = 1) = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \quad \text{Equation 6.1}
\]

\[
1 - p_{ni} = P(X_{ni} = 0) = \frac{1}{1 + e^{\beta_n - \delta_i}} \quad \text{Equation 6.2}
\]

where

- \( X \) is a random variable defining success or failure on an item,
- \( P(X_{ni} = 1) \) is the probability that person \( n \) is correct on item \( i \),
- \( P(X_{ni} = 0) \) is the probability that person \( n \) is incorrect on item \( i \),
- \( \beta_n \) = person proficiency on the variable scale,
- \( \delta_i \) = item difficulty on the same variable scale.

The difference indicated by \( \beta_n - \delta_i \) represents the distance of a person from an item on which they have a 50% chance of being successful. This distance is expressed in logits, which are log–odds and \( \log \left( \frac{p_{ni}}{1 - p_{ni}} \right) = \beta_n - \delta_i \).

In determining the estimates of person proficiency and item difficulty, first the items are calibrated, and their relative difficulties are determined from the raw scores that the persons receive on the items. On each scale of difficulty, the mean item location is set to zero logits. The proficiencies of the individuals who respond to those items are then estimated from these item difficulties, and they too are expressed in logits (Andrich, 1988, 2009a). The processes used to generate these estimates result in a common scale on which both the item difficulty and the measures of person proficiency are located. In situations where all persons have answered all items, the rank order of persons according to their raw scores is preserved in the production of their proficiency estimates.
6.2.2 Reading the scale

The RUMM2030 software (Andrich et al., 2016) provides several diagrams which show features of items and the behaviour of the respondents on the items. The category probability curve (CPC) depicts the relationship between person proficiency and the probabilities of receiving scores of 0 and 1 for that particular item. The scale along the horizontal axis is a conjoint scale (Callingham & Bond, 2006), and it indicates item difficulty as well as person proficiency. In theory, the achievement scale stretches infinitely in both directions, but the locations for most persons are between -3 and +3 logits.

The CPC for Item 2 from the current study is shown in Figure 6.1. At the lower end of the proficiency scale, the individuals have a high probability of scoring 0 on the item, and at the higher end the probability of being incorrect is nearly 0. The greater the proficiency of the person, the less likely they are to be incorrect and the more likely they are to be correct. The point where the curves cross is known as a threshold, and it represents the location where the probabilities of two adjacent scores (in this case 0 and 1) are equally likely and where $\beta_n = \delta_i$. The persons whose estimate of proficiency can be read from the horizontal axis at the threshold have a 50% chance of being correct on the item represented in the graph. With dichotomous scoring there is one threshold and it coincides with the item location and is referred to as the item’s difficulty.

Further details of the item are discussed in Chapter 9 along with the other items which were used to collect data in the test provided to the students.

![Figure 6.1 Category probability curve (CPC)](image-url)
In the CPC for Item 2 (Figure 6.1), the blue curve represents the probability of scoring 0 and the red curve models the probability of scoring 1. The item difficulty is defined as the location at which a person of that proficiency has 0.5 probability of answering correctly. In this example, Item 2 has a difficulty estimate (location) of 0.238 logits and persons with proficiency estimates of 0.238 logits have a 50% chance of being correct on this item. The graph also shows that, according to the model, a person with a proficiency estimate of -1 logit has about 20% chance of being correct on this item and a person with a proficiency estimate of 3 has a less than 5% chance of being incorrect.

6.2.3 Identification of item difficulty

Several diagrams relating to item difficulty and person proficiency can be generated using the RUMM2030 program (Andrich et al., 2016). All the figures provided in the rest of this chapter show graphs which have been produced using this software. The graphs were produced during the analyses of responses to items in either my previous study (Burfitt, 2014) or in the study described in this current report. The items were created to assess the mathematical achievement of secondary students.

The software generates the item characteristic curves (ICC) which model the students’ correct responses to the various items and, which are parallel. The items’ locations are given in the legend and can be estimated by reading the horizontal axis at the point where the probability of being correct is 0.5.

**Figure 6.2** Item characteristic curves (ICC) for Items 4, 19, and 20
In Figure 6.2, the ICC are shown for Items 4, 19, and 20; the item locations are -1.48, 0.17, and 1.06 respectively. The arrows have been added to ease the reading from the horizontal axis. The further to the left that the ICC is located, the easier the item.

Another diagram which gives insight into the behaviour of the students on an item is shown in Figure 6.3. The graph shows a continuous line which represents the theoretical ICC for that item, and the dots on or near the line (●) represent observed data. In this example, the sample of 859 students has been divided into five classes of approximately equal numbers and the mean person location for each class has been determined. The red line markers on the horizontal axis indicate these five means. The black dots corresponding to these red markers represent the proportion of individuals in that class who were observed to be successful on that item. In this analysis and representation of data for Item 2, the proportion of individuals who were successful in each class was similar to what was predicted by the model, as seen by the closeness of the five dots to the theoretical curve.

The graph also shows that at the highest level of proficiency, the proportion who were correct was just below that predicted by the model, and in the middle of the proficiency scale the proportion who were correct was slightly higher than predicted by the model. For the other three classes the proportion that was correct coincided with what was predicted by the model. When these proportions are close to the ICC, as in Figure 6.3, the responses are said to fit the model. One measure of item fit which is often used along with other statistics to check that the data fit the Rasch model, is the fit resolution (FitRes). This measure is given in Figure 6.3 and for Item 2, the FitRes = 1.024.

![Figure 6.3 ICC with observed responses for Item 2](image-url)
6.2.4 Polytomous scoring

Items can be scored polytomously when part marks are given for a response which is not fully correct or for individual parts of a question where only some of the parts are answered correctly. An example of this occurs when a student is given 0 for the incorrect response, 1 for a partially correct response, and 2 if their response is deemed to be fully correct.

Using the same symbols as for Equations 6.1 and 6.2, the formula used to describe the probability that a person with a proficiency level symbolised as $\beta_n$ will score $x$, where $x = 0, 1, 2, \ldots, m_i$ is given in Equation 6.3 where $m_i$ represents the maximum score of item $i$. The equation has been produced by adapting one provided by Van Wyke (2003, p. 29).

$$P(X_{ni} = x) = \frac{e^{-\tau_{1i} - \tau_{2i} - \ldots - \tau_{xi} + x (\beta_n - \delta_i)}}{\sum_{x' = 0}^{m_i} e^{-\tau_{1i} - \tau_{2i} - \ldots - \tau_{x'i} + x' (\beta_n - \delta_i)}}$$  

Equation 6.3

The curve showing the probabilities of participants obtaining the different scores 0, 1 and 2 is the category probability curve (CPC) and an example is provided in Figure 6.4. The curve shows that students with low proficiency, who are located further left on the horizontal axis, have a higher chance of scoring 0 than of scoring 1 and very little chance of scoring 2 in this item, Item 12. For persons located half a logit either side of 0, the probability of scoring 1 is higher than that of scoring 0 or 2.

![Category probability curve (CPC) for Item 12](image)

*Figure 6.4 Category probability curve (CPC) for Item 12*
In the highest proficiency group, where the person location is 1 logit or more, persons are more likely to score 2 than 1, and have very little chance of scoring 0. For polytomous scoring there is more than one threshold, and when scores are 0, 1, or 2, there are two thresholds. The first threshold is located where the probabilities of scoring 0 and 1 on the item are equal, approximately -0.4 for Item 12, and the second threshold occurs where scoring 1 has the same probability of scoring 2, approximately 0.8. For Item 12 shown in Figure 6.4, the probabilities of the different scores follow a logical order in that the greater the proficiency of a person, the more likely he or she is to score a higher mark. When this logical ordering occurs, the thresholds are said to be ordered.

When polytomous scoring is applied and the thresholds are ordered, then these item thresholds are “exactly the values of the difficulties of the items of a compatible resolved design in which there is a dichotomous response with respect to each of the thresholds” (Andrich & Styles, 2009, p. 74). In other words, the one item with polytomous scoring is equivalent to two items, each with dichotomous scoring.

6.2.5 Targeting

When items are created to measure student achievement, some items are needed to provide opportunities for students at each end of the proficiency scale to demonstrate their competence. A lack of challenging items could indicate that too many students get the most difficult items correct and then they cannot be separated into different proficiency groups. This matching of items to person abilities is referred to as targeting (Andrich et al., 2016).

The person-threshold distribution shown in Figure 6.5 is a histogram which simultaneously displays the person distribution in terms of class intervals, on the same scale as the item difficulty for which the mean of the item locations is set to 0 logits. In this example the items are well aligned to the proficiency range as there are persons from locations -1.8 to 3.6 and items from -1.8 to 2.8. Most of the items have locations in the middle of the proficiency range where most of the students are also located. The mean person location for a “well-matched” test would be close to 0 (Bond & Fox, 2007, p. 62) and in this test the mean of 0.104 indicates good targeting. This is supported by the standard deviation of 0.795 which is close to 1. While this distribution shows a good alignment of items to persons, the inclusion of extra items of a more challenging nature would allow the very able students to demonstrate their proficiency.
Figure 6.5  Person-item threshold distribution

6.2.6  Linking

The location estimates for person proficiency and item difficulty can only be compared with other estimates from within the test where the estimates are derived. They are relative in value rather than absolute. To compare achievement from one test to another, a sufficient number of items need to be common to both tests: Tests are then said to be linked (Callingham & Watson, 2004). With linked tests it is not necessary for all individuals to complete the same items for comparisons to be made and for individuals to be placed in the same scale; by linking, tests can be adapted for a wider range of proficiencies than when linking is absent.

For both simulated and real data, Kim (2006) investigated linking in MC items to determine the number of link items needed to achieve precision. Using information for linking from the item’s correct response only (CO) was compared with using information from all options (AC). Kim used an item response model that models all response options in an MC item and produced a category characteristic curve for each option. In Kim’s studies the linking items were selected at random. With real data for a test of 36 items, Kim found that 15 items were needed for linking with the CO method and 5 with the AC method. While these studies did not involve the use of Rasch modelling, nor was the quality of the distractors considered for the linking process, they indicate a potential benefit which could be considered for further research on MC items. If it could be shown that both the correct response and an informative distractor within a single item could be used for linking, then the overall number of the common items to which all students respond, could be reduced.
6.3 Fit of data to the Rasch model

Before we can claim the benefits of using the Rasch model to measure item difficulty and person proficiency, it is important to check that the observed data fit the model, and that the assumptions hold; that is, the assumptions upon which the model is based (Wu & Adams, 2013). According to Andrich (1985), different tests of fit are used because there is not any one test that can check for all situations where the data do not fit the model, and each test checks for “different violations of the model with different power” (p. 48).

Smith and Plackner (2009) describe a variety of factors which disrupt the measurement process, and they recommend that a family of fit statistics be implemented to address the inadequacy of any single test for misfit. They remind researchers that it is necessary to minimise measurement disturbances to obtain the best estimates possible. Measurement disturbances include the inflated effects due to guessing, item bias favouring subgroups of populations, and inter-dependence between items. The Rasch model incorporates requirements of relative unidimensionality and invariance which were described earlier in this section. Bias in the estimates is increased when these requirements are not met, and it occurs when there is more than one attribute or dimension being measured, or when the observed responses do not match the expected scores as predicted by the model.

Given the probabilistic nature of the model, a misfit of up to 5% of the items is expected (Callingham & Bond, 2006) and this may not be problematic. The scale may still be useful and an examination of the items to explain the misfit may be informative in providing insight for the creation of better items for future assessments. There may be diagnostic information in the items which have been identified as misfitting the model and this may indicate unforeseen ways by which the students have interpreted the item or understood the item’s content.

Extreme misfit of a large proportion of items, which is unlikely if the test is well designed and constructed, may render the proficiency estimates unstable and the data may need to be discarded. In some situations, one might assume the disruption to measurement affects all persons, for example, guessing, and to improve the scale a further analysis targeted at reducing guessing is performed. Where there is dependence between items or a bias towards a sub-group of the population, the responses can be modified or re-organised, and the analysis can be repeated with the altered responses (Smith & Plackner, 2009). This may involve creating two items from one by splitting the responses into two groups based on performance, or even combining the responses from two items to make one much larger item.
The RUMM2030 software (Andrich et al., 2016) provides routines which generate graphical and statistical output for detecting and identifying the relative fit of the responses to the model. Tests of fit are chosen to suit the types of responses and their scoring. These tests are considered separately and together to assess the level of confidence in the reliability of the estimates for item difficulty and person proficiency. Some of these tests are outlined in the following section.

**Person separation index (PSI)**

The person separation index (PSI), analogous to classical test theory reliability of the ratio of the true variance of persons’ scores relative to the total variance, is provided in the summary statistics when using the RUMM2030 software. If the PSI is close to 0 then the person estimates are not well spread across the continuum. This could be the result of the scores being close in value, persons with high and low scores being in the same section of the scale, or the presence of many dimensions in the test construct. The fit to the model may appear acceptable but the “power in detecting that items do not fit is very low” (Andrich et al., 2016, p. 9). The closer the value of the PSI is to 1, the greater the power of the items to separate persons on the measured variable, and to produce useful estimates of person proficiency.

**Graphical evidence of misfit**

Examination of the ICC can show how well the data fit the model. When dots representing the observed mean person estimates for each class coincide with the continuous curve representing the theoretical values as predicted by the model, the fit is described as excellent. The item presented earlier in Figure 6.3 showed relatively good fit. The curve for Item 59, shown in the first graph of Figure 6.6 shows that the observed means for a class coincides with the theoretical value curve in only one of the five classes: there is only one “dot” on the curve. Two dots are above the curve and two are below. For Item 56, the second graph in the same figure, only one class mean location for the observed data is close to the theoretical mean, and compared with Item 59, other observed means are further away from the curve. Neither of these two items appear to be a good fit for the Rasch model.
Figure 6.6  ICC for ill-fitting items

**Probability of chi-square statistic**

The RUMM2030 software provides an approximate chi-square statistic which is best used as an order statistic and in conjunction with other statistics and the ICC plots (Andrich et al., 2016). This item statistic represents the sum of the squares of the standardised residuals (differences between observed and theoretical expected values) for each class but it is affected by sample size and the number of class intervals formed. The probability of obtaining the chi-square value is reported on the ICC for each item. For Items 59 and 56, as shown in Figure 6.6, these probabilities are 0.026 and 0.000 respectively. These probabilities indicate that the items misfit at the 5% level of significance and that the difference between the observed and the theoretical means is larger than would occur by chance (Andrich et al., 2016).
To identify possible concerns, qualitative examination of such items is worthwhile. Considering both the ICC, and the probabilities for Items 56 and 59, it is noted that the item with the greatest differences between the observed scores and the theoretical scores, as seen in the distances of the points from the theoretical curve, is the one with the lowest probability. This test of fit is not specific for any particular misfit, and the values of the statistics are affected by a variety of other factors; this test is best used to compare items rather than to decide the level of fit of individual items.

*Graphical consideration of discrimination*

One of the assumptions of the Rasch model is that all items discriminate equally (Haladyna, 2004) and hence the theoretical curves of the ICC as shown previously in Figure 6.2, are parallel. This slope or gradient of the ICC is an average of the empirical discriminations of the items and is referred to as the item discrimination. Thus, discrimination describes in general the rate of change of the ICC (Andrich, 2009, p. 3).

The ICC for Items 40 and 71 are shown in Figure 6.7 and the observed responses in each case follow a pattern relative to the theoretical curve. For Item 40 the gradient of a curve joining the observed responses is much steeper than the curve of the theoretical ICC, and this item is classified as being over-discriminating. In comparison, the gradient for Item 71 is much lower, the slope is not as steep, and such an item would be said to be under-discriminating. An examination of the observed means makes it quite clear as to which groups are achieving higher or lower scores than predicted by the model.

*Fit residuals*

It is inappropriate to consider the fit of individual test items in isolation as the items work together, and one badly fitting item can affect the apparent fit of the others. The fit residual can be used as an order statistic to determine the relative fit of the different items. The fit residuals for each item are given on the ICC and for Items 40 and 70 the fit residuals (FitRes) are -3.94 and 1.678 respectively. With a theoretical mean of 0 and a standard deviation of 1, a very low (negative) fit residual indicates that an item may be over-discriminating, and a high fit residual indicates an item that is possibly under-discriminating. A value outside the range from -2.5 to 2.5 signals the possibility of misfit and indicates that the item should be examined qualitatively to determine what contribution it is making to the scale of achievement (Andrich et al., 2016).
It is possible that over-discriminating items are adding nothing to the measurement of the construct, while under-discriminating items might be assessing more than the dominant trait (Andrich, 2009). While it is worthwhile to check the items to see if they can be removed or improved, it is also important to use the fit residuals as relative indicators of the discriminating power of the items (Wu & Adams, 2013).

**Figure 6.7** Items with different discrimination
Local dependence

Local dependence between items or sets of items is seen as a violation of the Rasch model and it can reduce the accuracy of the estimates and hence affect the person distribution (Marais & Andrich, 2008a, 2008b). There are two ways by which local independence can be violated and these are referred to as trait dependence and response dependence. Trait dependence can occur when items have some common features which may relate to the content, structure, or stem. With response dependence, the individual’s response to one item has a bearing on their response to a following item (Andrich & Kreiner, 2010). Using simulated data, Marais and Andrich (2008b) investigated the influences on the scale and the PSI when the responses were not statistically independent. Their results indicated that trait dependence reduces the variance in person estimates and indicates lower reliability, whereas response dependence causes the estimates and the reliability to be inflated.

If two items are statistically independent, then there should be no relationship between them and no correlation between the residuals of the observed and expected responses (Andrich & Kreiner, 2010). The software RUMM2030 provides routines to analyse correlations to support the identification of dependence. Where dependence is suspected two further actions can be used for confirmation (Andrich & Kreiner, 2010; Marais & Andrich, 2008a). First, the items can be combined to form one large polytomous item, retaining all the responses. If the PSI is lowered as a result, then the presence of dependence is supported. A second action which may resolve any dependence between items is to create new items based on the actual responses to the items where dependence is suspected, and then check to see if their fit has improved. Any feasible action deemed useful to improve the fit of items to reduce local dependence is considered very worthwhile to improve the accuracy and dependability of the estimates which have been produced. Item creators can also assist by being aware of the influence of local dependence on the determination of the scale, and by checking that the structure and wording of their items and marking keys do not facilitate item dependence.
6.4 Differential item functioning

With invariance built into the model, estimates of item difficulty are independent of the respondent characteristics and there should be no significant differences in the expected scores of the subgroups at the various proficiency levels on the scale (Salzberger, Newton, & Ewing, 2014). In other words, persons of the same proficiency should get statistically equivalent scores on an item. Differential item functioning (DIF) is the term used to describe the presence of significant differences where they are not expected. The graph in Figure 6.8 displays the ICC for Item 36 divided according to gender, red for female and blue for male. The items used in these figures are sourced from my previous study of percentages (Burfitt, 2014). For four of the five classes the expected value in Item 36 at the class mean is similar for males and females as seen by the closeness of the two curves. There is little indication that where males and females have the same proficiency estimate, either gender has a much greater expected score on this item.

Figure 6.8 ICC showing no DIF by gender

By comparison with Figure 6.8, the graph displayed in Figure 6.9 indicates that DIF for year group is occurring for Item 27 which is also from the same study (Burfitt, 2014). For most of the proficiency continuum the ICC for the different year groups are well spaced and the Year 7 and 8 students are consistently outperforming the Year 9 students at nearly all levels on the continuum. At the mean of the second class interval, about 0.15 logits, the expected score for Year 7 students is about 0.55, for Year 8 students is about 0.7 but for Year 9 students the expected score is about 0.3. Clearly some other factor is influencing these expected scores.
When items show DIF they are said to be biased (Taylor & Lee, 2012) because performance is influenced by factors other than the construct. This reduces the dependability of the estimates and hence the scales of item difficulty and person proficiency. In their development of the numeracy scale as described earlier, Looveer and Mulligan (2009) investigated the role of items with DIF as link items. They found that the removal of link items with large DIF led to a better separation of the year groups and to a more obvious developmental scale. They concluded that items, which effectively change their features between tests or year groups, should not be used as link items.

When DIF for gender is detected in items for assessment, any actual differences in achievement by males and females can be distorted and difficult to identify (Salzberger et al., 2014). Andrich and Hagquist (2012, 2015) described a process for resolving real and artificial DIF in a dichotomous item, and this involved creating two items by splitting the item into the subgroups showing DIF. In the case of DIF for gender, this would result in having one item with the responses from the males and then a second item with the responses from the females. Item splitting is done sequentially to all the items showing DIF, starting with the one showing the largest DIF, and before further action is taken, the fit is further examined to determine the result, and the effect on the remaining items.
6.5 Accommodating guessing

As described in Chapter 3, guessing in MC items increases the number of correct responses. Not all persons guess in equal measure, and it is expected that students of lower proficiency will guess more frequently and increase their total scores on a test more than the students at the highest level of proficiency, who may not increase their total scores at all. The bias due to guessing in MC items has been found to also produce bias in the item difficulty estimates of the CR response items located in the same test (Andrich & Maris, 2018). By removing the effects of guessing, more accurate estimates of item difficulty can be determined, and the validity and reliability of the estimates of person proficiency can be improved.

Removing the effects of guessing results in greater accuracy of item difficulty and person proficiency estimates, and it is a fairer outcome when there is no greater advantage from guessing for one group of the individuals who are sitting the test. Guessing may be random or systematic and it is not possible, without further interrogation, to determine the extent of randomness and hence the level of partial knowledge that the student has when they guess the correct response. The techniques which are described in this section, and which relate to the removal of guessing from the responses to the MC items, do not cater for these two different types of guessing. There is no guessing parameter in the Rasch model but processes to identify and eliminate guessing have been developed.

Recognising guessing

Two of the graphs which are produced by the RUMM2030 software (Andrich et al., 2016) can be used to detect the presence of guessed responses: the ICC and the distractor curves. The ICC for an item which shows features of guessing is given in Figure 6.10 (Burfitt, 2014).

![Figure 6.10](image)

*Figure 6.10* Observed responses indicative of guessing
Guessing is indicated on the ICC when the observed responses do not follow the expected model (Andrich et al., 2012). The observed proportions in the class intervals at the lower end of the proficiency scale will be higher than expected, and at the higher end where guessing is less likely, observed scores will be lower than expected according to the ICC. This evidence of guessing is seen in Figure 6.10. When the same scale is used, the ICC would be further to the left when guessing is occurring because the item is easier with the inflation of scores due to guessing. The location where the probability of a person being correct on the item is 50%, is at a lower position on the proficiency scale when the item contains guessed responses.

When there are equal proportions of persons of low proficiency who are selecting each of the distractors, guessing is likely. This suggests that little is known about the item at that level and the persons of low proficiency are selecting an option at random. It follows that this could occur at all levels of the scale as the propensity to guess is governed by the person’s proficiency in relation to the item’s difficulty (Andrich et al., 2012). However, it is more likely to occur at the levels where proficiency is lower.

The selection of distractors in Item 15, as shown in Figure 6.11 indicates that guessing is likely. At the lower level of proficiency, where the observed class mean for person location is about -1.4, one option is virtually ignored and the probabilities for selecting each of the other three is the same, about 0.3. This does not occur at the other class intervals and it implies that the guessing, which is occurring between the correct answer (Option 2) and the other two distractors (Options 4 and 5), is only occurring at the lower end of the proficiency scale.

![Figure 6.11 Distractor selections indicative of guessing](image_url)
Removing guessing

In developing a process to remove the effects of guessing, Andrich et al. (2012) assumed that the level of guessing depended on the relationship between the proficiency of the individual and the difficulty of the item. Persons are more likely to guess the response to a MC item when their level of proficiency on the scale is substantially lower than the item difficulty. This research by Andrich et al. (2012) involved several analyses of the responses. The initial analysis produced estimates for item difficulty and person proficiency, and the results provided information to indicate where guessing was most likely.

In the next analysis, the responses of all persons, who have a lower probability than total random of getting an item correct, were removed for that item and this analysis is referred to as a tailored analysis. Those who are likely to have guessed are those who have a probability of less than 0.25 (on a four-option MC item) of being correct. If there is no guessing in the responses to the items, then the estimates from the tailored analysis will not be significantly different from the estimates produced in the first analysis. Andrich et al. (2012) found that the tailored analysis produced more accurate estimates of item difficulty and person proficiency but noted that not all persons’ responses were included, and this is a concern. A subsequent analysis used these improved estimates, and the responses of all individuals, to generate a proficiency scale to use for reporting the achievement of all persons who sat the test.

To detect any problems with the items, and to examine the fit of the responses to the model, testing should be carried out using the estimates from the tailored analysis which does not include guessed responses (Andrich et al., 2012). Statistical routines provided by the RUMM2030 software can be used to test the significance of the differences between the estimates generated in each of the analyses. These analyses were conducted in my previous study (Burfitt, 2014) and an improved fit to the model was indicated by the change in the fit residuals with tailoring; there were fewer extreme values and a decrease in the standard deviation. When the responses were tailored and the effect on proficiency estimates studied (Andrich & Marais, 2014; Andrich et al., 2015), the more capable students gained more than the less able as they were rewarded for being relatively more successful on what were more difficult items. The authors showed that removing guessing is important for the recognition of the achievement of the more capable students and to reflect more accurately the relative growth across the year groups.
6.6 Partial credit and distractor analysis

Partial credit

In the study of partial credit for CR items mentioned earlier in this chapter, the thresholds for only some of the items were ordered (Van Wyke, 2003). The category probability curves (CPC), shown previously in Figure 6.4, provided evidence of ordered item thresholds. In the CPC shown in Figure 6.12, the thresholds are ordered for Item 17 but not for Item 13. These two items form part of the current investigation of MC items and the CPC are provided to illustrate threshold ordering. Further details of these items are provided in Chapter 7.

For Item 13 there is no range of values on the proficiency scale where a score of 1 is the most likely outcome, and an increasing level of proficiency is not mapped to the probability of an increasing score. The equal probability of getting the two lower scores, 0 and 1, should relate to persons of lower proficiency than those who have an equal probability of getting the two higher scores, 1 and 2. This is not occurring in Item 13 where the thresholds are said to be reversed. Reversed thresholds indicate that persons with higher proficiencies have a lesser amount of the attribute being tested than persons of lower proficiency and this “contradicts the underlying principle of the Rasch model” (Van Wyke, 2003, p. 21).

Figure 6.12 Category probability curves for Items 17 and 13
Distractor analysis

The selection of distractors provides valuable information about the behaviour of students and their understanding of item content. The RUMM2030 software provides distractor curves which enable researchers to study the frequency of option selection across the proficiency scale. In my previous investigation (Burftitt, 2014), the proportion of students in the middle of the proficiency scale who selected the incorrect option for Item 60 was higher than that predicted by the analytical model, and the proportion of students who had proficiency estimates greater than 1 and who provided a correct response, was much lower than that predicted by the model (except for the highest class mean). These observed proportions are shown graphically by the positions of the curves for Options 1 (blue) and 2 (red) as shown in Figure 6.13, where the black curve models the probability of a correct response.

Distractor curve for Item 60

<table>
<thead>
<tr>
<th>Item 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>The price of a maths text book has risen 25% to a new cost of $100. The old price must have been:</td>
</tr>
<tr>
<td>a. $75</td>
</tr>
<tr>
<td>b. $80</td>
</tr>
<tr>
<td>c. $100</td>
</tr>
<tr>
<td>d. $125</td>
</tr>
<tr>
<td>e. None of the above</td>
</tr>
</tbody>
</table>

Figure 6.13  Distractor curves and content structure for Item 60
Regardless of the provision of NOTA as an option, which was later found to contravene guidelines for writing MC items, very few students chose the last three options in Item 60. Students probably knew that the answer had to be less than $100. The proportion of students selecting $75 is high throughout the low and average proficiency range, perhaps thinking that $75 + 25\% = $100. This incorrect selection for the item indicated that the students were thinking of fixed or absolute change rather than proportional change. This suggests that the selection of one of the distractors implies that the students choosing it have more knowledge than those who, while also incorrect, selected the other distractors. Students in the study were not given a score for their selection of this option and it is worthwhile to investigate the feasibility of giving such students partial credit for their selection and to examine the effect this would have on the development of the proficiency scale.

Option 1, as shown in Figure 6.13, would be classified by Andrich and Styles (2011) as possibly an informative distractor, or a distractor with information, and potentially worthy of partial credit as outlined earlier in this chapter. Andrich and Styles (2011) provided a procedure for detecting informative distractors which they developed using simulated then real data. Their real data came from the responses of approximately 30,000 Grade 5 students to 39 Mathematics items which included 25 MC questions. After removing some responses where it was likely that the students had guessed, they divided the population into three classes based on the total scores. Items with distractors in the middle class, and with frequencies better than expected by chance, were considered for rescoring and for the allocation of partial credit for that distractor.

Distractor graphs (as shown in Figure 6.13) were also examined by Andrich and Styles (2011) using 10 class intervals as these provided more detail of distractor selection. Using this process nine items were selected for rescoring with 2 for the key, 1 for the distractor with information and 0 otherwise. The rescored responses, together with the original responses to the remaining items were re-analysed with the polytomous Rasch model. To confirm the presence of informative distractors, Andrich and Styles (2011) examined the relative size of the item threshold values and the changes in the values and probabilities of the chi-square statistic. They also analysed the ICC, CPC, and threshold probability curves.
Polytomous scoring for informative distractors

If a distractor has information, polytomous scoring will provide a better fit to the model than the use of dichotomous scoring. The first threshold (0/1) will have a lower value than the second threshold (1/2) and on the CPC curve there will be a portion of the scale in the middle of the continuum where the probability of scoring 1 is higher than the probabilities of scoring 0 or 2. The observed means will be closer to the theoretical ICC, the chi-square statistic will be lower and its probability higher than with dichotomous scoring. The position of the maximum information value should be located at a lower value on the continuum after rescoring an item with partial credit (Andrich & Styles, 2011). Routines in the RUMM2030 software provide these outputs and Figure 6.14 provides an example of the graph showing the information function produced by this software: In this distribution the maximum value of the information function is 16.12 and it occurs at a location of 0 logits on the conjoint scale.

![Person-Item Threshold Distribution](image)

**Figure 6.14** Graph showing location of maximum information

Andrich and Styles (2011) have provided valuable insight into the statistical and graphic processes that are available for identifying and confirming the presence of such information in distractors. This needs to be endorsed by an examination of the item and the justification of the award of partial credit based on the substantive skills required to select the distractor to be scored. Crediting such information can increase the accuracy of the estimates of person achievement and item difficulty, and it can be likened to having two items instead of one. Andrich and Styles (2011) concluded that creating items with informative distractors would be challenging but the benefits were worth the effort.
Smith (1987) suggests a further benefit likely to eventuate from an attempt to create items with information in the distractors would be an improvement in item creation. Writers would need to consider the difficulty of each distractor in relation to the group of participants likely to make such an error. They would also need to consider the possible effect of distractor difficulty on the level of guessing of both the correct response and of the informative distractor. Sideridis et al., (2016) also studied the behaviour of the distractors with a view to improving construct measurement. They too suggested the qualitative examination of an item to identify the knowledge demonstrated when certain distractors are chosen. They proposed several approaches to remediate informative distractors and recommended the method proposed by Andrich and Styles (2011) as described in this section.

This review of the application of Rasch measurement theory for the analysis of student responses indicates that there are many good products and routines that can be used to report on student behaviour when using MC items. Rasch modelling provides a conjoint scale for the estimates of item difficulty and student proficiency which can be interpreted with confidence when the data fit the model. The RUMM2030 software provides useful and informative displays of the outputs and routines that are suitable for testing the fit of data to the model.

Of particular benefit for the analysis of responses to the items investigated in this study, is the provision of the distractor curves and the category probability curves by the RUMM2030 software (Andrich et al., 2016) when the Rasch model is used for the analyses. The distractor curves can be used to identify the options which contain information, and which indicate the partial knowledge of skills associated with the development of proportional reasoning. In this way, the distractor curves indicate which options should be considered for partial credit scoring, and they can be used to confirm the theoretical presence of information in those distractors. The CPCs provide clear evidence of the presence of ordered thresholds and hence, they can be used to confirm the success of partial credit scoring.

The provision of the scales of item difficulty and person proficiency, as well as the detection of distractors with information, are essential for answering all of the research questions which form the focus of this investigation. To answer these questions, it is necessary to identify distractors with information, develop proficiency scales for achievement, remove guessing from responses in MC items and identify the success of awarding partial credit. The application of the Rasch model to student responses using the RUMM2030 software provides an ideal platform for this study.
CHAPTER 7: RESULTS: Identification of partial knowledge

Introduction

To design and create MC items which have informative distractors, the focus of the first research question, it was necessary to identify the skills and understandings which could be described as partial knowledge of proportional reasoning. From the research studies, I had identified four different types of partial knowledge for reasoning about proportions: additive thinking, ignoring the proportion, providing an incomplete solution, and giving a reasonable estimate. Students think additively when they see increases only as addition and do not recognise the relevance of multiplication. They can make incorrect assumptions or ignore the whole amount when considering only a part of the amount, and they can do part of a task but not complete it. At times, students can estimate a reasonable solution, but they are unable to compute the exact amount. This chapter provides details of an investigation to confirm the existence of partial knowledge in both MC and CR items and this will assist with the identification of the skills that could be used to construct distractors with information.

For the investigation, the responses of students to items that were designed to assess aspects of proportional reasoning were collected and analysed: These secondary data were provided by sources in the United Kingdom and in Australia. Even though the responses had been collected some years earlier, the amount of data available for analysis was substantial and thus the evidence on which the findings were based was deemed sufficient for reliable conclusions to be drawn. Furthermore, all the students who provided responses for analysis were in similar school grades and aged 11–14 years; the same age group as the students doing the online test.

The Rasch model was applied to the students’ responses to these CR and MC items in order to determine the validity of awarding partial credit for partial knowledge, and to identify the effects of scoring partial knowledge on student proficiency: the focus of the second research question. During the estimation of student proficiency, guessing was removed from the students’ responses to MC items and the results provided information to answer the third research question about the effect on the proficiency scale of achievement in proportional reasoning, when guessing is reduced in MC items. While it seems intuitive that removing guessing in MC items and giving credit for the responses which indicate a student is further along the learning pathway will produce more accurate measures, these practices should also satisfy measurement criteria and produce more precise estimates of item difficulty and person proficiency.
7.1 Partial knowledge in constructed-response items

Currently, students have better opportunities to demonstrate their partial knowledge in CR items than in MC items. MC items are not usually written to specifically test student partial knowledge and the options from which students select a response represent a more limited range of answers than the students can generate when constructing their own responses. CR items may be written to provide opportunities for students to demonstrate their partial knowledge, and this is recognised when the response is credited with more than a single mark, and part marks are allocated when the students provide part of the response. This marking system could be further developed if part marks were allocated for a response that represents partial knowledge in the form of a misconception or as one of the types described earlier.

Rather than the student gain part marks only when their solution is incomplete, the student could gain part marks when their response indicates they have part understanding of the concept being tested. This process was shown to be successful by Van Wyke (2003) for an item in which the students were asked to identify the angle through which a bike turned. The correct response was 135° and this was awarded 2 marks while the response of 45° was awarded 1 mark. The latter response indicated that the students could determine angle sizes, but they were unable to recognise which was the angle of turn. Further analysis of CR items, specifically testing the skills necessary for the development of sound proportional reasoning, could confirm the types of knowledge described earlier. The results could provide ideas to use as a basis for writing informative distractors for MC items and this should improve the quality of both the items and the information they generate.

7.1.1 Constructed-response items in the ICCAMS test

The secondary data for this part of the study came from tests which were used in 2008 and 2009 and which tested some of the skills necessary for the development of proportional reasoning. Over the two years, students in Years 7, 8, and 9 (aged 11–14 years) participated in Phase I of the Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS) project funded by the Economic and Social Research Council in the United Kingdom (Hodgen, Brown, Coe, & Küchemann, 2012). The data provided for analysis included the students’ responses to all questions for two of the ICCAMS tests, the codes for all responses and their relevant keys (J. Hodgen, personal communication, December 1, 2016). The records were checked and edited to be read by the RUMM2030 (Andrich et al., 2016) software, resulting in 4950 records available for analysis using the Rasch model.
The ICCAMS tests, provided in Appendix 7.1 consisted of CR items on two topics, ratios and fractions; there were no MC items (J. Hodgen, personal communication, December 2, 2016). For the ratio test, students were asked questions on reasoning about proportions: for example, adjusting quantities in recipes, dividing quantities proportionally, and responding to questions on percentages. One item resembled the question on eels used by Piaget (as cited by Hart, 2004) and another involved determining a third ratio from two given ratios. Scale diagrams were provided, and the students had to use ratios, proportions, or missing value routines to identify missing parts or lengths. With the fractions test, there were items requiring the identification of fractional parts of rectangular shapes and knowledge of the area of a rectangle. Students needed to use subtraction, multiplication, and division with fractions, and to provide missing numbers to make equivalent fractions. Students were not penalised for incorrect spelling nor for the lack of, or incorrect use of, appropriate units.

7.1.2 Partial credit in constructed-response (ICCAMS) items

All 17 questions, or parts thereof on the ICCAMS tests were given an item number as tabled in Appendix 7.2. Most questions and parts of questions were scored as separate items, but this was not logical for Questions 3, 4a, and 6 of the Ratio test and F16 in the Fractions test. In Question 3 and in the parts of Question 6, the correct responses required a combination of two or three numbers (as in a ratio) and it was inaccurate to consider only part of the combination. The incomplete responses to these questions did not indicate partial knowledge because all of the numbers were needed to determine the logic of the student’s thinking. For Questions 4a and F16 it was not the quality of the response to each part of the question that indicated full or partial knowledge, but rather the provision of a correct response to each part of the question.

Three of the analyses of the item responses are relevant for this study. First, all 38 items were scored dichotomously with no credit given for partial knowledge. The estimates produced in the first analysis were needed for subsequent analyses, including a comparison with the estimates produced when partial credit was awarded. Second, items believed to provide opportunities for students to show partial knowledge, based on the content of the item, were scored with partial credit. Third, the items were scored either dichotomously (correct or incorrect) or polytomously (with partial credit) based on the results obtained in the second analysis. For this third analysis the items, for which partial credit scoring did not work and the thresholds were disordered, were changed back to dichotomous scoring. For all analyses, missing responses were treated as if the students had not been presented with the item.
For the second analysis, the responses to all items were classified as being correct, partly correct, or incorrect, according to the expected student behaviour in proportional reasoning that was identified in the literature. Partly correct responses were believed possible for 30 of the 38 items and these were identified as belonging to one of the four behaviours described in Chapter 5: additive thinking (15 items), proportion ignored (5 items), incomplete solution (7 items) and reasonable estimate (3 items). The table provided in Appendix 7.2 gives the responses that were deemed partly correct, and the type of behaviour associated with that response for each of the 30 items. For this second analysis the items believed to have a partly correct response were scored polytomously with 0 (incorrect), 1 (partly correct) and 2 (fully correct) and for the other eight items dichotomous scoring was retained.

Partial knowledge in the form of additive thinking was expected in Item 17 which has been used in several studies of proportional reasoning (Hodgen et al., 2012; Misailidou & Williams, 2003). Item 17 states that Mr Short is 6 paperclips or 4 matchsticks tall while Mr Tall is 6 matchsticks tall. When asked to determine Mr Tall’s height in paperclips, students using additive thinking responded with 8 paper clips. In the second analysis a score of 1 was given for this response of 8 and a score of 2 was awarded for the correct answer. The category probability curve (CPC), given in Figure 7.1, shows that the thresholds are ordered for this item, and the scoring is functioning as expected (Van Wyke, 2003). The threshold at which there is an equal chance that the score is 0 or 1 is lower than the threshold where there is an equal chance that the score is 1 or 2. Persons of higher proficiency are more likely to score 2 than 1, and persons of medium proficiency are more likely to score 1 than 0 or 2.

![Figure 7.1 Category probability curve (CPC) for Item 17](image_url)

*Figure 7.1  Category probability curve (CPC) for Item 17*
Additive thinking was also expected in Item 9 where students were asked to identify the number of sprats needed to feed an eel which was 10 cm long after being told that an eel which was 15 cm in length was fed nine sprats and that the number of sprats depended on the eel’s length. According to Hart (2004, p. 90) this item was adapted for the ICCAMS test from Piaget. If the response was 4 sprats, 5 less than the 9 sprats, then this example of additive thinking was awarded a score of 1 in the second analysis, while the correct response of 6 was awarded a score of 2. The category probability curve seen in Figure 7.2 shows that the thresholds are disordered, and that increasing ability is not associated with an increasing probability of achieving a higher score. Furthermore, a score of 1 is not the most likely score anywhere on the proficiency scale.

The use of polytomous scoring was successful in nine of the 30 items, as indicated by the ordered thresholds in the CPC, so this scoring was retained for these nine items (Items 5, 15–21, 33) and for the other 21 items, dichotomous scoring was re-instated. To determine the effect of using partial credit and to compare the estimates from the two analyses, the scales of item difficulty need to have the same origin and unit, so a fourth analysis was used to generate the required estimates. In this analysis, the mean of the estimates for all items scored dichotomously was anchored (fixed) to the mean of the estimates of those same items as generated in the first analysis. The estimates for item difficulty generated in this fourth analysis were then used to determine the final estimates of person proficiency and to investigate the validity of awarding partial credit.

Figure 7.2 Category probability curve (CPC) for Item 9
7.1.3 Results of awarding partial credit

The summary statistics provided by the RUMM2030 software (Andrich et al., 2016) can be used to indicate if there is any misfit of items to the model. One index of reliability provided by the RUMM2030 software is the person separation index (PSI). The PSI was already high in the first analysis (0.8774) and was slightly higher when partial credit was awarded (0.87821). These high values indicate that the persons are well spread across the ability continuum, and that the power to detect the misfit of the data to the model is very good. This statistic and other summary statistics relating to item and proficiency estimates are provided in Appendix 7.3.

Changes in item difficulty

To confirm that the scales of item difficulty in the two analyses have the same origin, the item locations for the dichotomous items in the first analysis (no partial credit) and the final analysis (partial credit awarded) were compared. The two sets of estimates have a high correlation coefficient ($r = 0.99$) and the graph of their relationship has near-perfect linear slope. Furthermore, this strong relationship had a near-unit slope (gradient = 0.955). This relationship, shown in Figure 7.3, confirms that the two analyses have the same unit with the same origin and any further differences between the estimates from the two analyses can be attributed to the awarding of partial credit.

![Estimates of item difficulty - dichotomous items](image)

Figure 7.3 Item locations for dichotomous items in first and final analyses
In the graph shown in Figure 7.4, the item locations in the first and final analyses are compared for all items. The extra items have been added to the previous graph: those where partial credit scoring has resulted in ordered thresholds. These extra item estimates have been calculated by averaging the lower and upper two threshold values. The lower threshold value is the level of proficiency where there is an equal probability of scoring 0 and 1, and the upper threshold values is the level of proficiency where there is an equal probability of scoring 1 and 2. The average of these two values then represents the level of proficiency where the probability of scoring 1 is a maximum. As seen in Figure 7.4, these averages for the items are much lower when partial credit is given, indicating that the items are easier, and the students’ probabilities of scoring are higher when partial knowledge is awarded credit.

![Estimates of item difficulty - all items](image)

**Figure 7.4** Item locations for all items in first and final analyses

*Changes in distribution patterns*

Histograms showing the spread of persons across the proficiency scale, and the frequencies of item thresholds of different difficulties, are provided in Figure 7.5. The histograms show that the allocation of partial credit has resulted in a better alignment of the items to the proficiencies of the students, particularly in the middle part of the continuum in the range of -1 to 0 logits. Without partial credit scoring the distribution of persons is positively skewed, and with partial credit scoring the distribution is slightly negative but more symmetric about the mean. This change in the shape of the histogram is reflected in a decrease in the skew statistic, and an increase in the value of kurtosis. The values for these parameters are given in the summary statistics in Appendix 7.3
No partial credit

Partial credit

Figure 7.5  Person-item threshold distributions with and without partial credit

Change in the information function

In this thesis the information function, a feature of modern Rasch theory, has been used to demonstrate differences in precision for different situations. The histograms shown in Figure 7.5 are overlaid with the information function which theoretically indicates the amount of information given about the persons, and which in the Rasch model is the inverse of the square of the standard error (Andrich et al., 2016). When partial credit is not scored, the function reaches a maximum of 5.43 where the difficulty estimate is 0.4, and when partial credit is given it reaches a maximum of 7.94 where the difficulty estimate is 0. These maximum values are occurring where the measurement error is minimised. This information
function is redrawn below in Figure 7.6 using the same vertical scale to show more clearly the change in this function with the introduction of partial credit. The relationships are between person proficiency on the horizontal axis and the information function on the vertical axis. The size of the class intervals is 0.2 logits and the mean information function is represented for each class interval in the proficiency range -5 to 5.

![Information functions](image)

**Figure 7.6** Information functions for analyses with and without partial credit

The change in the information function when partial credit is given shows a noticeable increase in the amount of information provided for persons in the middle of the proficiency scale but little difference at either end of the scale. As the graph drawn in Figure 7.6 is based on the mean values for both variables it does not show the exact location and value of the maximum information functions as seen in Figure 7.5. The increased amount of information about person proficiencies which is available when partial credit is awarded, provides support for any conclusions made about the impact of rewarding partial knowledge.
7.1.4 Type of partial knowledge identified in ICCAMS tests

Types of partial knowledge

For the nine items with ordered thresholds, the partial knowledge shown by the students was identified as additive thinking in four items, proportion ignored for two items, an incomplete solution in two items and a reasonable estimate in the other item. The mathematical nature of these nine items, and of the items with disordered thresholds, is summarised in Appendix 7.4. The category probability curve for Item 17, shown previously in Figure 7.1, is typical of the category probability curves for all eight other items scored polytomously in the final analysis.

As well as for Item 17 described earlier, awarding partial credit for additive thinking fitted the Rasch model in three other items, Items 16, 20, and 21. In Item 16, the students had to apply the ratio \( 2 : 3 = x : 5 \) to a diagram and draw a line with length \( x \). While less than 5\% of the students were correct on Item 16, about 40\% of the students drew a line that was 4 units long and it appears that they have added the difference between 3 and 5 to generate the value of \( x \). For the other two items students were given two curved letters which were the same shape but a different size and the ratio of one section to a similar one was 8 : 12. When applying the ratios \( 8 : 12 = 9 : m \) or \( 8 : 12 = n : 18 \), the additive responses were 13 and 14 respectively. Less than 15\% of the students were correct on these two items but about 50\% of the students demonstrated additive thinking with the first ratio and 38\% with the second ratio.

Partial knowledge when ignoring the proportion, was demonstrated in Items 18 and 19 where students needed to determine unknown proportions. To identify one such proportion, students were given the ratios 1A : 5B and 3C : 10B, and were asked to calculate the quantity of A to match a given quantity of C. By indicating that the ratio of A : C = 1 : 3, students were ignoring the quantity of B in the two given ratios. Item 33, where partial credit was given for an incomplete solution, consisted of two parts and the first part was quite easy as it involved subtracting a simple fraction from 1 but in the second part the subtraction was from another fraction. In Item 15, also an incomplete solution, the response was identified as being incorrect because the line of a shape was drawn to scale but the gaps between the lines had not been scaled. In Item 5, where partial knowledge was demonstrated by providing a reasonable estimate, students had to state the amount of cream in a recipe for 6 people when the recipe stated that half a pint was needed for 8 people. Many students gave a unit fraction smaller than a half and such responses were scored with partial credit.
The items for which granting partial credit was successful, represented all types of partial knowledge and there was no discernible pattern in the nature and content of the items (Burfttt, 2017). It is possible that the non-integral nature of the ratios, which were 2 : 3 and 8 : 12 and for which one number is not a multiple of the other, triggered additive thinking in all four items where partial credit was justified for additive thinking. With these ratios, students may have difficulty recognising a scale factor and thus resort to addition rather than multiplication.

*Partial knowledge not confirmed*

In Items 6 to 13, derived from the studies of Piaget (as cited by Hart, 2004) and where students were asked to determine the number of sprats to feed eels of various sizes, additive thinking was considered possible. However, the award of partial credit was not justified when polytomous scoring was applied. It is possible that the students found the scale factor, though not integral, was more familiar as it was based on multiples of five (5, 10, 15) and this may have triggered multiplicative rather than additive thinking.

For other items where students had to create equivalent fractions, partial knowledge was not confirmed. Equivalent fractions are introduced in the middle primary years and reviewed annually; the content should be familiar for the students doing the test. It is more likely that students would have used additive thinking in relating the numerator to the denominator when they initially learned this concept, rather than at this later stage of their learning. All questions relating to percentages were best scored as correct or incorrect: Where either the proportion was ignored, or the solution was incomplete, awarding partial credit was not justified. It is possible (Burfttt, 2017, p. 5) that as students “learn about percentages any misconceptions that develop are not linked to knowing that percentages represent proportions.”

Opportunities for students to demonstrate partial knowledge by giving an incomplete solution was considered possible in three quite challenging questions on fractions. A qualitative inspection of these items gives some insight as to why partial knowledge was not confirmed. In two of the items, Items 37 and 38, the use of fractions for the dimensions of the rectangles would have made it difficult for the students to apply their understanding of rectangle area. Solving the problems were difficult because they initially required the students to recognise the product of two fractions. Had these values been whole numbers, it is possible that the students would have been more successful in solving the problems. These thresholds were naturally reversed as the first action, the identification of the operation, is relatively more difficult than the second action (Van Wyke, 2003).
7.1.5 Further outcomes

The proportion of students who were given partial credit for items which they attempted is considerable and in the ICCAMS tests the proportions of students receiving partial \((s = 1)\) and full \((s = 2)\) credit is shown in Table 7.1 where the items are listed in order of difficulty. There are many students who would benefit from the extra scores awarded for partial knowledge and such recognition would be well received by the students if the practice of granting these extra scores were routine in tests of mathematical concepts.

**Table 7.1 Proportions of students receiving partial and full credit**

<table>
<thead>
<tr>
<th>Item</th>
<th>18</th>
<th>33</th>
<th>17</th>
<th>21</th>
<th>5</th>
<th>19</th>
<th>20</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s = 2)</td>
<td>0.29</td>
<td>0.27</td>
<td>0.24</td>
<td>0.14</td>
<td>0.14</td>
<td>0.12</td>
<td>0.11</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>(s = 1)</td>
<td>0.34</td>
<td>0.50</td>
<td>0.49</td>
<td>0.38</td>
<td>0.47</td>
<td>0.31</td>
<td>0.49</td>
<td>0.23</td>
<td>0.40</td>
</tr>
</tbody>
</table>

The easiest items, as seen in Appendix 7.5 where the items are listed in order of difficulty, are some of the dichotomous items. For most of the items where the award of partial credit is justified, the skills associated with partial knowledge were located in the middle of this ordered list while the skills for the correct response had the highest difficulty measures. The only item not following this trend was Item 15 for which the skill identified as partial knowledge was quite difficult. The differences between the first and second threshold values, referred to as threshold differences or threshold distances, reflect the differences in difficulty between the skill required to demonstrate partial knowledge and the skills required to provide a correct response. The significance of the values of these threshold differences, which range from 0.847 logits (Item 15) to 3.12 logits (Item 20), is discussed in the following section where partial knowledge in MC items is considered.

For the ICCAMS tests, partial credit has been awarded in some items where the answer was not correct but where it indicated a possible stage of development or a misconception that indicated partial knowledge. The allocation of partial credit particularly benefited students in the middle of the proficiency range. Given that this scoring of partial knowledge is justified, we can claim that the resulting estimates of student proficiency are more accurate than if all responses had been scored dichotomously. The results of this part of the investigation have provided valuable diagnostic information about the features of partial knowledge of the skills associated with the development of proportional reasoning, and of the number of students with this partial understanding of concepts. Such information can be used to inform and enhance decisions relating to curriculum development and the planning of learning activities.
7.2  Partial knowledge in multiple-choice items

MC items are generally marked dichotomously, and a score is allocated only when students choose the correct option. All student responses to the distractors (incorrect options) are usually allocated a zero mark. Examples of MC items, where the students have found some distractors more attractive than others, have been provided previously. At times, this has occurred when one distractor, either an informative distractor or a distractor with information, resembles the correct option because it has more aspects of the correct response than the other distractors. By giving a score for this partial knowledge the amount of information gathered from students could be increased without changing the test conditions.

To investigate the presence of informative distractors in MC items currently used in tests, I analysed student responses to items from the numeracy sections of the national tests, the NAPLAN assessments. I chose to analyse the responses of students in Years 7 and 9 because of my considerable experience in teaching students of this age group, and because I was interested in studying their achievement. One aim of the analysis was to identify information which could be useful for the creation of MC items to provide opportunities for students to display their partial knowledge: the focus of the first research question. Another expectation was the identification of an analytical process to detect and measure this partial knowledge.

7.2.1 Conjectures about existing MC items

There is no evidence to suggest that the NAPLAN numeracy items are purposefully written to detect or score partial knowledge, nor is there evidence that the students’ scores are adjusted to remove possible inflation due to guessing. Considering this situation and the findings from research into the use of MC items to assess student achievement, four conjectures relating to improving the design, creation, and analysis of MC items, were investigated. The first conjecture is that there are already items in these tests that contain incorrect options for which students should receive a proportion of the item’s score. The second conjecture is that giving credit for this partial knowledge provides estimates of person proficiency which are more accurate and more precise. The third conjecture tested in this study is that the behaviour of the distractor curves can indicate where partial knowledge exists and can thus be used to indicate where credit should be awarded. Finally, the perception that males outperform females on MC items because they tend to guess more often is the focus of the fourth conjecture; namely, removing the effects of guessing and awarding partial credit, reduces the perceived disadvantage for girls in tests containing a large proportion of MC items.
By investigating these four conjectures, the findings can be used to formulate answers to the research questions which are integral to this study. The first and third conjectures relate to the first research question which is aimed at identifying how to design and create MC items with informative distractors. It is necessary to identify and construct informative distractors which can be used to improve the quality of information gathered when MC items are used in assessments of mathematics. The other two conjectures relate to the improvement in the proficiency scale which can be achieved when items are scored with partial credit and when guessing is reduced. The identification of ways to improve the proficiency scale is one of the aims of answering the second and third research questions.

7.2.2 The NAPLAN tests

The data available for the analysis consisted of the responses by the students in Years 7 and 9 to the Numeracy assessments for NAPLAN in 2013 and 2014. The items also were included in the material to be analysed. The data, which were provided by ACARA (A. Dow-Sainter, personal communication, February 4, 2016) included the responses of several thousand students to the two sections of the numeracy tests; one where calculators could be used, and the other where calculators were not permitted. The random samples taken from each of the four test populations provided over 2900 records to analyse for each test. The records included many factors but only two of these, the gender of the students and their responses to the items, were extracted from the data set. There were 64 items in each test and the number in MC format in any one test varied between 46 and 50. For the remaining items which were in CR format, the students needed to enter a number in response to a stimulus.

The NAPLAN items are designed to test the numeracy achievement of students throughout Australia in relation to the concepts, skills, and understandings described in the content of the Australian Curriculum: Mathematics (National Assessment Program, 2016). The three content areas, number and algebra, measurement and geometry, and statistics and probability are assessed within the four proficiency strands; understanding, fluency, problem-solving, and reasoning. While the focus of this component of the research is student performance in MC items which test the skills necessary to develop sound proportional reasoning, the responses to all items are included to obtain the best possible estimates of student proficiency.
7.2.3 Analysis of MC items

Four tests were available for each analysis using the Rasch model. Each test contained a total of 64 items which were either in MC or CR format and came from both sections of the NAPLAN assessment, that is the calculator and the non-calculator sections. The four sets of responses were from the students in Years 7 and 9 in both 2013 and 2014. The responses to all items were retained in all analyses in order to generate the most relevant measures of student proficiency but the results reported in this study only relate to the MC items which assessed the skills associated with the development of sound proportional reasoning.

For all the MC items associated with proportional reasoning, the distractors were examined to identify which ones, if any, could enable the students to demonstrate partial knowledge of the concept being tested. The type of partial knowledge that the students could show was classified according to the various mathematical behaviours described earlier, for example, additive thinking. This subjective analysis was informed by the research findings in the literature and by my own experience of teaching this content to students of the same age as the test participants. From the qualitative inspection of the content of the 66 MC items relating to proportional reasoning in the four tests, 25 items were described as definitely having an informative distractor, 32 items as possibly having an informative distractor and the other nine items as not having an informative distractor at all.

*Using distractor curves to identify informative distractors*

Following the qualitative examination of the distractors, the distractor curves produced by the RUMM2030 (Andrich et al., 2016) were examined to identify the potential for each item to allow the students to demonstrate their partial knowledge. All responses to the MC items were scored dichotomously, and from the behaviour of the distractor curves, the expected existence of partial knowledge in each item was classified as being definite, possible or not at all. This classification of the distractors was also subjective, but it was based on the findings of Andrich and Styles (2011) and Sideridis et al. (2016). Partial knowledge may be indicated when the curve for one of the distractors peaks for students in the middle ability range. It may also be indicated when the more able students are distracted from the correct answer and the curve showing the observed responses for the correct response is lower than that expected according to the model. At the same time the curve for one of the distractors is higher than for the other distractors. Some of the research of the identification and relevance of the distractor curves was outlined in Chapter 6.
The distractor curves for four of the items are shown in Figure 7.7. For the first item, Item 23, the blue curve for the correct response follows the theoretical model and at the same time the red curve for Option 2 appears to resemble the behaviour of an informative distractor. The red curve rises until the person location, which is the estimate of proficiency, is about -1 and remains relatively high until it descends quickly after the person location reaches about 0.8.

For Item 11, the blue curve for Option 1, the correct response, is lower than that predicted by the model after a proficiency estimate of about 1. At the same time the red curve for Option 2 suggests that the option represented by this curve is attractive to students at the higher levels of proficiency. Together, these two curves suggest the possibility that Option 2 is an informative distractor. For the other two items shown in Figure 7.7, Items 10 and 13, the distractor curves did not appear to match either description of informative distractors.

---

**Figure 7.7** Distractor curves indicating various levels of information

---

<table>
<thead>
<tr>
<th>Item 23</th>
<th>Item 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informative distractor: deemed highly likely</td>
<td>Informative distractor: deemed possible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item 10</th>
<th>Item 13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlikely to contain an informative distractor</td>
<td>Unlikely to contain an informative distractor</td>
</tr>
</tbody>
</table>

---

**Item 23**

Informative distractor: deemed highly likely

---

**Item 11**

Informative distractor: deemed possible
Informative distractors in the NAPLAN assessments

As reported earlier, the qualitative inspection of the 66 MC items which assessed skills associated with proportional reasoning indicated that there were 57 items with potentially informative distractors: 25 definitely and 32 items possibly. Only item content was considered for this classification.

These 57 items were then allocated to one of three groups according to the strength of the evidence from the distractor curves: evidence as described in the previous section. Based on the appearance of the distractor curves, the items were described as highly likely, possibly, and as unlikely to contain informative distractors. The curve for Item 23 (Figure 7.7) indicated that an informative distractor was highly likely while the curve for Item 11 (Figure 7.7) indicated that an informative distractor was possible. Items 10 and 13 were assigned to the group where the demonstration of partial information was considered unlikely because the distractor curves did not have notable peaks in the middle proficiency range, and the curve followed the theorised model at the higher proficiency levels of the continuum.

Of the 57 items with potentially informative distractors, as classified according to item content, there were 25 items described as highly likely, 20 items as possibly, and 12 items as unlikely to contain informative distractors based on the appearance of the distractor curves. These curves indicate the potential for recognising the existence of informative distractors. While this recognition of information in the distractor curves provides an opportunity to confirm and hence assess partial knowledge, caution is recommended. Using only the behaviour of the distractor curves without any consideration of item content is not widely used and more research is needed to confirm the reliability of such an approach.

The distractor curves can also be used to indicate which items lack informative distractors and hence, which items are unlikely to be suitable for partial credit scoring. An inspection of the distractor curves for the nine items which had been initially identified as lacking an informative distractor based on item content confirmed this status. These items were scored dichotomously and included in the determination of the item and person estimates but are not included in further discussion. The other 57 items, where informative distractors were indicated, were scored for partial credit, that is, polytomously, with a score of 2 for the correct response, a score of 1 if the informative distractor had been selected and zero otherwise. The category probability curves for these items were examined to determine the presence of ordered thresholds and hence, the success of polytomous scoring.
A summary of the success of the partial credit scoring, in relation to what was expected given the nature of the distractor curves, is given in Table 7.2. Without reference to item content, the distractor curves were first examined and then classified as belonging to one of three groups: the item having a strong, a medium, or no indication of partial knowledge. Having a strong indication of partial knowledge (PK) is linked to the description of the item being highly likely to contain an informative distractor whereas, no indication of PK is linked to the item being unlikely to contain an informative distractor. After the 57 MC items were allocated to one of the three groups, the category probability curves (CPC) were inspected to see if partial credit scoring had been justified: that is, the thresholds were ordered. The data in the table shows the strong link between the behaviour of the distractor curves and the success of partial credit scoring: Partial credit scoring is more likely to be successful when there is a strong indication of an informative distractor in the graph of the distractor curves.

<table>
<thead>
<tr>
<th>Evidence from distractor curves</th>
<th>Total number</th>
<th>Ordered thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Proportion</td>
</tr>
<tr>
<td>Strong indication of PK</td>
<td>25</td>
<td>22</td>
</tr>
<tr>
<td>Medium indication of PK</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>No indication of PK</td>
<td>12</td>
<td>2</td>
</tr>
</tbody>
</table>

PK = partial knowledge

**Category probability curves**

In the category probability curves for Items 23, 11, and 13, shown in Figure 7.8, the thresholds are ordered and there is a range of proficiency on the continuum where a score of 1 is more likely than a score of 0 or 2. This range is the distance along the horizontal axis between the two intersections involving the red curve. The first intersection occurs at the first threshold where the curve showing the probability of a score of 0 crosses the curve showing the probability of a score of 1. The second intersection is at the second threshold where the curves modelling the probabilities of scores of 1 and 2 cross. This range was described as the difference between thresholds for the CR items reported in the earlier part of this section and is henceforth referred to as the threshold distance. It is noted that the threshold distance, which is described in more detail later in this chapter, is greater for the MC items which are considered highly likely to have an informative distractor, than for the items described as only possibly having an informative distractor.
Ordered thresholds were observed as expected for Items 23 and 11 as there was evidence in both the item content and in the distractor curves (seen in Figure 7.7) of the presence of informative distractors. The disordered thresholds were expected in Item 10 because there was no evidence of the presence of informative distractors. The presence of ordered thresholds in Item 13 was unexpected, given the appearance of the distractor curves, but on further examination of the curve (Figure 7.7), it appears that the relevance of the size of the peak in the middle of the proficiency continuum has been underestimated and this is noted for the future identification of informative distractors.

Graphs showing the ordered and disordered thresholds in these category probability curves, for the four items with distractor curves drawn in Figure 7.7, are given in Figure 7.8. For all items, where the thresholds were disordered and polytomous scoring was evidently not working, the items were rescored dichotomously for all subsequent analyses.

**Figure 7.8** Category probability curves with ordered and disordered thresholds
Evidence of guessing

One of the research questions for this study concerns the effect of guessing on the measures of person proficiency. To identify this effect, it is necessary to have a process by which guessing is identified, either by observation or from a theoretical perspective. Observations of guessing can be made by inspecting the distractor curves. For example, the curves for Items 13 and 16 in the 2013 Year 7 NAPLAN numeracy assessment, shown in Figure 7.9, provide evidence of guessing among students of lower proficiencies. The proportion of students selecting each of the options is similar, ranging from about 0.16 to 0.3, at the beginning of the scale where the proficiency is about -2 logits. At this level of proficiency, it appears that all distractors are equally attractive to the students and this suggests that guessing is occurring.

![Item 13: Distractor curves](image1)
![Item 16: Distractor curves](image2)

_Figure 7.9_ Distractor curves indicative of guessing

To investigate the effects of removing guessing from the estimates of item difficulty and person proficiency, the identification of the extent and location of guessing was determined theoretically. Tailored analyses, such as described by Andrich et al. (2012) and as outlined in Chapter 6, were conducted for all MC responses in the four tests. In these tailored analyses, the responses of the students likely to have guessed on the item were recorded as missing; as if the students had not been presented with the item. Using the output from the first analysis, where all items were scored dichotomously, a tailored analysis was conducted before the analyses involving polytomous scoring, rescoring items where polytomous scoring did not work, and anchoring on the original dichotomous tailored responses were undertaken. Thus, the same series of analyses were applied to the data which included the guessed responses and to the data with the guessed responses removed. A summary of all analyses conducted on the NAPLAN data, and as described in this section, is given in Figure 7.10.
Figure 7.10  Summary of analyses of NAPLAN data
7.2.4 Effects of awarding partial credit in MC (NAPLAN) items

For each of the four cohorts (Years 7 and 9 in both 2013 and 2014), the outputs from granting partial credit were compared with those where there was no partial credit, and this was done for responses with guessing and responses without guessing. This resulted in eight situations for examining the effect of granting partial credit. For each situation, the scales for the values of estimates were set with the same origin, in an anchoring process which is further described in Chapter 6. This meant that the differences in the outputs with each comparison could be attributed to the polytomous scoring of items with ordered thresholds.

The PSI was quite high in every analysis, varying from 0.9261 to 0.9336 so the power to detect misfit to the model can be described as excellent. There was no pattern to the changes in the values of the PSI when partial credit was awarded and the largest change in value was only about 0.001. As expected, granting partial credit resulted in a decreased mean item difficulty estimate in all cases, a small decrease in the item standard deviation for seven of these, but no pattern in the change to the mean item fit residual. The different statistics for these eight situations are provided in Appendix 7.6.

Changes to the distribution of person proficiency

The effects of granting partial credit are best seen in the changes to the person-item threshold distributions. Without partial credit the items were well aligned with student proficiency and for all eight situations the frequency of item thresholds, in the part of the continuum just below the centre of the distribution, increased when partial credit was awarded. This improvement in the alignment suggests that the recognition of partial knowledge provides scoring opportunities that are even better targeted to the students’ proficiencies, particularly for the students whose proficiency estimates are just below the median estimate.

With partial credit the distributions of person proficiencies all show an increase in the skew and kurtosis values: The distributions are more peaked and there is increased symmetry about the mean. As expected with this narrower spread of frequencies, the standard deviation for all persons, as well as for males and females separately, decreased when partial credit was awarded.

The person-item threshold distributions for all eight situations where NAPLAN items have been analysed are provided in Appendix 7.7 and two of these histograms, which are typical of all eight graphs, are reproduced in Figure 7.11.
For these MC items from the NAPLAN tests, as for the CR items described in the earlier section of this chapter, there is a notable change in the information function and its location when partial credit is given for responses. Similarly, the greatest benefit of granting partial credit is for the persons in lower half of the proficiency scale, though not for the persons with the lowest proficiencies. For the MC items as well as the CR items, rewarding partial knowledge increases the amount of information gathered about the persons, and this results in more precise measures of their proficiencies.

### Year 7, 2013: No partial credit

![Person-item Threshold Distribution](image1)

### Year 7, 2013: Partial credit

![Person-item Threshold Distribution](image2)

*Figure 7.11*  Person-item threshold distributions with and without partial credit
Considering guessing

The removal of guessing from the responses appeared to reduce the advantage for males. In an analysis of the Year 7 responses to the 2013 NAPLAN numeracy test, the responses were tailored to remove those responses likely to have been guessed. This part of the study was conducted on responses where partial credit was not considered and all scoring was dichotomous. After tailoring, the estimates were anchored to allow the comparison with the original estimates. The mean proficiency estimate for males was significantly higher at the 1% level, with or without the guessed responses, but the difference in the proficiency estimates of the males and females was lower when the guessed responses were removed. This effect was worth noting for further investigation when analysing the data from the student test.

With the guessed responses included, the mean person proficiency decreased when partial credit was awarded, but with the guessed responses removed, the mean person proficiency increased when partial credit was awarded. Most of these changes were quite small. When considering males and females separately, there was no pattern in the changes to mean person proficiency when partial credit was awarded.

The removal of guessed responses needs to be considered with some caution. Interesting patterns many be observed, and significant results may be generated but it would not be well received by students and educators if not all of an individual’s responses were included for the determination of any measure of proficiency or for the ranking of any performance.

One result which clearly applies regardless of the presence of guessing in the response and that is the effect of granting partial credit on the maximum value of the information function and its location; the value of the information function increased considerably in all eight situations and its location on the proficiency scale in each situation was further to the left. This effect occurred for both year groups, in both calendar years as well as with and without the guessed responses in the data. This important finding relates to the effect on the proficiency scale when distractors with information about student learning are awarded partial credit: the focus of the second research question for this study.
Threshold distances

Granting partial credit does not affect all items equally, and the probabilities of obtaining scores of 0, 1, and 2 vary between items, as seen in the category probability curves in Figure 7.8 which was presented previously. The threshold distance, that is, the range on the continuum where the probability of scoring 1 is higher than the probability of scoring 0 or 2, varies between items, and this distance can be expressed in logits. When guessed responses were removed from the analyses, the item thresholds differed slightly but there was no change in the threshold distances for any of the polytomous items. A summary of the values of the threshold distances, and the number of partial credit items in proportional reasoning within each range of values, are provided in Table 7.3. The ranges of values were chosen to categorise the threshold distances reflecting the varying probabilities of persons of fixed abilities achieving particular scores.

When the threshold distance is about 1.1 logits, a person of ability \( \beta \) has a 50\% probability of getting a score of 2 rather than 1, and a 75\% probability of getting a score of 1 rather than 0 (Andrich, 2016). The odds of this person receiving a score of 1 rather than 0 is 3 to 1. A threshold distance of about 0.7 logits equates to the 50\% probability of getting a score of 2 rather than 1 and a 67\% chance of getting a score of 1 rather than 0 so the odds of scoring 1 rather than 0 are 2 : 1 while the odds of scoring 2 rather than 1 are at 1 : 1. When the logit difference is 0.4, the probability of getting a score of 1 rather than 0 is 60\%.

### Table 7.3  Threshold distances in items scored with partial credit

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Total number of items</th>
<th>Number of items with ordered thresholds</th>
<th>&gt;1.1</th>
<th>0.7–1.1</th>
<th>0.4–0.7</th>
<th>&lt;0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y7 2013</td>
<td>13</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Y9 2013</td>
<td>12</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Y7 2014</td>
<td>13</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Y9 2014</td>
<td>19</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>57</td>
<td>30</td>
<td>10</td>
<td>5</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

A good proportion of the items with ordered thresholds had threshold distances greater than 0.4 and it would seem logical that having two quite distinct thresholds values for these items would be equivalent to having extra items in the test. Further discussion of this suggestion occurs later in this report.
7.2.5 Implications of recognising partial knowledge in MC (NAPLAN) items

As with the responses to the CR items in the ICCAMS tests, the establishment of measurement scales with the same origin facilitates a comparison of the results of the various analyses and some of the differences in the outputs have been attributed to the granting of partial credit when scoring the responses. The conjectures regarding the allocation of partial credit for partial knowledge in MC items, as described earlier, can be thus be addressed.

*Partial knowledge in NAPLAN items*

The claim in the first conjecture is that the NAPLAN tests already contain items with incorrect options for which students should receive a proportion of the item’s score. The evidence gathered in the analysis of the students’ responses, those with likely guessed responses retained as well as those where guessed responses were removed, provides strong support for this proposal. As seen in Table 7.3, there were between six and nine items out of a maximum of 19 items in each test where polytomous scoring was used and the thresholds were ordered. These items constitute between 42% and 47% of the total number of MC items which assess the skills of proportional reasoning in each test. For items where the thresholds are ordered, polytomous scoring should be retained (Andrich & Styles, 2011) and the students should receive credit for their partial knowledge of the item’s content.

As well as polytomous scoring being more accurate for these items, and the rewarding of partial knowledge more appropriate, there is an increase in the amount of information available about student achievement when partial credit is awarded. Where the difference between ordered thresholds, that it, the threshold distance, is greater than 1.096 logits (D. Andrich, personal communication, September 20, 2017), it is equivalent to having two distinct items in the test as opposed to just the one. Of the 30 items successfully scored with partial credit, there were 10 items for which this distance was greater than 1.096. This is akin to increasing the opportunities to collect data about student achievement without increasing the demands of the test.

Research studies which describe the impact of threshold distance have not been located in the research literature and a comparison of the threshold distance was not investigated further in this study. Knowing the level at which this difference significantly affects the amount of information that is gathered, could provide further insight for improving both the creation of MC items, and the analysis of responses to the items.
MC items which test skills relating to proportional reasoning only constitute 35% of the total number of MC items in the four tests and the evidence for the existence of items where partial credit should be awarded is strong. It follows that some of the MC items which test other mathematical concepts are also likely to provide similar opportunities for students to demonstrate, and be rewarded for, their partial knowledge.

*Greater precision of estimates with partial credit awarded*

The results of this part of the study provide considerable support for the second conjecture: that giving credit for partial knowledge produces estimates of item difficulty and person proficiency which are more precise than when partial knowledge is not scored. As with the ICCAMS responses, awarding partial credit increases the number of aligned thresholds for persons of low to middle proficiency without a commensurate reduction in the thresholds which are aligned to persons of high and very low proficiency. This improvement in alignment, and the consequent increase in the amount of information available for persons with proficiency estimates in the middle of the continuum, supports the anticipated increase in the precision of the measurement achieved when partial knowledge is rewarded. Greater precision is obtained when partial credit is awarded, and the thresholds are ordered. When the distractor has information and the data fit the model for polytomous scoring, it is then “contradictory to the model” for dichotomous scoring to be used (Andrich & Styles, 2011, p. 69). It follows that when credit is awarded for the selection of informative distractors then the estimates thus generated are more precise than under a dichotomous scoring regime.

*Distractor curves indicate partial knowledge*

The third conjecture states that the behaviour of the distractor curves can indicate where the partial knowledge exists, and these curves can thus be used to indicate where credit should be awarded. The curves display the behaviour of the students in their selection of the options and can be used to identify informative distractors. The results of the distractor analysis support this conjecture. Over all four tests there were 25 items where the distractor curves were described as strong indicators of the presence of an informative distractor and this presence was confirmed by the successful polytomous scoring in 22 of these items. Examples of distractor curves which feature informative distractors are provided in Figure 7.12 where there is one item provided for each of the four threshold distances described in Table 7.3.
In each of the graphs shown in Figure 7.12, the curve representing the informative distractor is notably higher than for all other distractors. The obvious peaks in the curves of the informative distractors are below the middle of the continuum and this is evident in all four items. This observation supports the research findings by Andrich and Styles (2011) that this behaviour indicates the presence of information. The distractor curves for the other 18 items which arise from all four NAPLAN assessments and which are presented in order of threshold distance are provided in Appendix 7.8.

**Figure 7.12** Distractor curves for items with ordered thresholds
There were 22 items with distractor curves which clearly indicated the presence in the item of an informative distractor, and with successful polytomous scoring. For all of these items, the highest curve for any distractor in each of the graphs shown in both Figure 7.12 and Appendix 7.8 represents the option chosen for partial credit. This was the option chosen as a result of an examination of the item content, and with consideration of the partial knowledge skills of proportional reasoning, without any reference to the students’ responses to that item.

In all 22 items, the curve for this informative distractor is notably higher than the curves for the other distractors in the lower to medium range of the continuum. In about half of the items the curve for the correct response is lower than that predicted by the model at the higher estimates of person location, as seen in Item 20 of Figure 7.8. This indicates that the informative distractor is also attractive to students of the highest proficiencies. The prominence of these two features in items with ordered thresholds supports the conjecture that the distractor curves can be used to identify opportunities for scoring partial credit.

*Partial knowledge not confirmed*

Three of the 25 items which were considered as highly likely to be suitable for partial credit scoring because of the appearance of the distractor curves, which are shown in Figure 7.13, had thresholds which were not clearly ordered. With Item 18 the category probability curve indicated that the difference between the disordered thresholds was about 0.5 logits and the curves for Items 39 and 43 showed coinciding thresholds. These items had been considered suitable for partial credit because the red curve of the proposed informative distractor (Option 2) is higher than the curves of the other distractors in the low to middle proficiency range.

It is possible that the behaviour of the red distractor curve at the upper end of the continuum also needs to be considered. In all three examples this curve is near the horizontal axis at proficiencies greater than 1 and this suggests that this option is not at all attractive to students of high proficiency. The curves modelling the behaviour of the other two distractors may also indicate partial credit is not appropriate. In Item 18 a second distractor is also attractive for a large proportion of students in the lower half of the continuum and in Items 39 and 43 the curves show that the other two options are not attracting a large proportion of students anywhere along the continuum.
These other aspects of the behaviour of the distractor curves are not considered in this study but future research in this area would build on the findings of this study and further the use of distractor curves in the detection of options for which partial credit scoring is warranted.

**Figure 7.13** Distractor curves for items with disordered thresholds (Items 18, 39, 43)
Debatable graphic evidence of informative distractors

There were two items which, based on the items’ content, were thought to provide opportunities for students to show partial knowledge but after the inspection of their distractor curves, as shown in Figure 7.14, this was considered less likely. When polytomous scoring was applied, the thresholds were ordered and the distances between the thresholds were considerable at 0.7 and 0.8 logits. In both items, the informative distractor was quite an attractive distractor in the early part of the continuum, but not elsewhere.

![Figure 7.14 Distractor curves for items with thresholds ordered but not expected](image)

There were 20 items which were initially considered for partial credit and for which the distractor curves were identified as only possibly indicating the presence of an informative distractor. Only six of these items had ordered thresholds and the threshold distances ranged from 0.1 to 0.9. The curves for the potentially informative distractors in some of these items were lacking distinct peaks at the lower or middle sections of the continuum but they were below the theoretical curve at the higher end of the continuum. For other items the peak was obvious in the early part of the continuum, but the curves showed that the potentially informative distractor was not attractive to students of the highest proficiencies.

From this study of the behaviour of the curves of the distractors, it appears that information can be indicated when the curve for a distractor shows a distinct peak in the early part of the continuum, or the curve for the correct response is below the theoretical curve. While this link between the distractor curves and successful partial credit scoring is strong, these results indicate that further research is needed to provide more support for this conjecture if only the nature of the curve is used to indicate the presence of information.
**Consideration of gender**

For each year group in both 2013 and 2014, the performance of males was significantly better than that of females ($p < 0.001$) regardless of the award of credit for partial knowledge and regardless of the inclusion or otherwise of the guessed responses. There was little evidence to suggest that awarding partial credit reduced the perceived advantage for males in these NAPLAN tests. When guessed responses were included the effect of granting partial credit was to reduce the differences between the mean proficiencies for males and females but these differences were small, varying from 0.004 to 0.025 logits. In comparison, there was no pattern in the differences between the mean proficiencies for males and females when partial credit was granted, and the guessed responses had been removed. These results do not support the fourth conjecture, namely; the application of partial credit scoring to responses where guessing has been removed reduces the perceived disadvantage for girls in tests containing MC items.

These NAPLAN numeracy tests have a high proportion of items of the MC format which are believed to favour males because of the perception that males are more inclined to guess. A decrease in the differences between the mean proficiencies of males and females would thus be expected when guessing is removed, but there was no strong evidence of such a decrease. It is possible that a consideration of all MC items, rather than just those relating to proportional reasoning, for partial credit may have resulted in a different conclusion. The mean estimate for male proficiency was much higher than that for females in all eight situations and this indicates that this gender discrepancy may be caused by factors other than those studied in this analysis.

**Link items**

Ten of the items for which partial credit scoring was successful were link items: They occurred in the tests for both Year 7 and Year 9 students in the same calendar year. For three of these ten items, partial credit scoring was successful for both groups of students. Partial credit scoring of link items, and the use of responses other than the key in the determination of item estimates, can be applied to reduce the number of link items needed to achieve precision of item difficulty and person proficiency estimates (Kim, 2006). Further investigation of these features could further the study of the identification of guidelines for the creation of future link items.
Types of partial knowledge

The types of partial knowledge revealed in the NAPLAN tests resembled those identified from both the research literature and the analysis of the CR items in the ICCAMS tests. One further classification, Incorrect method, was added to describe three items which did not fit into any of the earlier categories. An example of this type of partial knowledge occurred in the recognition of the response to the sum of 0.3 and 0.99; partial knowledge was recognised when the students selected 1.02 for the correct answer. A summary of the partial knowledge skills and their types, as demonstrated by the students in the NAPLAN tests, is provided in Appendix 7.9. The most common type of partial knowledge identified in these tests belonged to the type described earlier as proportion ignored: This applied to 12 out of 30 items.

Inspecting the content of items for which partial credit scoring was justified provided further information for the future construction of MC items. A question of scale allowed students to demonstrate the type of partial knowledge classified as a reasonable estimate. In this question, Item 52, a model of the Sydney Harbour Bridge had a scale of 1 : 6000 and the length of the actual bridge was given as 1200 m; the options were 5 cm, 20 cm, 200 cm and 720 cm. Partial credit was given for the selection of 200 cm: 10 times the correct response and a reasonable estimate. Giving partial credit for the selection of 200 cm recognises that the students have used division and considered the different units, possibly using 10 cm as 1 m rather than 100. In selecting 200 cm, the students have probably made fewer errors in their thinking and have selected the best estimate relative to the correct response. A relatively large threshold distance (1.1923) supports the selection of this distractor for the award of partial credit.

There was relatively large threshold distance in Item 23 in Year 7, 2013 and the distractor and category probability curves are shown for this item in Figures 7.7 and 7.8. The students were given a table (see Table 7.4) showing the relationship between the mass and volume of barley. They were asked to identify the mass of barley for a volume of 105 cubic metres. Choosing 65 tonnes rather than 63 indicated additive thinking as the student equated an increase of 5 cubic metres to an increase of 5 tonnes. Selecting either of the other two choices, 165 and 175, indicated thinking quite unrelated to the pattern in the numbers provided. The item was a link item in 2013 and the threshold distance was much higher for the Year 7 students than for Year 9 (1.9 cf. 0.4). This decrease in the threshold distance suggests that what is partial knowledge in Year 7 may not be partial knowledge to the same extent in Year 9, and as would be expected, the Year 9 students have developed a better understanding of this concept than the Year 7 students.
Table 7.4  Stimulus for Item 23: Year 7, 2013

<table>
<thead>
<tr>
<th>Volume (cubic metres)</th>
<th>Mass (tonnes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>100</td>
<td>60</td>
</tr>
<tr>
<td>150</td>
<td>90</td>
</tr>
<tr>
<td>200</td>
<td>120</td>
</tr>
</tbody>
</table>

One item in the Year 9, 2014 assessment involved the identification of the amount of water each hen would drink in a week if 10 hens drank 3 litres per day. The difference between the thresholds in this item was quite large (1.475) and the skills required to pass both thresholds involved applying two rates. The application of a single rate, determining the total water consumed in one week, was awarded partial credit and when this rate was further applied to determine the volume per hen, the student scored full marks. This type of partial credit belonged to the group described as incomplete solution. The other two options did not consider the difference in the units, days and weeks, which is a relatively easy concept for students to recognise and process and hence, not deserving of partial credit.

Partial knowledge was predicted according to content, but not expected given the shape of the distractor curves (Figure 7.13), for Item 38 in the Year 7, 2013 assessment. The students had to identify the bunch of balloons that gave a person the best chance of selecting a black balloon. Partial knowledge was confirmed when the thresholds were ordered, and it was classified as proportion ignored when the bunch with 3 black and 4 white balloons was chosen rather than the bunch with 2 black and 2 white balloons. Students were awarded partial credit for choosing the bunch with the highest number of black balloons rather than the highest proportion of black balloons. In the other two options, the bunches had fewer black than white balloons and these options would probably have been selected by students who are guessing or who misread the question.

The description of just these five items provides useful ideas for the construction of MC items where the award of partial credit can be justified: a focus of the first research question for this study. Further guidance can be obtained by studying the skills associated with partial knowledge in the items where partial credit scoring was justified. These are provided in Appendix 7.9. The items from the NAPLAN numeracy tests for 2013 and 2014, which have been used in this study, are currently available from the ACARA (2016b) website.
7.3 Merit of awarding partial credit for partial knowledge

The CR items in the ICCAMS tests of ratios and fractions, and the MC items in the NAPLAN numeracy assessments, were not written specifically to reward partial knowledge, nor to provide opportunities for students to demonstrate their partial knowledge. However, in this study it has been possible to show, regardless of the authors’ intentions, that both types of items have provided students with the opportunities to demonstrate their partial knowledge. Giving credit for this partial knowledge provides more statistical information about student learning and produces more precise estimates of item difficulty and person proficiency. Such improvement in precision can be achieved without adding to the test demands on students.

The processes which have been used in this study for scoring and analysing MC items, and in the application of the Rasch model to the students’ responses, can be used in tests containing MC items for a greater variety of age groups and subject areas. Studying the distractor curves has provided greater insight into the extent to which the existence of partial knowledge can be detected and confirmed: This process is relevant wherever there are MC items. The size of the threshold distance indicates the disparity between the skills described as full and partial conceptual knowledge. Further investigation of threshold distances could provide qualitative and statistical information to assist with the placement of skills on difficulty continua for other mathematical concepts and for concepts in other subject areas.

The identification of what constitutes partial knowledge of proportional reasoning has been determined from the item content and from the analysis of the results when partial credit was awarded. This information can be used in the creation and confirmation of learning trajectories and the identification of the stages in the development of mathematical concepts. This would be particularly useful for teachers for their planning of curriculum and learning activities. Teachers and researchers also benefit from knowing when awarding partial credit does not confirm the proposed partial knowledge.

The responses identified as partial knowledge in the CR items could be used to write informative distractors for MC items and to support the construction of other distractors. Further investigation of students’ responses to CR items from previous and current tests could add to the amount of information available to write such distractors. The results of this part of the study indicate that it should be possible to deliberately create MC items which assess students’ partial knowledge. The application of these findings could increase the appeal of the MC format to educators in many disciplines and especially to teachers of mathematics.
CHAPTER 8: Construction of Year 8 test with MC items

8.1 Overview of test creation

Attending to the design, creation, and analysis of MC items can improve the quality of the information about achievement that can be gathered from responses to these items. This claim expresses the intended outcome of answering the three research questions formulated for this study. To gather evidence to support the claim, an online assessment consisting solely of MC items was designed, created, and implemented. The MC items were purposefully written, not only to assess student skills and understandings in the field of proportional reasoning, but also to facilitate higher order thinking and to allow the students to score credit for their partial knowledge. The test was designed, and the responses analysed, to reduce the bias in the estimates of student proficiency due to guessing. By improving the way MC items function to collect information, some of the concerns with their use can be addressed.

For the Year 8 test, a framework was designed and used to determine the test structure and format. The determination of a suitable framework resulted in the identification of guidelines which were implemented during the creation of the items. With these guidelines it was possible to identify the various dimensions, for example, the nature of the content and the degree of difficulty, which the test needed to address. Findings from earlier research studies were used to identify the content, format, and organisation of the MC items created for the test. The MC items were created by adopting or adapting examples identified in earlier research studies, and where further items were needed, I created them. All items were checked for compliance with the item-writing guidelines described in Chapter 3 and were then subject to reviews by other educators and students. These reviews are described later in this chapter.

The Year 8 test, which was conducted in Western Australia in November 2016, consisted of 60 MC items, organised in blocks of 10 items, and each student was given three blocks. Given the proposed method of analysis, it was not necessary to have every student respond to every item. The choice of thirty items for each student was to allow the teachers and students to complete the assessment within one mathematics lesson which is typically between 40 and 60 minutes in length. With 60 items it was possible to cover a wide range of the curriculum and therefore to have a broad set of items from which the construction of distractors deserving partial credit could be explored and checked. The division of the 60 items into the different blocks provided the opportunity to create blocks of different difficulty levels.
Every student did the first block of 10 items which assessed the standard curriculum for the test participants. Then the students were given two more blocks of 10 items and these were either allocated to the student at random or chosen according to the student’s previous performance. This resulted in two designs for the test, hereafter referred to as the non-adaptive and the adaptive designs, or the non-adapted and the adapted versions of the test. The version of the test provided to the student was chosen at random. In the non-adaptive design, the other two blocks were allocated randomly and for the adaptive design, the allocation of the other two blocks was conditional. Details of these designs are given later in the chapter.

The difficulty estimates from the non-adaptive design are biased because of the presence of guessed responses and the purpose of conducting this adaptive testing regime, was to generate proficiency estimates where the opportunity for the students to guess was reduced. These estimates could then be compared with the estimates from the post-hoc analysis of data where the responses were tailored; that is, the responses from the students, who had been identified as likely to have guessed their answer, had been removed.

By having both formats it was possible to compare the effects of experimental and post-hoc control of guessing on the scale in terms of the psychometric properties: notably the statistical information. It was also possible to see the different effects of these two approaches to controlling for guessing when confirming the success of constructing distractors deserving of partial credit. To detect or confirm that a distractor deserves partial credit, the correct response to an item should be relatively difficult for the sample of the population being tested. If the item is very easy, then the students will answer correctly, and the partial credit option is unlikely to be selected. If the item is very difficult, then most students might be guessing randomly. Thus, the success of constructing distractors deserving of partial credit is linked to both item difficulty and the influence of guessing.

The tests were conducted with Year 8 students from twelve schools in Western Australia and of these schools, there were at least three from each of the educational sectors: The Department of Education, Catholic Education, and the Association of Independent Schools. While the test was provided in an online environment, one school requested, and was provided with, pen and paper copies. The total number of valid records available for analysis was 1273. The responses from the students to the MC items were analysed by applying Rasch measurement theory, and with the use of the RUMM2030 software (Andrich et al., 2016). The analysis and software are described in the previous chapter and the results in the next chapter.
8.2 The test framework

A test framework was designed to guide the construction of the test which was used to collect student responses to the MC items; the frameworks described in Chapter 2 were not applicable. While some of these frameworks had been created to design tests for students of the same age, none of them had been created specifically for tests consisting solely of MC items which tested partial knowledge, and which addressed content relating to proportional reasoning. Some aspects of these frameworks were suitable for the data collection and these were included in the final design, in which each item was classified in terms of the *nature of the content assessed, the type of proficiency addressed, and the level of difficulty of the item*.

A table listing these aspects for each item is given in Appendix 8.1.

Four of the five dimensions recommended by Kind (2013a), that is, knowledge, conceptual knowledge, behaviour, and context were incorporated into the design of the framework; only student attitude was excluded. These four dimensions were selected because they mirrored the expectations described in the curriculum documents. Knowledge and conceptual knowledge were combined to form one dimension, namely *content*, and this was classified according to the skills identified as being essential for the development of sound proportional reasoning (see Table 4.3). The test content covered multiplicative thinking (number), fractions, decimals, percentages, ratios, rates, spatial proportion, and linear relationships. The behaviour dimension was aligned with the mathematical *proficiencies* described in the Australian curriculum: understanding, fluency, problem solving, and reasoning (see Table 4.4). Familiar contexts were chosen for the students. These included situations involving simple finance, sporting activities, the school environment, social events, and home life.

*Difficulty of items*

To identify levels of difficulty for item classification, the assessment frameworks described in Chapter 2 were considered. Dividing the items into categories of knowing, applying, and reasoning, as per the TIMSS framework, and as adapted from the Bloom’s (Bloom et al., 1956) taxonomy, was recommended by authors of MC items (R. Peck, S. Zoumboulis, & P. Rogers; personal communication, March 7, 2016). Similar categories with more detailed descriptions of the skills at each level are used in the creation of the NAEP tests in the United States (National Assessment Governing Board, 2014). Those categories of complexity are referred to as low, medium, and high.
The final levels of difficulty were labelled as for the NAEP framework, that is, low, medium, and high because these categories appeared to be the easiest to identify and apply and hence, the most efficient to implement for the proposed test.

Items requiring students to recall knowledge, procedures, and properties are deemed to be of a low difficulty level, as are items requiring students to recognise concepts and procedures. In easy items students are not required to devise original methods or solutions. Items of medium difficulty may require students to go beyond what is familiar, routine, or practised and, to decide which calculation to use and how to execute it. Such items may alternately require students to use more than one step in determining an answer, to be more flexible in their thinking, and to connect two or more ideas. For items of high levels of difficulty, students may be required to reason, justify, generalise, think creatively, analyse assumptions and arguments, perform unfamiliar and complex calculations, and use more abstract reasoning.

It is necessary to consider what is high, medium, and low difficulty in terms of the expected learning for that year group. When a student is first taught a concept, a high level of cognitive processing is required to comprehend and synthesise the associated ideas. After further study, the student will develop greater understanding of the concept and the ability to apply it. With increasing exposure, practice, and application, the concept becomes part of the student’s repertoire of knowledge (Andrich, 2002). It follows that what constitutes high level reasoning for a Year 7 student may have become a familiar routine for a Year 8 student.

Item 24 (Block 3, number 4) was designed as an item of high difficulty for Year 8 students. The students should have known the terms discount, original, and price but the question was probably unfamiliar as it required them to work backwards from the discounted price and to recognise the required mathematical operation. Students are usually presented with examples where they are given the original price and then asked to determine the discounted price. Unless the student is familiar with this type of question (unexpected for Year 8 students) they would need to apply several concepts to identify the correct response. For this item, shown in Figure 8.1, the student needs to realise the answer must be greater than $35.20, and they can thus eliminate the first and last options. The student also needs to know that the percentage increase is not the same as the percentage decrease and thus eliminate the second option. This leaves the third option as the correct response, but it is not intuitively attractive because it involves division by a percentage which is less than one but does not look less than one. Had the discount been set at 10% and the final price at $45, the item may have been easier.
Item 24

After a pair of shoes is discounted by 12%, the price is $35.20. The original price before the discount was given can be calculated by

a. $35.20 \times 88$

b. $35.20 \times 112$

c. $35.20 \div 88$

d. $35.20 \div 112$

*Figure 8.1* An item designed to be a difficult item for a Year 8 student

For Item 4, located in Block 1, calculating distance using speed and time would be more familiar for Year 8 students than Item 24 because the given numbers are easy to manipulate. These numbers should trigger the recollection of different methods for identifying the correct response by those students who do not remember the formula to use. Knowing that the response to this item, as seen in Figure 8.2, must be greater than 100 km allows students to immediately eliminate the first two options. Seeing 90 minutes as one and a half hours and 50 km as half of 100 should be very easy for most Year 8 students and this would direct them to choose the third option as the correct response. Items classified as being at a low level of difficulty for this test are not necessarily so easy that most students get them correct consistently; they are created to be at the standard required for Year 8 and should be familiar for most students.

Item 4

Milly drives her delivery truck from the farm to the depot at an average of 100 km per hour. This journey takes 90 minutes. What is the distance from the farm to the depot?

a. 90 km

b. 100 km

c. 150 km

d. 175 km

*Figure 8.2* An item designed at a low level of difficulty for a Year 8 student
8.3 Item allocation to students

The 10 items in each block were written for a nominated year group and at a specified level of difficulty for the students in that year of schooling. By doing this the test could be adapted for students of different abilities. The difficulty rating of each block was described in terms of the standard for that year group; the standard being the level which most students of that year should be able to achieve (ACARA, 2015a). By having items suitable for different school years and at different levels of difficulty for these years of schooling, it was possible to provide a variety of pathways for the adaptive design. It also provided an opportunity to collect a much greater set of responses from students doing the non-adaptive version of the test, responses which produced more biased estimates, but ones which could be used for the comparison with the less biased estimates generated from the adaptive design.

The standard for a year group is considered as being at a C grade level. The first block, 8L, contains items set at the Year 8 standard and as such, they are classified as of low difficulty. These items require students to recognise and recall familiar knowledge and procedures which they should have learned in Year 8 but, being of low difficulty does not imply that they are easy. Similarly, Block 4, 7L, consists of items which have been written at the standard level for Year 7 students and which involve recognition and recall of knowledge and procedures named in the curriculum for Year 7 students. Items of medium or high difficulty occur in Blocks 2, 3, and 5 and they may require students to perform procedures which are unfamiliar and more complex as described in the previous section, but which are based on the curriculum for their year group. The design of the blocks is presented in Table 8.1 and this shows for each block, the number and name, the item numbers, the year for which the curriculum content is written, and the difficulty rating.

<table>
<thead>
<tr>
<th>Block number</th>
<th>Block name</th>
<th>Item numbers</th>
<th>Curriculum year</th>
<th>Difficulty rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8L</td>
<td>1–10</td>
<td>8</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>8M</td>
<td>11–20</td>
<td>8</td>
<td>medium</td>
</tr>
<tr>
<td>3</td>
<td>8H</td>
<td>21–30</td>
<td>8</td>
<td>high</td>
</tr>
<tr>
<td>4</td>
<td>7L</td>
<td>31–40</td>
<td>7</td>
<td>low</td>
</tr>
<tr>
<td>5</td>
<td>7MH</td>
<td>41–50</td>
<td>7</td>
<td>medium/high</td>
</tr>
<tr>
<td>6</td>
<td>6L</td>
<td>51–60</td>
<td>6</td>
<td>low</td>
</tr>
</tbody>
</table>
Non-adaptive design

Every student was directed to complete the items in Block 1 (8L) and then, except for the students doing a paper copy of the test, the students were directed to either a non-adaptive design or an adaptive design. For the non-adaptive design, students were allocated at random to one of the 10 different tests, each test consisting of Block 1, plus two more blocks of 10 items. For example, one student could have been given Blocks 1, 4, and 5, and another student might have received Blocks 1, 3, and 6.

For all of these ten tests, the order of block presentation to the students followed a numeric order, thus Block 1 was always the first block offered to the student. Furthermore, if Block 6 was allocated to a student, then it was always the last block which was offered. For the students doing the paper copies of the test \( n = 175 \), equal numbers of copies of the ten versions of the non-adapted test were printed and randomly distributed.

Adaptive design

Even though the test was designed for Year 8 students, items were created to address the skills and understandings that are taught in the earlier years: Years 6 and 7. Some items were written at the standard level for both years and compiled into Blocks 4 and 6 to provide alternate pathways for the students allocated to the adaptive version of the test. Block 5, 7MH was also written to accommodate students doing the adaptive version of the test. It was used for the pathway for students who were successful on their first block but not on their second.

For the students selected for the adaptive design, the process described by Chuesathuchon (2008) was applied in determining the next items to be offered to the student. Success with the first 10 items in Block 1 was used to determine the subsequent subtest for the student, as it was not practical nor necessary to use success on a single item to determine the next item to be presented. Success in a block was defined as 6 or more items correct out of the 10 offered.

The structure for block allocation is shown in Figure 8.3. Success in one block meant that the next block offered to the student was the one to the left and on the lower level. An unsuccessful student was offered the block to the right at the lower level. There were four possible combinations of blocks in the adaptive design: one of these is 8L, 8M, 8H; and another is 8L, 7L, 7MH. The first of these combinations represented the items to present to the most proficient students and the 8L, 7L, 6L combination would be presented to the least proficient students.
As the students completing the test were in Year 8, and the test was conducted at the end of the school year, Block 1 was designed to contain items which should have been familiar for the students and which were classified at the standard (low level of difficulty) for most Year 8 students. Students who were successful in this block in the adaptive design were then provided with items that are theoretically of a higher level of difficulty, but which are still based on the Year 8 curriculum. Unsuccessful students were offered items based on the Year 7 curriculum in the expectation that they would find these easier.

Figure 8.3 Pathway of blocks in adapted version of Year 8 test

Using an adaptive design should have reduced guessing by the students (van der Linden & Glas, 2010) as the more difficult items were only offered to the students who demonstrated an ability to respond correctly to the easier items.

One constraint in designing the allocation of blocks was to have at least 100 responses to every item in each of the non-adaptive and the adaptive versions of the tests. To allow for the possibility that there were not equal numbers of students directed to the four different pathways of the adaptive test, a more conservative number of 150 students was chosen as the minimum number of students to respond to each item in the adaptive version. With the ten different pathways for the non-adaptive test, the more conservative figure of 200 was chosen as the number of students. To achieve these more conservative numbers, 1600 students were needed, and 600 of these students were to be directed towards the adaptive test. This constraint of having sufficient numbers of students responding to the items was to provide enough data from which the estimates of item difficulty and person proficiency could be determined.
8.4 Creating the MC items

To create the MC items to test student understanding of mathematical concepts, and to fit the framework designed for this test, several factors were addressed. These included determining the content and context of the items, identifying further detail of the proficiencies to be addressed, and deciding the degree of difficulty. Guidelines for creating effective MC items were followed, and examples of items used in earlier research studies provided ideas and inspiration to support item creation. The construction of the distractors is an important part of item creation, and in this study the identification of what constitutes partial knowledge was necessary to assist in the writing of informative distractors.

The test content refers to the skills and concepts which are necessary for students to develop if they are to have sound understanding of proportional reasoning. These skills and concepts are described in Chapter 4 and are summarised in Appendix 4.1. Within each block there are items addressing the different areas of content, for example, fractions and decimals, and this varies according to the curriculum for that year group. As recommended by Cumming and Maxwell (1999), the contexts were chosen to be familiar for the students and to allow the focus to be on the mathematics rather than on the situation outlining the mathematics. Within each block there are items addressing the four mathematical behaviours or proficiencies as described in the curriculum for that year group. These mathematical behaviours and the degrees of difficulty addressed in the items have been described previously. The content, proficiencies, and degrees of difficulty are provided for each block and for each individual item in Appendix 8.1, and they are summarised in Table 8.2 for Block 1 only.

**Table 8.2 Block structure for Block 1: Year 8 at the low level of difficulty**

<table>
<thead>
<tr>
<th>Block</th>
<th>Item</th>
<th>Content</th>
<th>Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>multiplicative thinking</td>
<td>understanding</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>fractions</td>
<td>understanding</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>multiplicative thinking</td>
<td>problem solving</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>rates</td>
<td>fluency</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>linear relations</td>
<td>understanding</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>percentages</td>
<td>fluency</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>percentages</td>
<td>fluency</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>spatial reasoning</td>
<td>reasoning</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>percentages</td>
<td>understanding</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>ratios</td>
<td>fluency</td>
</tr>
</tbody>
</table>
The guidelines described by Haladyna et al. (2002) were applied when writing the items for this test. The main ideas are presented in the stem; positive language has been used in the items; and all options are similar in length and format, presented vertically, and are independent of each other. Using a question format for all item stems as recommended by Statman (1988) was not considered necessary in this study as the mathematical items were designed with shorter stems and options than those discussed in that study. MC items typically contain between three and five options but research findings for the ideal number of options is inconclusive. Four options were chosen for the MC items in this study as this is the number of options typically experienced by the students selected for the Year 8 test.

Most of the items I have written, and others have been created by either adopting or adapting items identified in earlier research studies. Many items suitable to adopt for the Year 8 test were not in MC format but the values, contexts, and calculations were appropriate for the creation of the MC items. Items 19 and 23 (Block 2, number 9; and Block 3, number 3) were reported in research by Boudreaux, Kanim, and Brahamia (2015), and Items 8 and 27 (Block 1, number 8; and Block 3, number 7) were adopted from research reported by Kuo, Chen, Yang, and Mok (2016). Other items which have been adopted with very little alteration are Items 2 and 48: Block 1, number 2 (Hart, 2004); and Block 5, number 8 (Burfitt, 2014). A further ten items were created by adapting MC or CR items from a variety of research studies (Jitendra et al., 2011; Lamon, 2013; Lesh & Landau, 1983; Misailidou & Williams, 2003; Nabors, 2003; Parish, 2010). For these ten items, ideas from research studies were used for the content, but the context, units, or values were adapted for the target audience. The first 10 items, Block 1, are given in Figure 8.4 and all 60 items are provided in Appendix 8.2.

Identifying suitable distractors

All distractors, that is, incorrect responses, have been created to make the choices as plausible as possible to prevent students from selecting the correct response because they immediately recognise options that are highly unlikely to be correct. One distractor in each item was purposefully written to attract students who have some knowledge and understanding of the item’s content, but insufficient insight for them to identify the correct answer. This particular distractor is theoretically more attractive than the other distractors to the students who do not recognise the correct response. For some items, this distractor is written to appear more like the correct response than the other distractors, and in other items, this distractor addresses concepts that a student with partial knowledge will understand.
Creating a distractor which attracts students with limited knowledge of the item also means that the wording of the stem must be checked to ensure that, for a student who believes that this distractor is in fact the correct response, the expression makes sense.

With four options in each MC item there were three distractors to create but the findings from the research studies did not provide enough guidance to enable three distractors, each indicative of partial knowledge, to be written for each item. It was also considered to be quite a challenging task to achieve for MC items assessing mathematical skills. Hence, just the one distractor was created to detect partial knowledge rather than a hierarchy of distractors to match different levels of partial knowledge. The prior recognition of the various types of partial knowledge of proportional reasoning as outlined in Chapter 5: additive thinking, proportion ignored, incomplete solution, and reasonable estimate, supported the creation of the informative distractors. The fifth type of partial knowledge, incorrect method, which was identified during the analysis of the MC items described in Chapter 7, has been used to classify several behaviours in the Year 8 test. These include the application of familiar algorithms when it is not appropriate, as in Item 1, (Figure 8.4) where the students knew that the factor of four was correct, but they selected an increase rather than a decrease.

Knowledge of common misconceptions has been used to write distractors which are more plausible than others and thus described as having information. For Item 40 (Block 4, number 10), a common misconception, that the more you buy the cheaper it is per unit of mass, has been incorporated in the item as the most plausible distractor. Some informative distractors were written to include some, but not all, of the aspects of the content needed to identify the correct response to the item. Distractors, where the students who do not have full knowledge can recognise some of these correct aspects, could be suitable for the award of partial credit. For Item 5 (Figure 8.4), students might see that $2.50 is being added as the number of kilograms increases in the table, the last option, but do not have the full knowledge to select the relationship between the two variables. The option they have chosen is only partly correct.

For each of the 60 items created for the Year 8 test, one distractor has been purposefully written to allow students to demonstrate their partial knowledge. The skills necessary for the selection of the informative distractor, as well as the skills necessary for the selection of the correct response are summarised for the first ten items in Table 8.3. These skills are described in Appendix 8.3 where the type of partial knowledge is identified as belonging to one of the five types listed earlier. The key and the informative distractor are identified in Appendix 8.2.
### Table 8.3 Partial knowledge skills in Block 1 (Items 1 to 10)

<table>
<thead>
<tr>
<th>Item</th>
<th>Skills in identification of key</th>
<th>Partial knowledge (PK) skills</th>
<th>Type PK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>increasing the number of people that share a set cost results in a smaller contribution needed per person (inverse proportion)</td>
<td>quadrupling is linked to increasing to four times the original amount</td>
<td>IM</td>
</tr>
<tr>
<td>2</td>
<td>one sixth of three quarters of a circle is one eighth of a whole circle</td>
<td>one sixth of three quarters of a circle is one sixth of a whole circle</td>
<td>PI</td>
</tr>
<tr>
<td>3</td>
<td>there is a total of 105 g of carbohydrate in 100 g of falafel and 200 g of bread when the rates are 25 g /100 g and 40 g /100 g respectively</td>
<td>components are added without consideration of the proportion</td>
<td>PI</td>
</tr>
<tr>
<td>4</td>
<td>distance travelled in 90 minutes at 100 km per hour is 150 km</td>
<td>distance is calculated using 100 km per hour as 1 km per minute.</td>
<td>RE</td>
</tr>
<tr>
<td>5</td>
<td>rule to determine cost (from a table of values) is <em>number of kilograms multiplied by the cost for 1 kg</em></td>
<td>identification of recursion from a table of values</td>
<td>IS</td>
</tr>
<tr>
<td>6</td>
<td>65 $\times$ 100 $\div$ 450 is not a calculation to determine 65% of $450$</td>
<td>identifies multiplication by 450 but does recognise 65 $\div$ 100 as 0.65</td>
<td>IS</td>
</tr>
<tr>
<td>7</td>
<td>increasing from 1.2 million to 1.6 million is about a 30% increase</td>
<td>percentage increase is calculated relative to the final amount.</td>
<td>RE</td>
</tr>
<tr>
<td>8</td>
<td>if $A/B = 1/3$ and $B= 60$, then $A = 20$</td>
<td>identification of 30 cm as one third of 60 (most reasonable estimate)</td>
<td>RE</td>
</tr>
<tr>
<td>9</td>
<td>$744$ is the result of increasing $600$ by $24%$</td>
<td>identifies $44$ as $44%$ increase</td>
<td>PI</td>
</tr>
<tr>
<td>10</td>
<td>$3:5 = x:8$ does not mean that $3/5 = x/8$</td>
<td>Interprets the context to form a correct ratio</td>
<td>IS</td>
</tr>
</tbody>
</table>

**KEY**

- **IM**: incorrect method
- **IS**: incomplete solution
- **PI**: proportion ignored
- **RE**: reasonable estimate
1. If the number of people sharing the cost of building a Cat Refuge were to quadruple (multiply by 4), then the amount of money that each person needs to give will
   a. reduce to a quarter of the original amount
   b. reduce to a half of the original amount
   c. increase to four times the original amount
   d. increase to double the original amount

2. These two circles have the same area. Three quarters of the first circle is pink. What fraction of the yellow area (whole second circle) is \( \frac{1}{6} \) (one sixth) of the pink area?
   a. \( \frac{1}{6} \)  
   b. \( \frac{1}{4} \)  
   c. \( \frac{1}{8} \)  
   d. \( \frac{1}{24} \)

3. The table shows the carbohydrate content of 100 g of bread and falafel.

<table>
<thead>
<tr>
<th>Food</th>
<th>Carbohydrate (g / 100 g of food)</th>
</tr>
</thead>
<tbody>
<tr>
<td>falafel</td>
<td>25</td>
</tr>
<tr>
<td>bread</td>
<td>40</td>
</tr>
</tbody>
</table>

If you eat 100 g of falafel and 200 g of bread, then you take in
   a. 65 g of carbohydrate  
   b. 75 g of carbohydrate  
   c. 95 g of carbohydrate  
   d. 105 g of carbohydrate

4. Milly drives her delivery truck from the farm to the depot at an average of 100 km per hour. This journey takes 90 minutes. What is the distance from the farm to the depot?
   a. 90 km  
   b. 100 km  
   c. 150 km  
   d. 175 km

5. The table shows the number of kilograms of oranges and their respective costs.

<table>
<thead>
<tr>
<th>number of kilograms</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>cost in dollars</td>
<td>$2.50</td>
<td>$5.00</td>
<td>$7.50</td>
<td>$10.00</td>
<td>$12.50</td>
</tr>
</tbody>
</table>

The relationship between the cost and the number of kilograms is
   a. cost = number of kilograms x $2.50  
   b. cost = number of kilograms ÷ $2.50  
   c. number of kilograms = cost + $2.50  
   d. add $2.50 to the cost
Figure 8.4 (cont’d)  First ten items of Year 8 test (Items 6 – 10)

6. Which of the following expressions does NOT calculate 65% of $450?
   a. \(65 \times 450 \div 100\)
   b. \(65 \times 100 \div 450\)
   c. \(0.65 \times 450\)
   d. \(65 \div 100 \times 450\)

7. Oil production was forecast to be 1.2 million barrels per day. Instead, it reached 1.6 million barrels per day. This increase in what was forecast is closest to
   a. 4%
   b. 25%
   c. 30%
   d. 40%

8. For the diagram provided
   \[
   \frac{\text{radius of small circle}}{\text{radius of large circle}} = \frac{1}{3}
   \]
   The radius of the large circle is 60 cm. What is the radius of the small circle?
   a. 10 cm  b. 20 cm  c. 30 cm  d. 40 cm

9. The correct answer in a student’s homework was $744. The question could have been
   a. Increase $600 by 24%
   b. Increase $700 by 44%
   c. Decrease $700 by $44
   d. Decrease $800 by $166

10. Read the following problem and study the solution shown.
    For every 3 kg of plastic recycled, there are 5 kg of cardboard recycled. What weight of recycled plastic would you expect for 8 kg of recycled cardboard?
    Four lines of working are shown.
    What is the first line on which an error is made?
    \[
    \frac{3}{5} = \frac{x}{8} \quad 1.
    \]
    \[
    \frac{3}{8} = \frac{x}{5} \quad 2.
    \]
    \[
    24 = 5x \quad 3.
    \]
    \[
    x = 5 \div 24 \quad 4.
    \]
    a. 1  b. 2  c. 3  d. 4
8.5 Item review

Following the creation of the MC items for the test, three reviews to provide feedback on the items were designed and implemented. The purpose of conducting these reviews was to identify any problems with the items in terms of the ease with which they would be understood by the students. The reviews allowed the language of the items to be checked for the use of appropriate terms and allowed the presentation of the items to be checked for clarity and the use of appropriate diagrams. Two reviews involved inspection of the items by colleagues with experience of teaching the content to students of the same age as those from whom the data were collected. Using a peer review process, as in the case of the review by colleagues, was found to improve the quality of multiple-choice items and to increase test reliability (Malau-Aduli & Zimitat, 2012). For the third review, cognitive interviews were conducted with students and, as described in Chapter 3, this involved conversations with students about their interpretation of the test items.

*Item review by educators*

In the first review, teachers were asked to examine 10 items and to comment on their suitability for the target population. Responses of yes or no were sought for the following aspects: (1) the item can be mapped to the curriculum for that level, (2) it is clear what the mathematical expectation of the item is, (3) the language used is appropriate for the item and the students, (4) the item is presented clearly for the students, and (5) the item is far too difficult for students in Year 8. Thirteen teachers participated in this process and each item was examined by five different teachers. The results as seen in Appendix 8.4 indicate that for all aspects and all items, teachers provided positive support for the items as they were written, with between 80% and 100% agreement on each question. Where teachers indicated the item was not mapped to the curriculum, the item was checked and its presence in the test justified.

Comments provided by the teachers were used to edit the items, and this included bolding key words, re-positioning diagrams, providing easier numbers to manipulate, correcting errors in grammar and spelling, and reorganising options. Altering content included shortening sentences, removing redundant words, and adding clarifications. One item for which teachers recommended changes is Item 55 (Block 6, number 5). Sentences were rewritten in active tense, the more familiar term cost replaced fee and the emphasis on the price at which to sell the cakes was established by repeating the word in both final sentences. In Figure 8.5 the original text of the item and the changes made are shown.
For the second review the focus was on partial knowledge and two teachers, experienced in writing and editing assessments, were asked to identify the correct response and any incorrect options which, if chosen as the answer, would indicate that the student had some knowledge of the item’s content. The outcomes from this review, provided in Appendix 8.5, indicate that there was some agreement between the researcher and the two reviewers regarding the distractors which could be considered for partial credit.

Given the search of the literature that I had completed, I did not expect that the two reviewers would have had the same insight into partial knowledge of proportional reasoning as myself, so a high level of agreement was not anticipated. However, the process was beneficial in that the reviewers not only confirmed the correct responses, but also identified with justification further distractors that I could consider for partial credit during the analyses. Furthermore, the reviewers made suggestions to improve the items, and the adoption of these recommendations involved minor changes to the items including bolding key words and making numbers in the items easier to manipulate.
Item review by students

Individual interviews were conducted with ten Year 9 students who had volunteered to support the research. The main purposes of conducting interviews was to identify any possible misinterpretations of the items, to see if the items were interpreted as intended by the researcher, and to check that the items were clear for the students. Year 9, rather than Year 8, students were chosen for these cognitive interviews because they had had greater opportunity to develop the skills tested in the survey and theoretically had a better understanding of the items’ content. At the time the interviews were conducted, which was a few months before the main data collection, the Year 8 students had not been taught all the curriculum content covered in the test. By interviewing Year 9 students, the Year 8 students at the same school would have still been able to volunteer for the data collection. The sample size was sufficient given that conducting the interviews was only one of the review processes for the items.

Each student was asked 22 questions relating to two of the six blocks of items. As the items appeared on the screen, one at a time, I asked the student a question and waited for a response before forwarding the screen to the next item. There was a different interview question for each item. The planned 15 interviews did not materialise because some of the students who had volunteered to be interviewed did not bring the signed parental permission slips necessary to satisfy ethics requirements. Consequently, some blocks were reviewed more than others, but all blocks were reviewed at least twice. I created the interview questions, which are indicated in Appendix 8.6, by adopting or adapting those used in studies by Campanelli (1997, p. 21), Collins (2003, p. 235), Karabenick et al. (2007, p. 143), and Wininger, Adkins, Inman, and Roberts (2014). My personal experience of teaching students of this age enabled me to assess the ease with which students would respond to questioning and this experience was invaluable for preparing the probe questions for the interviews.

The responses from the ten students interviewed (see Appendix 8.6) indicated that only minor revisions to the items were needed. All students were able to read and paraphrase items; they could describe the item content and they indicated that they could eliminate and identify incorrect options. All students also claimed that they knew the mathematical terms used and that they could determine the action required to select a correct response. Furthermore, for different items each time, at least 80% of the students said that they could (a) understand the wording, (b) describe what they thought the item required, (c) describe what could be done to make the item easier, (d) provide ideas mathematically related to the item, (e) provide a sensible explanation for choosing an option, and (f) explain the difference between two
options. This same proportion of students, again with different items each time, claimed that the item was an easy one to understand and said it was easy to identify the correct response.

Given the recommendations made by the students, all items were examined, and revisions were considered where students had indicated the need. Diagrams were added, sentences reworded, and language simplified where the meaning of the item was not clear. Some sentences were shortened, and some repositioned with a new sentence on each line and with greater spacing between them than previously. An example of an item seen by the students, and edited as a result of their input, is provided in Figure 8.6. For this item one of the students had suggested that adding a diagram would make the item easier to understand.

### ITEM 43 as seen by students during the interviews

Phil is packing up to move. He has two boxes which are both rectangular prisms. He estimates the larger box is twice as high, three times as long and twice as wide. Approximately how many times greater is the capacity of the larger box?

- a. 24
- b. 12
- c. 14
- d. 7

### ITEM 43 as edited in response to student’s suggestions

![Diagram of two boxes](image)

Phil has two boxes which are both rectangular prisms. The larger box is twice as high, three times as wide and twice as long as the smaller box. How many times greater is the capacity of the larger box?

- a. 24
- b. 12
- c. 14
- d. 7

*Figure 8.6* Item 43 before and after editing

As most students showed a good understanding of the items and indicated that the test was at the appropriate level for Year 8 students, only minor revisions were made, and the basic mathematical content was unaltered.
8.6 Data collection

The process

The Year 8 test of 60 MC items was written using the Qualtrics survey program (Qualtrics, 2013), which is software licensed to the university and, which has all the features needed for this testing process. The program facilitates the creation of MC items with diagrams, colour, and mathematical symbols. Logic statements can be accommodated, and it is possible to branch to alternative pathways based on the students’ responses to one or more items. I had used the software in my earlier study (Burfitt, 2014) and found it to be easy to use, robust, reliable, and capable of storing data for several hundred participants. With the online and telephone support provided by the software company, I was able to solve all problems associated with branching and programming. For the data collection, the information gathered from each student, including the MC options chosen, was recorded and stored online. It was then possible to download all the details to a spreadsheet in a format suitable for further analysis.

The participants

Schools in Western Australia with substantial numbers of students in Year 8 were invited to participate in the data collection. For schools under the jurisdiction of the Department of Education and Catholic Education, permission was sought from those authorities for me to issue an invitation to their schools to participate in the research. Schools were selected from those responding to the invitation with the aim of collecting data from approximately 2000 students with equal numbers of males and females. It was impractical to select the students at random from within the schools as this would have created much extra work for the teachers. Furthermore, this was deemed unnecessary, because the focus of the research was on the quality and analysis of the MC items, rather than on generating measures to represent the achievement of all Year 8 students in the skills and understandings relating to sound proportional reasoning.

There were several reasons for choosing Year 8 students to respond to the MC items testing the skills necessary for developing sound proportional reasoning. For one, students by the end of Year 8 should have developed a sound understanding of proportional reasoning and the associated skills; these are indicated in the national curriculum (D. Siemon, personal communication, July 9, 2015). In my earlier study, the participation rates for Year 8 students were much higher than for Year 9 students and this prompted me to consider Year 8 students
for this data collection (Burfitt, 2014). Another reason for choosing Year 8 students is that Year 8 is the only schooling year in lower secondary where there is no external testing: state or national. It was expected that this would enhance teacher and student willingness to be involved. A further consideration was that Year 8 students would have undertaken three NAPLAN numeracy assessments in their earlier years and should be familiar with the MC format for mathematical items.

About six months after the students were expected to sit the test, they would have had their first opportunity to demonstrate the basic numeracy skills necessary to qualify for the WACE certificate through the quality of their achievement in the Year 9 NAPLAN assessments. The Year 8 test provided in this study could have been used as another preparation for that opportunity. Doing a test of this nature at the end of Year 8 allowed feedback to be provided to the schools at the beginning of the following school year. Teachers could have used the information provided to identify the misconceptions that were common among their students and to revise these concepts before the Year 9 NAPLAN numeracy assessment. The prospect of providing students with an independent NAPLAN-like preparation, and the opportunity to receive diagnostic feedback could have encouraged teachers to volunteer to be involved in the study. For both teachers and students, there were many benefits of doing the test at the end of Year 8 and this feature of the study design was purposefully implemented to attract a substantial number of respondents, and thus to enable valid conclusions to be drawn.

**Ethics requirements**

Approval to conduct this research on behalf of The University of Western Australia was granted by Human Ethics from the university’s Office of Research Enterprise. The procedures for collecting information from students complied with the national requirements for ethical conduct as outlined in the National Statement (National Health and Medical Research Council, Australian Research Council, & Australian Vice-Chancellors’ Committee, 2015). Following the granting of the approval, those responsible for research at the Association of Independent Schools, the Department of Education, and Catholic Education gave consent for me to approach the principals of their schools to request their participation in the data collection. All letters of communication to the schools, teachers, parents, and students were approved by the university.

One school agreed to participate in the conduct of the cognitive interviews with the Year 9 students, and for that part of the study written permission from each parent and child was required. For the main data collection, the information requested from the students was typical
of student data entry when using online mathematical software, and in accordance with the ethics approval, students could give passive consent to participate in the research by simply doing the Year 8 test. All participants from the schools, that is, the principal, students, teachers, and administrators, volunteered to be involved in this research. Parents were provided with information about the nature of the research, the requirements of the test, and their child’s role in supporting the research. Without any penalty and at any time, parents could choose to opt out if they did not wish their child to participate. Similarly, the students could opt out.

Information for participants

Twelve schools participated in the data collection for the Year 8 test of MC items and there were at least three schools from each of the three educational sectors. The students who expressed an interest in supporting the research were provided with a Participation Information Form, which is given in Appendix 8.7. This form explains that the aim of the research is to improve the way that multiple-choice items are used to collect data about student learning in mathematics. Participants were told that the test was to be done without the aid of calculators and that it consisted of only MC items for which they would be asked to select the best option.

There was no reference to research into guessing because this could have influenced the students’ selections of options; they may have thought there was a penalty for guessing and subsequently left answers blank. This action was deliberate and was informed by the research which was described earlier, and which reported a change in student behaviour when the students were told about the presence or absence of penalties for guessing. There was also no mention of crediting partial knowledge because it is preferable to simulate, as far as possible, the conditions under which the participants sit large-scale assessments involving MC items.

The link to a practice test was provided to all participating schools for distribution to teachers and students. The practice test was created to allow the participants to develop familiarity with the types of items in the test and with the process for responding to items in an online environment. The Qualtrics (2013) software was used to create and conduct the practice test which consisted of 10 MC items. Over 850 students accessed the practice test but data on the number of these students who also sat the Year 8 test were not collected.
Data collection

As well as selecting options for multiple-choice items in the data collection for the Year 8 test, participants were asked to nominate their school to allow feedback on achievement and on misconceptions to be provided to individual schools. This feedback was one of the benefits provided to the school for their participation, and it consisted of a report on the school’s performance on the test, an overview of the items in the test, and a summary of the misconceptions identified during the data collection. For the school to make use of this feedback, the reports were sent to each school early in the following school year and within three months of the data collection. In the test, students were also asked to enter their gender as there is evidence that guessing is more prevalent for males (Ben-Shakhar and Sinai, 1991) and it was a factor for consideration in the analysis. Students’ names were not collected as they were not necessary for the feedback or the analysis. No other information was collected from the participants.

The link to the test was provided to schools in November 2016, and except for the school using a paper version of the test, all test-related activity was electronic; no school visits were necessary. The recommended time for the test was forty minutes and this was considered ample time to complete 30 MC items and it allowed the test to be completed during a single lesson. November was chosen for the data collection as most schools would have covered the curriculum for Year 8 and the test material would have been familiar for most students. Further to this, there are fewer interruptions to the school day; fewer carnivals, visiting speakers, and school excursions during the last term of the school year. Schools could choose the days and times for the test but were told that the students would be unable to come back to complete the test on another day. Teachers were informed that any Year 8 student was welcome to attempt the test, that there was no need to balance out the number of participants for gender or perceived ability, and that the students should be supervised as for any normal mathematics test.

Except for one school, the test was conducted online through the Qualtrics survey platform and participants could enter their responses using mobile phones, iPads, or computers. Test instructions were provided online, and the multiple-choice items were presented one at a time to the students. A new item only appeared on the screen when the student indicated that they had finished the previous one. Students could review and change their responses as the opportunity to go back to the previous question was provided on each screen page. After the
30 items had been presented as an online survey, the students were given a further opportunity to have their responses excluded from the analysis.

Once the schools were given the link to the test, I checked the data each day and, where there was evidence that some students had completed the test, I downloaded and stored the data securely. The survey was only opened during school hours to reduce the opportunities for students to complete the test more than once; however, in my previous study (Burfitt, 2014) there was no indication that the data had been corrupted by students accessing the test outside school hours. The process for collecting data was efficient and did not appear to make unnecessary demands on either teachers or students.
8.7 Test participants

From the online data collection 1273 records were available for the analysis and this was considered a satisfactory number on which to base any conclusions. Each record was allocated an identifying number and contained, a school name, the student’s gender, and responses to 30 of the 60 items. Knowing the identity of the school allowed feedback to be provided and it gave some insight into the nature of the sample of the Year 8 population. The school name was not linked to the students’ responses for any of the analyses which were conducted on these data and which are reported in this study. Twelve schools participated in the data collection with at least three from each of the education sectors. The percentage of the students who were males was approximately 43%.

The proportion of students directed to the adapted version of the Year 8 test was 37.5%, but with one school choosing to only do the pen-and-paper version of the test, the final proportion of students responding to this version was slightly lower, with 412 students who provided responses for the adapted version and 861 for the non-adapted version. The number of participants completing each block for the two test designs is given in Table 8.4, and the number of participants completing each version of the two test formats is given in Table 8.5. For all blocks in both versions of the test, except for the most difficult block in the non-adapted version, there were at least 100 responses for each of the 60 items, a number considered sufficient for the planned analysis. There were only 43 students who provided responses in the adapted version to the 10 most difficult items.

<table>
<thead>
<tr>
<th>Block</th>
<th>Numbers of student participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Name</td>
</tr>
<tr>
<td>1</td>
<td>8L</td>
</tr>
<tr>
<td>2</td>
<td>8M</td>
</tr>
<tr>
<td>3</td>
<td>8H</td>
</tr>
<tr>
<td>4</td>
<td>7L</td>
</tr>
<tr>
<td>5</td>
<td>7MH</td>
</tr>
<tr>
<td>6</td>
<td>6L</td>
</tr>
</tbody>
</table>
### Table 8.5  Numbers of participants for different test designs

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Numbers of student participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers</td>
<td>Names</td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>8L, 8M, 8H</td>
</tr>
<tr>
<td>1, 2, 5</td>
<td>8L, 8M, 7MH</td>
</tr>
<tr>
<td>1, 4, 5</td>
<td>8L, 7L, 7MH</td>
</tr>
<tr>
<td>1, 4, 6</td>
<td>8L, 7L, 6L</td>
</tr>
<tr>
<td>1, 2, 4</td>
<td>8L, 8M, 7L</td>
</tr>
<tr>
<td>1, 2, 6</td>
<td>8L, 8M, 6L</td>
</tr>
<tr>
<td>1, 3, 4</td>
<td>8L, 8H, 7L</td>
</tr>
<tr>
<td>1, 3, 5</td>
<td>8L, 8H, 7MH</td>
</tr>
<tr>
<td>1, 3, 6</td>
<td>8L, 8H, 6L</td>
</tr>
<tr>
<td>1, 5, 6</td>
<td>8L, 7MH, 6L</td>
</tr>
</tbody>
</table>

Of the 12 schools involved in the Year 8 test, 11 were co-educational with approximately equal numbers of males and females. The other school was an all-girls school where there were only 36 Year 8 students who provided data for analysis. For all schools, the proportion of participants who were females responding in the non-adaptive design was 58%, and in the adaptive design was 56%. The imbalance of males and females suggests females are more willing to participate in this type of study and the proposal could be confirmed by further research in this area.

The unexpectedly greater proportion of females should not affect the interpretation of the results, as the research focus is on the quality of the information collected from responses to MC items, rather than describing populations for this proficiency. There are still considerable numbers of both males and females contributing data for analysis and these are believed to be sufficient for comparing the performances of the two gender groups. Because there was no attempt to obtain a representative sample of Year 8 students, the proficiency distribution may not reflect the achievement of Year 8 students in this state or country. However, given the nature of the test and the numbers of students providing their responses, it is possible to use the results to describe a potential hierarchy of learning for the relevant skills and understandings associated with the development of proportional reasoning. This hierarchy of learning could be a useful tool for teachers and for further research.
8.8 Preparing responses for analysis

If responses to MC items include correct responses which have been guessed, then the scores for individuals will be higher than they would be otherwise, and the items will appear easier. To maximise the accuracy of the estimates generated for item difficulty and person proficiency, the bias due to guessing was removed from the students’ responses. In this study, guessed responses were removed in the process described as tailoring and this involved treating responses, likely to have been guessed, as if the items had not been presented to the student. The analyses conducted during tailoring, and the justification for conducting these analyses are further described in Chapter 6.

8.8.1 Tailoring processes

In their report on the assessment of guessing, Andrich et al. (2012) chose a cut-off of 0.3 for tailoring and all the responses of persons, where the probability of being correct on the item was less than 0.3, were converted to missing data. For my test of 60 MC items, with each item having four options, the probability of selecting a correct response by random guessing is 0.25. Before deciding the criterion for the cut-off value for all tailored analyses, the effects of using 0.2, 0.25 and 0.3 were compared for both the non-adaptive and the adaptive designs. The students responding to items in the non-adaptive design were expected to guess more than those responding in the adaptive design, and by comparing the effect of tailoring on both designs, it was possible to provide evidence to support this claim. All tailoring to identify a suitable value for the cut-off was performed on the responses scored without partial credit and the changes to the estimates of item difficulty were examined.

Relative item difficulties from the original and tailored analyses cannot be compared directly because the mean item difficulty is constrained to be zero in each analysis and, any increase in difficulty for some items is offset by decreases in difficulty for other items. It is still possible, however, to compare the extent of the changes to the item difficulties which occur when different cut-off values are tested. The process described by Andrich et al. (2012, p. 425) was used in this study to test the effect on the item estimates of changing the cut-off value for the tailoring. A standardised statistic was calculated for each item by dividing the difference in the item difficulty estimates between the non-tailored and tailored analyses, by the square root of the difference of the squares of their standard errors. Where the standardised statistic was greater than 2.56 or less than -2.56, the item estimates were considered significantly different at the 1% level.
8.8.2 Tailoring criteria

In the non-adaptive design, the numbers of items significantly affected by tailoring was similar when the cut-off values were 0.25 and 0.3 and this number exceeded those affected when the cut-off value was 0.2. In the adaptive design, the number of items affected by tailoring was similar for all three cut-off values. The actual numbers of items affected are shown in Table 8.6, where it shows there are now 59 instead of 60 items. The responses to Item 11 in the non-adaptive design were excluded from the analysis because the item was deemed to be an extreme item when tailored at cut-off values of 0.25 and 0.3 and further analysis was not possible.

The similarity of the numbers of items significantly affected, when the criteria of 0.25 and 0.3 were applied to the responses for both designs, indicated that using 0.25 as the cut-off value was effective in converting most of the responses likely due to guessing, to missing data. However, 0.3 was selected as a conservative value for all tailoring in the analyses of the responses collected from the Year 8 test, to ensure that a considerable number of the responses, which were likely to have been guessed, were removed from the data. Using a value of 0.3, did nonetheless, require the removal of Item 11 from all analyses.

<table>
<thead>
<tr>
<th>Cut-off values</th>
<th>Non-adaptive design</th>
<th>Adaptive design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Significant $(p &lt; 0.01)$</td>
<td>Not significant</td>
</tr>
<tr>
<td>Comparison (one tailored)</td>
<td>0.2</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>39</td>
</tr>
</tbody>
</table>

It is worth noting that the number of items which were significantly affected by tailoring in the adaptive design for the test is much smaller than the number of items significantly affected by tailoring of responses in the non-adaptive design. Furthermore, this occurred for both cut-off values. This supports the claim that the use of the adaptive design has contributed to a reduction in the propensity of the students to guess their responses and hence to a reduction in the bias of the estimates, due to guessing, which have been determined.
8.9 The Year 8 test

The Year 8 test has been constructed in order to collect responses from students for the purpose of addressing the research questions which were proposed for this study. The findings from the various research studies have been used to guide the design and creation of a test framework which was applied during the process of developing the online test. The test consisted of items which address the content and proficiencies of the Australian curriculum: Mathematics as is stipulated for Year 8 students as they develop in the skills associated with proportional reasoning. In applying the framework, the MC items were created by adopting and adapting items and ideas from earlier research studies.

For each MC item, one of the distractors, known as an informative distractor, was purposefully constructed in order to allow the student to demonstrate their partial knowledge of the item’s content, and hence to deserve partial credit. In total there were four options for each item; the correct response, the informative distractor and two other distractors which were written to be plausible but not informative. This design for the item construction was to facilitate the discovery of a solution to the first research question concerning the construction of MC items with informative distractors. By having items with informative distractors, the effect on the proficiency scale of the application of partial credit scoring could be investigated: the focus of the second research question.

The items were organised into six blocks of 10 items and all students were only offered three blocks; 30 items in total. All students, regardless of which blocks they were offered, answered the same 10 items at the beginning of the test and these common, link items allowed all items to be calibrated concurrently and all students to be located on the same scale. There were two designs for the test; one a non-adaptive design with the random allocation of blocks of varying difficulties, and the other an adaptive design. In the adaptive design, students answered the second and third blocks of items depending on their success or failure on the previous block. The purpose of the adaptive design was to provide alternate pathways for students of varying abilities, to reduce guessing and to examine the effect on these actions on the measures of person proficiency: the focus of the third research question.

The design and construction of the Year 8 test has been for the purpose of generating results to support finding solutions to the three research questions which were formulated previously.
CHAPTER 9: RESULTS: Scoring MC items with partial credit

Introduction

As described in the previous chapter, and in summary, to confirm that the function of MC items could be improved by reducing guessing, and by designing distractors deserving partial credit, an online test consisting of 60 MC items was conducted. There were two designs for the test; a non-adaptive design where blocks of 10 items were allocated at random, and an adaptive design where blocks were allocated according to success on a previous block. It will be recalled that the adaptive design was implemented in order to minimise the effect of guessing. However, because of some violation of local independence, this design is known to produce biased estimates of item difficulty that tend to be more stretched at the extremes of the item difficulty (Eggen & Verhelst, 2011). The non-adaptive design does not have this disadvantage of a violation of local independence. However, because of guessing, the estimates in the Rasch model tend to be biased in a way that the more difficult the item, the more the difficulty estimates are regressed towards the mean. Thus, the bias in item difficulty estimates from the two designs tend to be in opposite directions. Conveniently, as described by Andrich et al. (2012), it is possible to control for guessing in the non-adaptive design by conducting post-hoc analyses where the responses are tailored and the responses where guessing is likely are removed.

The two designs therefore, the non-adaptive tailored and the adaptive designs, provide the opportunity to assess the validity of the inferences from each: first in removing the effect of guessing, second in assessing the success of constructing distractors which permit partial credit. To the degree that the two designs give similar results, to that degree they confirm the validity of each. The evidence presented regarding the benefits of having items with partial credit includes the improved definition of the scale which has two thresholds for each item deserving partial credit, and the implications this definition of the scale has on improving the estimates, and the precision of the estimates, of person proficiency.

A considerable number of students, over 1000, provided their responses to the test and with this number of responses it was possible to conduct analyses to investigate the prevalence of guessing and to determine the success of constructing distractors for partial credit scoring. As a result of this study it has also been possible to identify the skills associated with the partial knowledge of concepts relating to proportional reasoning.
The results of the study reported in this chapter contrast to the studies of the impact of guessing and the identification of the skills associated with the development of partial knowledge using data from the ICCAMS and the NAPLAN tests. First, the inclusion of an adaptive design permitted an empirical, rather than a post-hoc control of guessing. Second, one distractor was constructed to provide an opportunity for the student to demonstrate partial knowledge of proportional reasoning, rather than a post-hoc examination of responses to detect the possibility of the student having partial knowledge of the concept.

The rest of this chapter is structured as follows. The next two sections provide details of the analysis of items before partial credit was considered. Section 9.1 describes how these details provide the background to the success of the adaptive design and it includes a discussion of the relative difficulties of the blocks of items. In Section 9.2 the impact of the post hoc control of guessing in the non-adaptive design is described and this guessing control is compared with the control that is evident in the adaptive design.

Section 9.3 presents the items which are confirmed to have a distractor that deserves partial credit. Seventeen items were confirmed to justify partial credit, with one item sufficiently extreme that it was excluded from the partial analyses. Section 9.4 provides the substantive interpretation of the distractor for each item that is shown to deserve partial credit from the Rasch model analysis. Section 9.5 presents the relative difficulties of the 42 items which did not have a distractor with partial credit. The difficulties of these items were taken to define the scale. Section 9.6 provides the analyses of the items which were identified for the award of partial credit, as well as a description of the process of placing the threshold estimates of these items onto the same scale as those of the dichotomously scored items defined in Section 9.5. Section 9.7 presents the impact of awarding partial credit on the estimates and precision of person proficiency in comparison to when partial credit is not awarded.

Section 9.8 considers three subsidiary features of assessment with MC items that could be investigated from the results of this thesis. First, the impact of controlling for guessing on comparisons of performances between males and females. Second, an examination of gender DIF and response dependence, two important requirements of an empirical scale and properties of the Rasch model, are presented. Both essential criteria, invariance across gender and lack of response dependence, are confirmed. The third feature relates to a learning trajectory for the development of proportional reasoning. Using the results of this study, it is possible to identify a hierarchy of learning and a probable order for learning some skills associated with sound proportional reasoning.
9.1 MC items before tailoring or partial credit scoring

Identification of responses for analysis

All 60 items were initially scored dichotomously with one mark for the correct response and zero otherwise. Because the estimates for item difficulty and person proficiency are known to be biased in both the non-adaptive and adaptive designs, the responses of the students doing the non-adaptive design of the test were analysed separately from those doing the adaptive design. The results of the analyses of these two designs were compared. Both sets of responses were also combined and analysed as one single set of responses. For completeness, the results of this analysis were compared with those from the individual sets of responses, and the differences in item difficulty and person proficiency were noted. Because the blocks of items had been constructed to reflect varying levels of difficulty, the mean estimates of item difficulty for the blocks were considered.

9.1.1 Results of comparing estimates from both test designs

With the bias known to exist in the estimates, not all the statistics generated from the three analyses provided the best possible estimates, but they provided information indicating where further investigation was warranted. Even though the estimates have not been anchored to a fixed origin and a direct comparison is not valid, the results can still be compared. This is justified because all 60 items are the same in each analysis and the item difficulties add up to zero giving the same origin; any differences can be attributed to the designs.

The statistics reported in Table 9.1 show that the person separation index is high in all three analyses and this indicated that the items were well aligned with person proficiency, and that the estimates of person proficiency were well spread across the continuum. The mean estimate of proficiency for males exceeded that of females in all three analyses, and the difference was significant, to at least the 5% level, in each analysis.

Some of the statistics, which were generated during the separate analyses of the three sets of responses, differ considerably between the non-adaptive and the adaptive design. For the adaptive design, the maximum value of the information function is lower, and its location is lower on the scale. The mean proficiency estimates for all persons, and for males and females separately, are much lower in the adaptive design than in the non-adaptive design. Further to this, the standard deviation is notably higher for both person proficiency and item difficulty in the adaptive design.
Table 9.1  Dichotomous scoring statistics for three test situations

<table>
<thead>
<tr>
<th></th>
<th>Non-adaptive</th>
<th>Adaptive</th>
<th>Both non-adaptive and adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Item estimates:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.89</td>
<td>1.47</td>
<td>0.94</td>
</tr>
<tr>
<td><strong>Information function:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum value</td>
<td>12.8</td>
<td>10.8</td>
<td>12.6</td>
</tr>
<tr>
<td>Location</td>
<td>0</td>
<td>-0.4</td>
<td>0</td>
</tr>
<tr>
<td><strong>Person estimates:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (standard deviation)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- All persons</td>
<td>-0.31 (0.94)</td>
<td>-0.59 (1.28)</td>
<td>-0.36 (0.96)</td>
</tr>
<tr>
<td>- Male</td>
<td>-0.23 (0.98)</td>
<td>-0.42 (1.39)</td>
<td>-0.26 (1.01)</td>
</tr>
<tr>
<td>- Female</td>
<td>-0.38 (0.89)</td>
<td>-0.74 (1.16)</td>
<td>-0.43 (0.91)</td>
</tr>
<tr>
<td>Person/Item separation index</td>
<td>0.781</td>
<td>0.884</td>
<td>0.794</td>
</tr>
</tbody>
</table>

The differences in the person proficiency and standard deviation for both item and person estimates indicate that the measures of item difficulty vary in the two designs. Given that the items are the same in both designs of the test, it would not be fair to allow the estimates of item difficulty to vary for students according to which version of the test they were offered. Such variation in these item difficulty estimates would affect the estimates of the proficiencies and hence the positions of the students on the proficiency scale. This provides further evidence of the need to remove the bias in the difficulty estimates of the two designs before using these estimates to finalise a single scale of proficiency for all persons.

9.1.2 Relative difficulty of blocks

For both designs of the test, the mean item difficulties of each block of 10 items were compared, in order to check if the empirical order of block difficulty matched the theoretical order of difficulty. This comparison of the empirical and theoretical ordering of block difficulty has been made using estimates determined from responses which were scored dichotomously, and which have not been tailored to minimise the effects of guessing and there has been no anchoring to achieve a common origin. The mean item difficulties for each block of 10 items are shown in Table 9.2 where the blocks are ordered according to the theoretical order of difficulty specified in the test framework. These results suggest that the attempt to create blocks of a hierarchy of difficulty has been somewhat successful and that it has been possible to assess some higher-order thinking using MC items.
Table 9.2  Mean item difficulty: Blocks in both designs

<table>
<thead>
<tr>
<th>Test designs</th>
<th>Block 6</th>
<th>Block 4</th>
<th>Block 5</th>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6L</td>
<td>7L</td>
<td>7MH</td>
<td>8L</td>
<td>8M</td>
<td>8H</td>
</tr>
<tr>
<td>Non-adaptive design</td>
<td>-0.38</td>
<td>-0.52</td>
<td>0.28</td>
<td>-0.50</td>
<td>0.65</td>
<td>0.48</td>
</tr>
<tr>
<td>Adaptive design</td>
<td>-1.10</td>
<td>-1.13</td>
<td>0.29</td>
<td>-0.74</td>
<td>1.03</td>
<td>1.66</td>
</tr>
</tbody>
</table>

Table 9.2 shows that the mean item estimates for the six blocks extend over a much greater range in the adaptive design of the test than in the non-adaptive design: 2.8 logits compared to 1.4. In the easier blocks, Blocks 6L and 7L, the mean item estimates for the blocks are much lower for the adaptive design than for the non-adaptive design and in the more difficult blocks, Blocks 8M and 8H, the mean estimates are much higher in the adaptive design than in the non-adaptive design. This suggests that there are factors other than the content of the item, for example, the presence or absence of guessing, which are influencing item difficulty.

There are some similarities in the relative difficulties of the blocks in both designs, but the order of difficulty differs. For both designs, the mean estimates for all three blocks which were constructed at the low level of difficulty (i.e. the standard) are lower than the estimates for the three blocks set at medium and high levels of difficulty. However, the range of difficulty did not follow the order of school years. For these low difficulty items, the highest mean was for the Year 6 content in the non-adaptive design and for Year 8 content in the adaptive design. The mean item estimates, as seen in Table 9.2, follow the constructed difficulty order better for the adaptive design than for the non-adaptive design.

For both designs the mean estimate of Block 5, which tests Year 7 mathematical content at a medium to high level of difficulty, is much greater than that of Block 1, the standard for Year 8. According to these results, the Year 8 students have found the Year 7 concepts, which were tested at the medium and higher levels of difficulty, to be much harder than the Year 8 concepts tested at the standard level. Two possible explanations for these results are offered. First, contrary to expectation, the students in Year 8 have not developed familiar routines for the high-level concepts for the Year 7 curriculum content. Second, the level of difficulty requires students to use higher order skills, for example, interpreting diagrams and graphs, and using a two-step solution process. This presumes that such higher order skills determine the difficulty rather than the content itself.
9.2 Guessing in both test designs

Guessing is expected whenever students are offered MC items for which they do not know the correct answer. In the non-adaptive design for the Year 8 test, the allocation of the second and third blocks at random has resulted in difficult items being presented to the students and hence the possible need to guess. By comparison, there should be less guessing in the responses to the adaptive design of the test as a smaller proportion of the students are faced with items that are too difficult for them. Thus, the expectation is that there is less guessing in the adaptive design than in the non-adaptive design.

9.2.1 Guessing in the non-adaptive design for the test

Impact on item difficulty estimates

The graph given in Figure 9.1 shows the effect of tailoring the responses in the non-adaptive design and this clearly indicates that there was notable guessing in the responses of the students doing the non-adapted version of the test. In this figure, the estimates of the tailored responses in the non-adaptive design are plotted against the estimates of the untailored responses for the non-adaptive design.

![Graph showing comparison of item difficulty estimates](image)

**Figure 9.1** Comparison of item difficulty estimates from responses for the non-adaptive design with estimates from the same responses which have been tailored to remove guessing

\[
\text{SD (tailored)} = 1.19 \\
\text{SD (non-adaptive)} = 0.82 \\
y = 0.66x - 0.44
\]
To generate the item difficulties for this comparison, the original responses of the students were tailored by removing the responses to each item of the students who were likely to have guessed on that item. The criterion chosen for tailoring the responses was to use a cut-off value of 0.3 whereby all responses where the students were deemed to have less that a 0.3 chance of being correct on the item were converted to missing data. With tailoring, Item 11 was identified as an extremely difficult item and had to be removed before the analysis could continue. The tailoring process is described in further detail in the previous chapter.

The tailored estimates used on the horizontal axis of the graph in Figure 9.1 were calculated when the responses were tailored using the cut-off value provided. The origin for the estimates for the non-adaptive design was equated to the same origin as the tailored estimates to allow the comparison, and thus, the estimates for the non-adapted design are described as being origin-equated. The equating involved the selection of the 10 easiest items from the calibration of the tailored estimates; the assumption being that there is no guessing in these easy items and that their difficulty estimates are unaffected by tailoring (Andrich et al., 2012). These 10 item estimates were then used as anchors for the same items in the calibration of the estimates from the non-adaptive design. For the ten items chosen for this anchoring, as seen in Figure 9.1, their difficulty estimates are unchanged, and the assumption of no guessing holds.

As expected, tailoring resulted in a greater range of item difficulty and increased item difficulty estimates. With the removal of guessing, the range of difficulty changed from [-2.23, 1.29] to [-2.27, 2.74], the greatest increase being at the higher difficulty level. Figure 9.1 shows that the increase in item difficulty is notable when the origin-equated estimates for the non-adaptive design exceed -1 and that the increase in the tailored estimates continues to increase as the estimates for the non-adaptive design increase. This evidence supports the assumption that there is a higher proportion of guessing in the more difficult items and that as the item becomes more difficult, guessing increases.

The standard deviations of the estimates indicate that there is greater variation in the values of the tailored estimates than there is in the values of the origin-equated estimates of the non-adaptive design. The correlation between the two sets of estimates is high at 0.956 and the equation for the line of regression of the tailored estimates on the estimates for the non-adaptive design is \( y = 1.34x + 0.6 \). This shows that, on average, for every logit rise in the difficulty estimates of the non-adaptive design, tailoring increases the difficulty by 0.34 logits. After being tailored, most items became significantly more difficult; 37 items at the 1% level of significance and 5 items at the 5% level of significance.
Impact on block difficulty

The removal of guessed responses in the non-adaptive design has resulted in much higher estimates of item difficulty for the more difficult items and little change for the very easy items. The increase in the difficulty estimates as a result of tailoring the responses was due to the removal of responses where students, who were identified as unlikely to be correct on the item, had provided a correct response and this unlikely selection was attributed to guessing.

Evidence of the greater increase in difficulty for the more difficult items as a result of tailoring was seen in Figure 9.1 and further evidence is given in Table 9.3. This table shows some of the changes to the blocks of items for the non-adaptive design when the responses are tailored. The increases in mean item difficulty were much higher in the blocks where more challenging items had been constructed: increases of 0.54 to 0.80 in Blocks 8M, 7MH and 8H. In comparison, the increases in mean item difficulty for all three blocks constructed at a standard level of difficulty were less than 0.3.

Table 9.3  Mean item difficulty: Blocks for tailored responses

<table>
<thead>
<tr>
<th>Item features</th>
<th>Blocks in order of theoretical difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Block 6</td>
</tr>
<tr>
<td>Mean item difficulty</td>
<td>-0.49</td>
</tr>
<tr>
<td>Increase in difficulty</td>
<td>0.29</td>
</tr>
</tbody>
</table>

The relative increases in item difficulty means for the blocks support the claim that it is in the blocks, which are designed to be more difficult, where the greater proportion of guessing is occurring. Blocks of items which were relatively easy for most participants were less affected by guessing. The reason the blocks were easier is that the items in these blocks contain content material which should have been quite familiar to the students. Table 9.3 shows that, in the tailored analysis, the difficulty estimates for items in the more difficult blocks increased by a greater amount than the estimates in the easier blocks.

While these results provide some evidence of the relative difficulty of the blocks and the amount of guessing occurring in them, they need to be treated with some caution. The means are based on 10 items, which should be reliable, but the existence of extreme item difficulties has not been considered. Furthermore, these estimates of item difficulty are based on dichotomous scoring which was later shown to be less than ideal for some of these items.
Impact on person proficiency estimates

The impact of guessing on person proficiency can be seen by comparing the histograms shown in Figures 9.2 and 9.3. Each histogram displays the distribution of persons according to their proficiency, and also the distribution of items according to their difficulty. Changes to the item estimates were described in the previous section but further detail of the effects on these estimates of removing guessing can be seen by comparing the item threshold frequencies at the base of each histogram.

Figure 9.2   Distribution of person proficiency: responses not tailored

Figure 9.3   Distribution of person proficiency: responses tailored
When guessing is included in the students’ responses, as pictured in the histogram shown in Figure 9.2, student proficiency ranges from -3 to 4 and when guessing is removed, as shown in Figure 9.3, the range increases by at least 2 logits. Many have lost the benefit of guessing and their measures of achievement are lower. There are many more students with much lower proficiency estimates after guessing is removed. There are also more students with higher proficiency estimates after guessing is removed because these more capable students have received the benefit of being successful on the difficult items which have higher difficulty estimates after guessing is removed. Removing the responses of some students for the determination of proficiency would not be an acceptable practice to adopt in high-stakes, national and state testing, and the examination of the results of the alternative way to reduce guessing is the focus of the next section.

9.2.2 Using an adaptive design to reduce guessing

It was expected that the implementation of an adaptive design for the test would reduce the amount of guessing by the students and this expectation has been confirmed by the results of this analysis. The relationship between the tailored estimates from the non-adaptive design and the estimates produced from the adaptive design (not tailored) is seen in Figure 9.4.

\[
SD \text{ (tailored)} = 1.19 \\
SD \text{ (adaptive)} = 1.40 \\
y = 1.05x - 0.002
\]

*Figure 9.4* Comparison of item difficulty estimates from responses for the adaptive design with estimates from the non-adaptive design which have been tailored to remove guessing.
It can be seen in Figure 9.4 that there is greater variation in the values of the estimates from the adaptive design than for the tailored estimates. However, the equation for the regression of the estimates for the adaptive design on the tailored estimates, \( y = 1.05x - 0.002 \), and the high correlation coefficient (\( r = 0.89 \)) indicate an overall similarity in the two sets of estimates.

Figures 9.1 and 9.4 together show that the estimates of the more difficult items in the adaptive design are more closely related to the tailored estimates than are the estimates from the non-adaptive design. This confirms that the implementation of the adaptive design has reduced the amount of guessing by the students.
9.3 Potential for partial credit

All 60 MC items in the Year 8 test were written to capture partial knowledge but, before examining the effect on the proficiency scale of granting credit for this partial knowledge, it was necessary to confirm the success of scoring this partial knowledge in each item. To decide which items should be scored with partial credit, three item features were considered: the ordering of thresholds when items are scored polytomously, the distractor curves produced in the analysis with dichotomous scoring, and the content of the item. The results of studying threshold ordering and distractor curves to confirm the items for which partial credit scoring was appropriate is described in the remainder of this section, Section 9.3. In the subsequent section, Section 9.4, the list of items selected for partial credit scoring is provided, and the choice of items is justified. As a result of confirming which items allowed the students to demonstrate their partial knowledge, it was possible to also confirm the theoretical partial knowledge proposed for the different item content. A summary of the partial knowledge skills, and the extent to which they were demonstrated in the students’ responses, is also provided in Section 9.4.

9.3.1 Ordered thresholds

Key evidence of the success of awarding credit for partial knowledge is the presence of ordered thresholds which indicates that the greater the proficiency of a student, the more likely they are to receive a higher score. To identify the items with ordered thresholds, all 60 items were scored for partial credit according to the theoretical partial knowledge created for each item. Polytomous scoring was applied with a score of two for a correct response, zero for an incorrect response, and one for the selection of the option designated as indicating partial knowledge. This process was conducted separately for each of the three sets of responses: those for the non-adaptive design, those for the adaptive design, and then the combination of responses to both designs. The process was also conducted with the same three sets of responses after tailoring when the responses of the students, for whom the probability of being correct on the item was 0.3 or less, had been converted to missing data.

When partial credit was allocated to all items there were 25 items with ordered thresholds in the non-adaptive design and 29 items with ordered thresholds in the adaptive design. Compared to the non-adaptive design, seven different items had ordered thresholds in the adaptive design and three items, which had been ordered in the non-adaptive design, were disordered in the adaptive design.
When all responses were analysed together, threshold ordering was very similar to the ordering shown for the non-adaptive design of the test. For this combined set of responses, the number of items with ordered thresholds was the same but in just two items, Items 52 and 58, the threshold ordering was the reverse of what it had been in the non-adaptive design. Of the 60 items, there were 22 items which had ordered thresholds regardless of which set of responses was analysed (i.e., common to all three sets) and a further 10 items which had ordered thresholds for only one or two sets of responses. These results are shown in Table 9.4.

Within each of the three sets of responses, the non-adaptive, the adaptive, and the two designs together, the threshold ordering did not vary when the responses were tailored. The same items had ordered thresholds in each set of responses, tailored or otherwise.

Table 9.4   Items with ordered thresholds

<table>
<thead>
<tr>
<th>Items with ordered thresholds in both tailored and non-tailored responses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
</tr>
<tr>
<td>Block 1 (8L):</td>
</tr>
<tr>
<td>Block 2 (8M):</td>
</tr>
<tr>
<td>Block 3 (8H):</td>
</tr>
<tr>
<td>Block 4 (7L):</td>
</tr>
<tr>
<td>Block 5 (7MH):</td>
</tr>
<tr>
<td>Block 6 (6L):</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

| **B**  | Items with ordered thresholds for only one or two sets of responses |
|------------------------------------------------|
| Non-adaptive design | Adaptive design | Non-adaptive and adaptive designs considered together |
| Items 45, 52, 59 | Items 10, 23, 28, 29, 47, 58, 60 | Items 45, 58, 59 |
|             | 10 items in total |

During the analysis of the responses from only the adaptive design and when all items had been scored with partial credit, the data could only be analysed after Item 21 was removed. Item 21 was a very easy item, but it was declared to be extreme when partial credit was awarded for the responses in the adaptive. A close examination of the raw data showed that only one student doing the adaptive version of the test provided an incorrect response for Item 21 and that student did not select the distractor designed for partial credit.
The items with ordered thresholds varied in item difficulty but they were mostly items that had been created to be more difficult. Of the 32 items with ordered thresholds, 16 items were positioned in Blocks 2 and 3, which had been constructed to be the most difficult blocks. The thresholds of the nine easiest items were disordered, while those of the 12 most difficult items were ordered. This classification of easy and difficult items was based on the values of the difficulty estimates for the dichotomously scored responses in the non-adaptive design.

The application of a post-hoc process to remove the responses for students who most likely guessed their answers, that is, tailoring, has not affected the relationship between ordered thresholds and the relative difficulty of the items. The lack of opportunity to identify partial credit in the easiest items is indicative of the expected mastery of the easier skills and concepts. For the more difficult items, where conceptual knowledge and understanding are less likely to be fully developed, the increased facility to demonstrate partial knowledge is expected. The results of these analyses have provided evidence of this proposal.

**Threshold distances**

For the 25 items which had ordered thresholds for responses in the non-adaptive design, the number of items in different ranges of threshold distances is given in Table 9.5. The threshold distances for each item were the same for both the tailored and the non-tailored responses.

**Table 9.5  Number of items with ordered thresholds**

<table>
<thead>
<tr>
<th>Range of threshold distance</th>
<th>&gt; 1.1</th>
<th>0.69 – 1.1</th>
<th>0.41 – 0.69</th>
<th>&lt; 0.41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of items</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

A description of the relevance of the threshold distance, which is the difference between the values at each of the two thresholds, was provided in Chapter 7. In brief, the greater the distance between the two thresholds, the higher the probability that a student will score 1 rather than 0 when they have a 50:50 chance of scoring 1 or 2. Where the threshold distance is 1.1 logits or more, a person with a 50:50 chance of scoring 1 or 2, has a probability of at least 75% of getting a score of 1 rather than 0. When the threshold difference is about 0.69, the probability of scoring 1 rather than 0 is 67%.
The ranking of the items according to their threshold distances was much the same for both the non-adaptive and the adaptive designs for the test. The similarity of these rankings gave general support for the selection of the items for partial credit. The small variation in the order may be due to the reduced number of students responding to a few of the more difficult items in the adapted version of the test. It was the size of the threshold distance, rather than the ranking which influenced the confirmation of items for partial credit, and how this distance was used is explained in the following section.

9.3.2 Distractor curves

To confirm the list of items for partial credit scoring, the distractor curves and the category probability curves (CPC) for all 60 items were examined. The distractor curves were available after the items were scored dichotomously and the CPC curves were available after the items were scored polytomously, that is, with partial credit. There were two advantages of studying these curves. First, it was possible to identify any distractors that had been incorrectly identified for partial credit and second, there was the opportunity to verify the process of using distractors curves for the confirmation of partial knowledge, as had been done with the NAPLAN data.

Identifying different distractors

For just two of the 60 items, Items 30 and 44, there were indications that the distractors which should have been awarded partial credit were not those constructed to test partial knowledge. This can be seen in the distractor curves and the CPC provided in Figures 9.5, 9.6 and 9.7.

Figure 9.5  Distractor curves for Item 30
In Item 30, partial credit had been given for the selection of Option 4, which was an example of additive thinking for inverse proportion. However, it was the green curve for Option 3, as seen in Figure 9.5, which typified the behaviour of an informative distractor. In this item, Option 3 was an example of the *incorrect method* of partial knowledge. The students who selected this option had possibly considered the situation as being one of direct proportion and applied the scale correctly for the incorrect type of proportion.

The red curve for Option 2 in Item 44, as shown in Figure 9.6, indicates the attraction of this distractor. However, it was Option 4 which had been chosen for partial credit. It seems that the less able students discounted the possibility of a person’s height being 110% of what it was before, and they selected the option describing the new proportion as less than 100%.

To confirm that partial credit was justified for these new options, rather than for the options which had been created for partial credit scoring, the two items were rescored, and partial credit was given for selection of these new options. For both items, the thresholds, which had previously been disordered, were then ordered. This change in the ordering is best seen in the changes to the CPCs and for Item 30 these CPCs have been provided in Figure 9.7. The CPC on the left shows that the thresholds were disordered when partial credit was given for the selection of Option 4, whereas the CPC on the right shows that the thresholds were ordered when partial credit scoring was given for the selection of Option 3.

*Figure 9.6  Distractor curves for Item 44*
Verifying the use of distractor curves

The inspection of the distractor curves for all 60 items gave similar results to those obtained when the MC items in the NAPLAN test were examined. The same two behaviours of the curves were used to confirm the distractors with potential for partial credit. First, the curve for one of the distractors is positioned notably above the curves for the other distractors for part or all of the range of person proficiency, and second, the curve for the correct response is below that expected according to the model. If there was graphical evidence of one of these two behaviours, then the item was classified in terms of the strength of this evidence. Diagrams showing distractor curves which indicate different levels of potential for partial credit scoring were provided in Chapter 7; namely, Figure 7.7.

Using these principles, all items were given one of the following classifications:

- A: strong indication of partial credit
- B: medium indication of partial credit
- C: possible indication of partial credit
- D: no indication of partial credit

For each of the four classifications, the number and proportion of items with ordered thresholds, and hence successful partial credit scoring, is provided in Table 9.6. All 32 items which had been identified as having ordered thresholds in either design for the test are included in this table. As with the NAPLAN data, the behaviour of the distractor curves was a good indicator of the potential for successful partial credit scoring. These results verify that the distractor curves for the options which were designated by the researcher to warrant...
partial credit, can be used for the reliable prediction of the presence of partial knowledge. If the thresholds are ordered, the substantive partial knowledge in the distractor is confirmed.

**Table 9.6 Identifying partial knowledge from distractor curves**

<table>
<thead>
<tr>
<th>Distractor curves: Indications of partial knowledge</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: strong indication</td>
<td>Total number of items</td>
</tr>
<tr>
<td>B: medium indication</td>
<td>6</td>
</tr>
<tr>
<td>C: possible indication</td>
<td>11</td>
</tr>
<tr>
<td>D: no indication</td>
<td>21</td>
</tr>
</tbody>
</table>

9.4 Items for partial credit

9.4.1 Finalising the items

Of the 60 items written for the Year 8 test, 18 items were selected for partial credit scoring. The size of the threshold distance, the extent to which the thresholds were ordered, and the nature of the distractor curves were the features used to determine which items had provided opportunities for students to demonstrate their partial knowledge. Given that the items were initially constructed to reward partial knowledge, the content of the items was included for consideration when the evidence from the statistical measures was inconclusive.

All items with threshold distances greater than 0.4 logits were accepted for partial credit as this is the range of values for which a person with a 50% probability of obtaining a score of 2 rather than 1, has at least a 60% chance of scoring a 1 rather than a 0. There were 17 items in this category, 15 of which had ordered thresholds for each set of responses, the non-adaptive, the adaptive, and the two designs together. None of these 17 items had been classified as lacking an informative distractor based on the behaviour of the distractor curves.

Item 30 was added to this list because rescoring the partial credit option, as described earlier, resulted in a threshold distance exceeding 0.4 logits. This was not the same for Item 44 which had also been rescored after the examination of the distractor curves; there the thresholds were ordered but the threshold distance was considerably less than 0.4 logits. Item 44 was therefore retained as a dichotomous item.
All eight items with positive threshold distances less than 0.4 logits were rejected for partial credit. The smallest threshold distance for any of these items was 0.15 logits and this translates to only a 53% probability of scoring 1 rather than 0 while having a 50% probability of scoring 2 rather than 0. The distractor curves for four of these eight items had been classified as *strongly indicating* the presence of an informative distractor but these were not the four items with the highest threshold distances; nor were the items’ threshold distances near the 0.4 logit cut-off value used in making the decision to reject items for partial credit.

A further inspection of the distractor curves for the rejected items showed that one item had possibly two informative distractors, and that two items had higher than expected proportions of students who provided a correct response at the higher proficiency levels. There was evidence of guessing in one item, and in another item two of the distractors were not at all attractive to most of the students. While in the previous analysis of the NAPLAN MC items, all items with ordered thresholds were accepted for partial credit scoring, a more conservative approach has been adopted for the Year 8 test and this has provided an opportunity to be confident in any conclusions drawn from the results of awarding partial credit in MC items.

**Students gaining partial credit**

The number of students who would have gained partial credit from the polytomous scoring of this selection of items is substantial. The proportion of the total number of students ($n = 1273$) selecting the distractor scored with partial credit varies between 29% and 74% for individual items, and the values for each item are given in Figure 9.4. For each item, more than a quarter of the students have chosen the option designed to allow students to demonstrate their partial knowledge. This is greater than the proportion that would result if the students were choosing options at random. It indicates some rationale for the students’ choices and supports the claim that such selections are purposeful and indicative of partial understanding of the item content.

Awareness of the extent to which students are demonstrating their partial knowledge can have significant benefits for the students and their teachers. Crediting this partial knowledge not only adds to the accuracy of the measurement of student achievement but it provides the student with a deserved recognition of their understanding: a recognition that would be well received by the student. By studying the skills associated with the partial knowledge concepts, and by being aware of the extent of the partial knowledge skills of their students, teachers can adjust their teaching activities to address any misconceptions or misunderstandings upon which this partial knowledge is based.
9.4.2 Features of items

A summary of the skills for complete and partial knowledge is provided in Figure 9.8.

<table>
<thead>
<tr>
<th>Item</th>
<th>Skill for full credit</th>
<th>Skill for partial credit</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Recognise inverse proportion with decreasing by a factor of four</td>
<td>Uses a factor of four</td>
<td>40%</td>
</tr>
<tr>
<td>7</td>
<td>Increasing from 1.2 million to 1.6 million is about a 30% increase</td>
<td>Percentage increase is calculated relative to the final amount</td>
<td>36%</td>
</tr>
<tr>
<td>11*</td>
<td>Percentage of components unchanged when amount of product changes</td>
<td>Sharing 72% between 3 people results in 24% per person</td>
<td>74%</td>
</tr>
<tr>
<td>12</td>
<td>Recognise that adding 25% changes the proportion</td>
<td>Recognises 75% as the complement of 25%</td>
<td>52%</td>
</tr>
<tr>
<td>14</td>
<td>Highest cost per kilogram is represented by the steepest set of points</td>
<td>The highest cost per kilogram is represented by the highest point</td>
<td>30%</td>
</tr>
<tr>
<td>15</td>
<td>Doubling the length of the side of a square increases the area fourfold</td>
<td>Doubling the length of the side of a square doubles the area</td>
<td>70%</td>
</tr>
<tr>
<td>16</td>
<td>Speed is total distance travelled divided by total time taken</td>
<td>Calculates the average of two speeds</td>
<td>54%</td>
</tr>
<tr>
<td>17</td>
<td>Calculates overall percentage using total of parts out of possible total</td>
<td>Calculates the average of two percentages</td>
<td>60%</td>
</tr>
<tr>
<td>20</td>
<td>Uses common multiple for A:C given A:B = 1:6 and C:B = 2:9</td>
<td>Compares two ratios in terms of the first members of each pair</td>
<td>38%</td>
</tr>
<tr>
<td>26</td>
<td>Multiplies percentages to determine successive percentage increases</td>
<td>Adds percentages to determine successive percentage increases</td>
<td>49%</td>
</tr>
<tr>
<td>30</td>
<td>Recognises inverse proportion with a factor of 3</td>
<td>Applies the given factor of 3 to direct proportion</td>
<td>39%</td>
</tr>
<tr>
<td>38</td>
<td>Sharing treats according to weights means that 10:15 = 12:18</td>
<td>Uses additive (add on) thinking when applying a ratio in context</td>
<td>39%</td>
</tr>
<tr>
<td>39</td>
<td>Applies a factor of 9 to the ratio 3:1 using multiplication</td>
<td>Identifies the incorrect factor to the ratio 3:1</td>
<td>29%</td>
</tr>
<tr>
<td>41</td>
<td>2 m is 0.4 of 5 m</td>
<td>Recognises 3 m is 0.5 of 6 m</td>
<td>38%</td>
</tr>
<tr>
<td>45</td>
<td>Considers original values before adding time elapsed to each term</td>
<td>Adds the increase over time to each term of the ratio</td>
<td>32%</td>
</tr>
<tr>
<td>46</td>
<td>Identifies whole amount needed for fraction comparison</td>
<td>Knows comparison of fractions requires common denominators</td>
<td>46%</td>
</tr>
<tr>
<td>51</td>
<td>Considers size as well as number of partitions of a model</td>
<td>Counts number of partitions in a fraction model</td>
<td>45%</td>
</tr>
<tr>
<td>59</td>
<td>If one hour is four fifths, then one fifth is 15 minutes</td>
<td>Calculates one fifth of one hour</td>
<td>38%</td>
</tr>
</tbody>
</table>

*Figure 9.8 Skills associated with the correct response and with the selection of the distractor for partial credit. The proportion of students selecting the partial credit distractor is given.

*Item 11 was removed from further analysis because it was deemed an extreme item when the responses were tailored.
The partial knowledge which the students have demonstrated in this test consists of a variety of skills associated with the development of proportional reasoning. There does not appear to be any pattern regarding the types of partial knowledge demonstrated, nor are there any particular skills associated with the demonstrated partial knowledge. For the following discussion of the concepts which indicated the potential for measuring partial knowledge, all items with ordered thresholds are included. The findings from this aspect of the study can be applied to the future construction of MC items and can be used by teachers for the planning and sequencing of learning activities.

There was no discernible link between the type of partial knowledge in the item and the success of polytomous scoring and hence the award of partial credit. The proportion of items of each type, for which awarding partial credit was successful, varied from 30% to 67%. The smallest proportion occurred for the type of partial knowledge designated as incomplete solution and the largest proportion was for partial knowledge identified as additive thinking. Given the inconsistent views about additive thinking behaviour as a stage in the development of proportional reasoning, which were described earlier, this is an interesting finding which suggests further research is worthwhile.

While the summary in Figure 9.8 provides details of the skills required to score credit in the partial knowledge items, details for all items are given in Appendix 8.3. All items, their correct responses, the potential partial knowledge skills, and the type of partial knowledge is provided in this appendix: The items, where scoring partial knowledge was successful, are highlighted in the table. It is not practical, nor the focus of this study, to give a detailed description of all the skills associated with partial knowledge of proportional reasoning as demonstrated by the Year 8 students in the test; one example from each group is provided.

Using an incorrect method could be demonstrated in Item 1, where the students interpreted the item as an example of direct rather than inverse proportion. The students demonstrated their partial knowledge by recognising the correct scale factor but not the correct process to calculate the answer. They chose an increase rather than a reduction.

The response which was classified as an incomplete solution in Item 46 was selected by 46% of the students. For this item, the students had to compare the values of two fractional amounts of money, but they were not given the total amounts from which to calculate the proportions. When teaching fractions, students are usually told that to compare the size of fractions they need to create equivalent fractions with common denominators to facilitate the
comparison. The option describing the need to create equivalent fractions represents some of the information which might be useful but is not necessary to respond correctly to this item. Selection of this option is also associated with the failure to recognise the need to know the total amount before any final amounts can be calculated.

Item 7 was constructed to assess the type of partial knowledge described as a reasonable estimate. Students were asked to identify the percentage increase when daily oil production rose from 1.2 million to 1.6 million barrels. Of the incorrect options 4%, 25%, and 40%, the closest value to the correct answer is 25% and this attracted a score of one. The other two options were unlikely to represent the percentage increase given the daily increase was 0.4 million barrels but were used as distractors with the idea that the students may have linked the 0.4 million increase with 4% or 40%. The selection of 25% could have also indicated that the students were calculating the proportion in relation to the 1.6 million barrels rather than the 1.2 million barrels.

In Item 16, students needed to recognise that the time taken should be considered when calculating an overall mean speed from two given average speeds. Failure to do so meant that the students who selected the option where the two given speeds were averaged were ignoring the proportion. Of the total number of students, 56% chose this option.

The remaining type of partial knowledge, described earlier as additive thinking, could be demonstrated in Item 38 where 39% of all students chose the distractor with additive thinking imbedded. The students were told that there were two dogs who were given treats according to the ratio of their weights. The students were asked to recognise the number of treats to give to the dog which weighed 15 kg when the dog weighing 10 kg received 12 treats. The selection of 17 for the number of treats, indicated that the students added the difference in the weights of the two dogs to the 12 to determine the number of treats for the larger dog.

Further identification of the skills that can be defined as partial knowledge of proportional reasoning can be achieved by a qualitative examination of the content of the items and of their informative distractors. There is extensive information available in this study for such an examination, but it would be impractical to include a complete description and such a detailed inclusion is beyond the scope of this research.
9.5 Estimates for dichotomous items

To identify a common proficiency scale on which to place all students, the parameters to describe all items were determined. This was necessary because there were two designs for the test, and for each of these designs, the parameters which described item difficulty were different and were biased. This bias has been discussed previously and it is summarised here to explain how the item parameters for the scale were identified. For the common scale, the parameters were determined first for the dichotomous items and then later for the items scored with partial credit.

9.5.1 Comparing estimates from both designs

There were two sets of difficulty estimates produced during the analyses of the responses to the dichotomous items; those from the non-adaptive design and those from the adaptive design. The original estimates from the non-adaptive design were biased because of the effects of guessing and in particular the estimates for the more difficult items were much lower than they should have been. The bias was reduced by tailoring the responses for the non-adaptive design; that is, by removing data for students who had been identified as probably having guessed their response to the item. As a result of tailoring, there were two sets of estimates from the non-adaptive design to consider for the dichotomous items on the final scale: the original estimates and the tailored estimates. To compare these two sets of estimates, they were each put on the same origin by the process known as anchoring; this process uses the very easy items that had no guessing. There were 10 such items and the anchoring occurred on the mean of these items in the tailored analysis.

From the adaptive design, there was a third set of estimates that could be used to determine the item parameters for the common scale. Even though the bias due to guessing had been reduced by the adoption of this design, there was still bias in these estimates as explained previously. The estimates for the dichotomous items in this adaptive design were placed on a scale with the same origin as the original and tailored estimates from the non-adaptive design. This was achieved by anchoring the origin of the estimates for the adaptive design to the mean of the same 10 items that were used to anchor the original estimates from the non-adaptive design. As a result of the anchoring, all three sets of estimates were origin-equated and could be compared for any differences.
The reduction in the proportion of guessed responses with the implementation of the adaptive design can be seen clearly in Figure 9.9, where the item difficulty estimates from both the non-adaptive and the adaptive designs for the test are placed on the same graph. The two sets of estimates are the same as those in Figures 9.1 and 9.4.

![Comparing estimates from both test designs](image)

**Figure 9.9** Comparison of item difficulty estimates from responses for the non-adaptive and adaptive designs of the Year 8 test with estimates corrected for guessing

The positions of the points on the graph in Figure 9.9 show that, for the more challenging items, the item difficulty estimates generated in the adaptive version of the test are much higher than those estimates for the same items when determined for the non-adaptive design without tailoring. For the students, the item difficulty does not change between versions of the test and it is necessary to determine which set of estimates should be used to generate the student proficiency estimates in order to place all students on the same scale, regardless of which version of the test was attempted.

### 9.5.2 Confirming item estimates

The item difficulty estimates, generated by tailoring the responses from the non-adaptive design of the test, are taken to be the best measures of difficulty to use for the dichotomous items in the determination of the estimates of person proficiency. The justification for this choice is given below.
The graph in Figure 9.9 confirms the results of Andrich et al. (2012) concerning the extent of guessing in MC items and the influence that this has on item difficulty estimates. For the final scale to determine measures of person proficiency, the choice of estimates is between those from the adaptive design and those of the tailored estimates from the non-adaptive design. To decide which of these estimates should be used several factors were considered. The first factor related to the relative fit of the items and some fit statistics are provided in Table 9.7. For completeness, the statistics for all three sets of estimates are provided.

**Table 9.7 Comparing fit for three sets of estimates**

<table>
<thead>
<tr>
<th>Set of estimates</th>
<th>Non-adaptive Tailored</th>
<th>Adaptive (non-tailored)</th>
<th>Non-adaptive Non-tailored</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Person Separation Index</strong></td>
<td>0.78</td>
<td>0.88</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>Item difficulty</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0</td>
<td>-0.0025</td>
<td>-0.44</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.20</td>
<td>1.41</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>Item fit residual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.39</td>
<td>-0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.67</td>
<td>1.59</td>
<td>2.25</td>
</tr>
<tr>
<td><strong>Number of items</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over-discriminating items</td>
<td>0</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Under-discriminating items</td>
<td>7</td>
<td>6</td>
<td>11</td>
</tr>
</tbody>
</table>

The PSI (person separation index) was quite high; thus, the power to detect misfit was good. As expected, there was greater variation in the estimates from the adaptive design than in the estimates from the non-adaptive design. There was not much difference in the mean or standard deviation of the fit residuals, which were described in Chapter 6 as indicators of item fit and to be used in conjunction with other fit statistics. Table 9.7 shows that the numbers of over-discriminating and under-discriminating items were similar for both sets of estimates. The criteria to identify item discrimination was the value of the fit residual: where it was less than -2.5, the item was described as over-discriminating, while under-discriminating items had fit residuals exceeding 2.5. An examination of these fit statistics, as summarised in Table 9.7, did not provide any clear justification for using either set of estimates.
The tailored estimates had a substantial effect of guessing removed and were not affected by response dependence. Therefore, it seemed that these tailored estimates were the most suitable to use as the item difficulty estimates of the dichotomous items for the final scale. Furthermore, these tailored estimates were calculated from the responses of many more students than had responded to items with the adaptive design. To most of the items for the adaptive design, more than 100 students provided a response, but there were fewer than 100 students who were offered Block 3 in this adaptive design. Furthermore, only 58 students doing the adaptive design were offered the pathway with Blocks 1, 4, and 5.

9.6 Estimates for partial credit items

Once the difficulty estimates for the dichotomous items had been fixed, it was necessary to identify the best values for the parameters to describe the difficulty of the items to be scored with partial credit. This was done to provide a single scale on which all students doing the test could be located: those doing the adaptive design and those doing the non-adaptive design. For each polytomous item with ordered thresholds, there are two values which define the item. The first value is the difficulty estimate at the first threshold, and this occurs where the probability of obtaining a score of 0 is equal to the probability of obtaining a score of 1. The second value is located where the probability of obtaining a score of 1 is equal to the probability of obtaining a score of 2. These values are described as threshold estimates.

Eighteen items were chosen for partial credit scoring, but Item 11 had been identified as an extreme item during the tailoring process and was deleted from all further analyses. During the investigation to identify suitable threshold estimates for the partial credit items, two other items showed disordered thresholds for responses to items in the adaptive design and both items were rescored dichotomously. As a result, there were 15 items with partial credit scoring and 44 items with dichotomous scoring.

9.6.1 Considering the estimates

Three sets of threshold estimates were investigated as potential sources of information to guide the selection of the parameters for the partial credit items. These estimates had been generated from the original responses to the items in the non-adaptive design and the adaptive designs and from the tailored responses in the non-adaptive design. After comparing these estimates and observing the effects of using these estimates on the scale of difficulty, the estimates from the tailored responses of the non-adaptive design appeared to provide the best values to use for the generation of the threshold estimates for the partial credit items.
Anchoring the estimates

To identify the threshold estimates for the partial credit items, the estimates for the dichotomous items chosen earlier were anchored. Two forms of anchoring were investigated: mean item anchoring and individual item anchoring. For mean item anchoring, the mean estimate of the 10 easiest items was fixed and for individual item anchoring, each of the estimates for the 10 easiest items were fixed. When mean item anchoring is used, the origins of the two sets of estimates are equated and when individual items anchoring is used, not only are the origins equated but also, the unit is fixed for the two sets of estimates. The effects of these two methods of anchoring were also considered in the investigation to determine the best set of estimates of the partial credit items for the final scale.

Investigating possible threshold estimates

The focus for this part of the investigation was to compare the threshold estimates of the partial credit items from the tailored responses in the non-adaptive design with the threshold estimates of the non-tailored responses from the adaptive design. For both designs the estimates of the dichotomous items are fixed, and the estimates for the partial credit items have been determined in the two ways described previously. The bias due to guessing is reduced in the estimates from the adaptive design, as explained previously, and hence, there is no need to tailor these responses. The comparisons used to determine the threshold estimates of the partial credit items involved the same sets of responses that were used for the comparison of the estimates from the dichotomous items and this was described in a previous section.

The graphs showing the relationships between the first and second threshold estimates (un-centralised thresholds) for each of the 15 partial credit items, and for each design are provided in Figures 9.10 and 9.11. In both graphs the values on the horizontal axis are the estimates of the tailored responses from the non-adaptive design and the values on the vertical axis are the estimates from the adaptive design. For both sets of estimates the same anchor file was used. Mean item anchoring was used to generate estimates of items for the graph provided in Figure 9.10 and individual item anchoring was used for the graph provided in Figure 9.11.
With mean item anchoring, as shown in Figure 9.10, the first threshold estimates of eight of the 15 items are lower in the non-adaptive design than in the adaptive design while the first threshold estimates for four items are higher. For the second threshold, the estimates are higher or equal in the adaptive design for all but two items. The scale in the non-adaptive design is much smaller than that in the adaptive design for just the partial credit items as well as for all items. The range of the estimates is 3.77 for the non-adaptive design and 6.04 for the adaptive design when all items, both dichotomous and polytomous, are included.

For individual item anchoring, as seen in Figure 9.11, there are six first threshold estimates that are greater in the non-adaptive design than in the adaptive design and the same number that are smaller. Most of the second threshold estimates are higher in the adaptive design than in the non-adaptive design. Again, the range of difficulties is smaller for the estimates in the non-adaptive design when only the partial credit items are considered but when all items are considered the difference in the ranges is less pronounced. The range of the estimates is 4.58 [-2.27, 2.3] for the non-adaptive design and 4.89 [-2.58, 2.31] for the adaptive design when all items, both dichotomous and polytomous, are included.
Three concerns emerged during these analyses. First, the threshold estimates from the adaptive design showed greater disordered than those of the non-adaptive design, even when the threshold distance was quite large in the latter. This was observed when partial credit was first allocated to responses in both designs and occurred for various subsequent analyses. Second, the use of mean item anchoring reduced the range of the threshold estimates for the tailored responses to the dichotomous items. As a result of this reduction of the range, these dichotomous items from the tailored responses had difficulty estimates which did not reflect the relative range of difficulty that had been identified for the same responses prior to this form of anchoring. The third concern related to the use of individual item anchoring. When individual item anchoring was used, two more items which had been identified for partial credit scoring had disordered thresholds for responses to both designs of the test, and this had not occurred in any previous analyses involving these two items.

*Figure 9.11* Comparison of first and second threshold values (T1 & T2) in the non-adaptive design with the thresholds in the adaptive test design with individual item anchoring of the dichotomous items (SD = standard deviation)

Non-adaptive: SD (T1) = 0.51   SD (T2) = 0.79  
Adaptive:   SD (T1) = 1.06   SD (T2) = 1.05
9.6.2 Threshold estimates

The difficulty estimates of the items which retained dichotomous scoring, and which had been obtained from the tailored analysis from the non-adaptive design, where the items had been scored dichotomously, were chosen as the basis for establishing the scale for the threshold estimates for the items which were rescored as partial credit. Being from the tailored analysis, the effect of guessing on these estimates had been largely removed, and therefore, they were less biased than the estimates from the responses in both the adaptive design and the non-adaptive design without a tailored analysis.

The estimates from the adaptive design for the dichotomous items were considered less reliable than the estimates from the tailored analysis from the non-adaptive design because of the numbers of students completing some of the items and because of the proportion of missing responses (omission rates). Only 43 students completed the most difficult pathway in the adaptive design (Blocks 8L, 8M, & 8H) and, for five of the ten items in the final block (8H), the rate of omission was about 10%. For the adaptive design, the overall rate of omission was 2.4%, compared to 1.8% for the non-adaptive design. While this discrepancy is unexpected, the overall rates of omissions in the two designs are quite low.

There are two values used to define the partial credit items when the scores can be 0, 1, or 2: the first threshold value and the second threshold value. The significance of these threshold values has been described in a previous chapter. For the final scale on which to place the students, the two thresholds need to be identified. The second threshold value represents the estimate of the difficulty of selecting a correct response which receives a score of 1 in dichotomous scoring, or in this study, 2 for the polytomous scoring. Therefore, the value that I chose for the second threshold estimate for each partial credit item was the item difficulty estimate of the dichotomously scored form of the item from the tailored responses in the non-adaptive design. This scoring retained the scale position of the second threshold to be the difficulty of the dichotomous form of the item.

To calculate the first threshold values for the partial credit items, I subtracted the threshold distance for each partial credit item from this second threshold value. Two options were considered for this threshold distance, both based on the non-adaptive design. These options were the estimates from tailored and non-tailored analyses. Because the threshold distances for the 17 items with partial credit were stable across these two analyses, I decided to use the estimates from the tailored design. Further detail of the identification of these threshold distances, and the justification for their use is provided in Appendix 9.1.
In summary, the final scale of item difficulty was based on the estimates from the tailored responses for the non-adaptive design when all items had been scored dichotomously. These estimates were used in two ways. First, the dichotomous items on the final scale retained the same difficulty estimates from the tailored analysis. Second, the second threshold values on the final scale were allocated the difficulty estimate of the corresponding items from the tailored responses. Then, estimates of the first threshold for these partial credit items were calculated by subtracting the threshold distances, which had been determined when the tailored responses in the non-adaptive design were scored with partial credit, from these newly identified second threshold estimates.

9.6.3 Fixing the scale

After the final scale of item difficulty had been fixed, it was possible to place all students, regardless of which test design they had been offered, onto the same proficiency scale. Having a common scale is necessary for comparing the impact of awarding partial credit, or for comparing student performance according to some other variable. Attention to the development of an accurate scale is also essential for the provision of precise measures of student achievement and to establish an accurate ranking of performance.

Fixing the scale using the estimates based on the tailored non-adaptive design meant that the items, which had previously been removed from partial credit scoring because of disordered thresholds in the adaptive design, could be reinstated as partial credit items. Item 11 was still excluded because the analysis could not continue with Item 11 when the responses were tailored. For the final scale, there were 17 items scored with partial credit and these are listed and described in Figure 9.8.

To examine the impact on the student performance of awarding partial credit in MC items, the distribution of student proficiency when partial credit was not scored was compared with the distribution when 17 items were scored with partial credit. The distributions were identified after all students from both test designs were placed on the common scale which was formed using the estimates which had been determined using the process described above. This meant that both proficiency distributions could be compared and that any difference could be attributed solely to the effect of awarding partial credit for 17 of the items. For the following discussions of the impact of awarding partial credit, this fixed scale has been used to determine measures of proficiency for all students.
9.7 Impact of awarding partial credit

The results of this investigation, as described previously, have shown that it is possible to place all persons on the same scale even though they have responded to different designs for the administration of the same test items. There is evidence of an improvement in the precision of the estimates of person proficiency when partial credit is awarded in MC items.

9.7.1 Person proficiency estimates

To examine the impact of awarding partial credit, two sets of statistics to describe person proficiency were determined: the statistics when all items were scored dichotomously and those when some items were scored with partial credit. These were described in more detail in an earlier section. The results given in Table 9.8 show that the mean proficiency estimates for all persons, as well as for males and females, increased when partial credit was given. The increase in proficiency was similar for males and females but regardless of partial credit scoring, the performance of the males was significantly better ($p < 0.01$) than that of females. With the responses of all persons now included, it is important to remember that the responses from the non-adaptive design include the guessed responses, which had been removed in some of the analyses described previously.

<table>
<thead>
<tr>
<th>Table 9.8 Awarding partial credit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Information function</td>
</tr>
<tr>
<td>Max value</td>
</tr>
<tr>
<td>Location</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Mean estimate for person proficiency</td>
</tr>
<tr>
<td>(standard deviation)</td>
</tr>
<tr>
<td>All persons from both designs</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Non-adaptive design (not tailored)</td>
</tr>
<tr>
<td>Adaptive design</td>
</tr>
<tr>
<td>Person reliability index PSI</td>
</tr>
<tr>
<td>Person skew</td>
</tr>
<tr>
<td>Person kurtosis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Information function</th>
<th>No partial credit</th>
<th>Partial credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max value</td>
<td>11.4</td>
<td>15.2</td>
</tr>
<tr>
<td>Location</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean estimate for person proficiency (standard deviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All persons from both designs</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Non-adaptive design (not tailored)</td>
</tr>
<tr>
<td>Adaptive design</td>
</tr>
</tbody>
</table>

| Person reliability index PSI | 0.813 | 0.833 |
| Person skew                  | 0.44  | 0.42  |
| Person kurtosis              | 0.58  | 0.62  |

The PSI was high for both scales of proficiency and it indicated that there was a good range of items which were available to separate persons on the scale of achievement.
9.7.2 Person proficiency distributions

Scoring partial credit resulted in an increased number of students with proficiency estimates between 0 and 1 logits. This can be seen from the comparison of the histograms showing person distributions in Figures 9.12 and 9.13. The increased frequency was expected but the location of this increase had not been previously considered. The persons appearing to benefit from partial credit scoring are those below the median because they have gained credit and hence higher proficiency estimates.

*Figure 9.12*  Person-item threshold distribution: dichotomous scoring

*Figure 9.13*  Person-item threshold distribution: partial credit given
When partial credit was given, the maximum value of the information function, also shown in Figures 9.12 and 9.13, increased from 11.4 to 15.2 and the location of this function was the same on both proficiency scales. There was also an increased number of item thresholds, as shown by the item frequencies below the axis, particularly on the scale where the location estimates range from -0.2 to 1.0 logits. This increased number of thresholds coincided with the increased frequency of persons, and this translates to an increased precision of the proficiency estimates in that part of the scale. The partial credit scoring had a negligible effect on the values of the skew and kurtosis values for person distribution. This is reflected in the similar nature and shape of the two distributions as shown in Figures 9.12 and 9.13.

9.7.3 Test design

The students sitting the non-adapted version of the test have outperformed the students sitting the adapted version and their proficiency estimates are higher regardless of the recognition of partial knowledge. This result is expected because there is guessing in the responses and students are benefitting from the inflation of their scores. The difference in performance of the students on the two versions of the test was only significant at the 5% level. There was very little change in this difference in performance when partial credit was scored and the level of significance of the difference between the two test designs was unaltered.

The histograms showing the student proficiency distributions for each version of the test, with and without partial credit scoring are provided in Figures 9.14 and 9.15. These histograms indicate that for the students doing the adaptive design, there is an increase in the number of students with proficiency estimates higher than -1 logit when partial credit is scored. For the adaptive design the increase is more noticeable for proficiency estimates above 0.8 logits. One possible disadvantage for students doing the adaptive design might be related to the number of students who were offered each of the different pathways and, in particular, the small number of students offered the most difficult items in the test.

The effects of scoring partial credit are independent of the test design and reliable conclusions about partial credit scoring and tailoring for guessing can still be made. However, future research is necessary to ensure that students are not disadvantaged by being offered an adaptive design rather than a non-adaptive one.
Figure 9.14  Histogram of person proficiency for both test designs: no partial credit

Figure 9.15  Histogram of person proficiency for both test designs: partial credit
9.8 Further considerations

9.8.1 Impact of guessing on person proficiency

The impact of removing guessed responses on the estimates of item difficulty was presented in an earlier section and in summary, the mean estimate increased when tailored estimates were used and the range of difficulties increased. The range and mean estimate of person proficiency also increased, but the extent of this effect was different for males and females.

Gender differences

Results show that removing guessing has reduced the advantage for males in this Year 8 test. The mean proficiency estimate of males was significantly higher at the 5% level than that of females in both versions of the test and while the performance of males was still higher when guessing was removed, the difference was not significant. The proficiency estimates for both test designs and for the tailored responses are provided in Table 9.9. As expected, the mean person proficiency estimates have increased with tailoring and these results also confirm the reduced amount of guessing in the responses for the adaptive design.

Table 9.9 Gender differences with guessing

<table>
<thead>
<tr>
<th>Source of estimates (guessed responses)</th>
<th>Sets of responses</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-adaptive</td>
<td>Adaptive</td>
<td>Tailored</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(included)</td>
<td>(included)</td>
<td>(removed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mean proficiency estimates and standard deviation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>-0.62 (0.98)</td>
<td>-0.37 (1.37)</td>
<td>-0.46 (1.37)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>-0.77 (0.89)</td>
<td>-0.68 (1.15)</td>
<td>-0.59 (1.21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>0.15</td>
<td>0.31</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Level of significance (t-test)</strong></td>
<td>Probability of significant difference</td>
<td>0.027</td>
<td>0.014</td>
<td>0.168</td>
<td></td>
</tr>
</tbody>
</table>

As described earlier in review of the literature, the advantage for males is often linked to their perceived tendency to take greater risks than females and hence, to guess when they do not know the answer. In the Year 8 test, the rate of omissions for all responses was the same for both genders: The proportion of items unanswered by both males and females was 2%. However, in the non-adaptive test, the rate of omission for females was 1.6% and for males it was 2.1%. These values suggest that the advantage for males in the adaptive and non-adaptive
designs is not related to their greater willingness to take risks and guess responses. The removal of the guessed responses reduces the gap in performance between males and females and the results indicate that males are more successful with guessing than females. Further research is needed to confirm and explain this suggestion.

Some evidence of the different effects of removing guessing on the estimates of male and female proficiency can be seen in the histograms provided in Figures 9.16 and 9.17.
It appears from these histograms that the frequency of females with proficiency estimates between -1 and 0 increased when the responses were tailored to remove guessing and the proportion with higher estimates did not decrease. Similar effects for males are not as noticeable in these histograms. This change in frequency, for which no further explanation is available, might be responsible for the decrease in the significance of the gap in gender achievement when responses are tailored.

The histograms provided in Figures 9.16 and 9.17 also show the numbers of students at lower proficiency levels who have lost the advantage of guessing, as well as the numbers of students who have higher proficiency estimates when responses are tailored. An explanation for this change in the proficiency scale with tailoring was provided in a previous section.

9.8.2 Checking fit

DIF

The item estimates were checked for two violations of the Rasch model, DIF and local dependence. Only one item showed DIF. DIF for gender occurred in the dichotomous item, Item 8, regardless of partial credit scoring for the other items. This DIF is shown in Figure 9.18 where there is a much lower performance for females in three of the class intervals towards the lower end of the scale.

The item was one of the easiest items and one where partial credit was not justified. The students were asked to recognise the relative length of the radius of a small circle when given the radii as relative fractions. Given that, (a) there was DIF in only 1 of 60 items, (b) there were only notable differences in achievement in the middle of the continuum, and (c) there is
no likely explanation for such DIF, this outcome can be attributed to chance. One could not conclude that this item provided an opportunity for males to be more successful than females.

Three items showed DIF for test design when there was no partial credit scoring, and the same three items, plus one extra item, showed DIF when the partial credit scoring was applied. In all items there were class intervals where students of particular proficiency had not been offered the item. This occurred for Items 51 and 59 which were in the easiest block, and which were not offered to students with high proficiency estimates. Similarly, Block 3 in the adaptive design was only offered to students of high proficiency and the DIF indicated in Item 27 was not unexpected, given that the less proficient students were not offered the item.

**Response dependence**

An examination of the correlation of the residuals revealed the presence of two items for which there was possible response dependence. An examination of the items indicated that they were not structurally dependent, and this suggests the correlation is due to chance or to another factor, other than the students receiving a clue from one item that would help them answer another. It is unlikely that items requiring students to manipulate percentages and speeds would help a student to recognise the length of a ruler that is one-third of a given length. The other correlation was between an item on adjusting a proportion and an item on averaging percentages. Item content has not indicated any reason for possible response dependence and with the small number of correlations and the low correlation coefficients, it is possible that these relationships are all due to chance alone. We can conclude that the item estimates are independent of each other as required for appropriate fit to the Rasch model. Confirming the fit of the responses to the Rasch model implies that the item estimates are reliable, and therefore, any conclusions drawn from interpreting differences in person proficiencies can be accepted with confidence.

9.8.3 Development of a learning trajectory

It is possible to use the scale for threshold difficulty, which has been developed in this study, to identify a learning trajectory: a probable order in which students might develop the knowledge and understandings associated with becoming skilled in concepts associated with proportional reasoning. For the identification of the learning trajectory, which is not one of the intended outcomes of this research, both partial credit and dichotomous items would be used to describe the probable order of development.
Part of the Item Map, provided by the RUMM2030 software, is reproduced in Figure 9.19. It shows the order and degree of difficulty of the thresholds for the partial credit items. The correct response to Item 17 has a threshold difficulty estimate of 2.2 logits while the partial knowledge created for the item has a threshold difficulty estimate of 0.4 logits. The skills associated with each threshold can be described and ordered for this item and for all items, thus creating an order of difficulty and a probable order in which students develop these skills: in other words, a learning trajectory. The inclusion of the skills associated with partial knowledge provides considerable extra information about student learning, and such inclusion would be of great benefit for teachers in their planning of learning activities.

| 2.6 | x | 26.2 |
| 2.0 | x | 17.2 |
| 1.0 | x | 15.2 |
| 0 | x | 38.2 |
| -1.0 | x | 39.2 |
| -2.0 | x | 14.1 |
| -2.6 | x | 15.1 |

Figure 9.19  Item map showing the order of threshold difficulty and the distribution of students at each class interval (× = 7 students, #.2 = second threshold, #.1 = first threshold)
9.9 Summary: Effects of awarding partial credit

Initially, all 60 items were scored dichotomously and the expected bias in the item estimates for the two test designs was confirmed. With this scoring, the performance of males was significantly better than that of females in both the non-adaptive and the adaptive designs. The blocks of MC items had been constructed to reflect varying levels of difficulty for the Year 8 students responding to the test items. The students’ results indicated that the relative difficulty of the blocks of items was similar to that proposed in the test framework.

All items had been constructed to contain partial knowledge and for the next part of the study, polytomous scoring was applied with a score of 2 allocated for the correct response and 1 for the response deemed to demonstrate partial knowledge. For 18 of the 60 items, partial credit scoring was successful, and the skills associated with the informative distractors in those items were confirmed as partial knowledge of the associated concepts. The selection of items for partial credit scoring was informed by the presence of ordered thresholds, the appearance of the distractor curves, and the size of the threshold distance. With the allocation of partial credit, many students gained extra scores for their selection of the informative distractor.

The item difficulty estimates generated from the dichotomously-scored responses to the two test designs were different and all students could not be put on the same scale. Hence, two scales were constructed, and the first scale consisted of the item difficulty estimates from the tailored analysis. For tailoring, all responses likely to have been guessed had been removed. For the second scale, the item difficulty estimates from the tailored analysis were used for the dichotomous items, and the threshold estimates for the polytomous items were determined by adjusting the item difficulty estimates from the tailored analysis. All persons, regardless of which test they had been offered, were placed on both scales and the different outcomes were attributed to the allocation of partial credit for the 17 polytomous items (Item 11 had been removed during tailoring). The removal of guessing appeared to reduce the advantage for males in the Year 8 test and the award of partial credit provided more information about student proficiency but did not provide a perceived advantage for either males or females.

These results show it has been possible to provide partial credit for MC items and to counter the effect of guessing in student responses. Further discussion of the impact of these two outcomes and their associated effects are provided in the following chapter.
CHAPTER 10: FURTHER DISCUSSION AND CONCLUSION

In this chapter, the research questions and a summary of their answers is given before the overall summary which includes the final discussion and conclusion. The results of the analysis of the responses to the CR and MC items in the ICCAMS test, the NAPLAN numeracy tests, and the Year 8 test, have been considered in answering the three research questions posed for this study.

Research question 1

For distractors to contain information about student achievement in mathematics, how should multiple-choice items be designed and created?
Identification of what constitutes both partial knowledge of a concept, and high levels of thinking, assists in the creation of distractors with information about student learning. The form of the partial knowledge may involve a misconception, part-understanding of a concept, or the recognition of the correct process but the inability to use the process. The inclusion of high-order thinking in a MC item provides greater opportunities to create distractors with information than are possible when only low levels of thinking are assessed.

Research question 2

To what extent is the proficiency scale for achievement in proportional reasoning affected when distractors with information about student learning are awarded partial credit?
When informative distractors are awarded partial credit, the information gathered about student learning is increased significantly, particularly for students in the middle of the scale of achievement. The results of this research confirm that the award of partial credit for partial knowledge of proportional reasoning is supported by the analytical model and that there is extra information about student learning, while still satisfying measurement principles.

Research question 3

To what extent is the proficiency scale for achievement in proportional reasoning affected when guessing is reduced in MC items?
The results of countering the effects of guessing in MC items in this study confirm earlier research which showed that the scale of proficiency is stretched in both directions when guessing is removed from the responses to the items. In the earlier research, the reduction in guessing had been achieved by post-hoc tailoring but in the current study, this reduction was also achieved by the inclusion of an adaptive design for the test.
Overview of research

The focus of this research has been to investigate ways to improve the function of MC items for the assessment of mathematical understanding of lower secondary students. The MC format for testing skills and knowledge has become more prominent in recent years, and it is both feasible and worthwhile to continue to develop guidelines for the construction and analysis of more effective MC items. It is important that the items work well to provide accurate measures of student achievement and indications of the quality of student learning. While MC items are challenging to create, they are able to cover a wider range of the curriculum than CR items and they are attractive for students, easy to score, and allow impartial scoring of the responses. Along with CR items, MC items continue to have a vital role in the assessment of student learning.

Several concerns with the use of MC items have been raised by educators and particularly by teachers of mathematics; some of these have been addressed in this thesis. The first of these concerns relates to the perception that MC items only assess knowledge and low levels of skills, and that they cannot test the application of knowledge and other higher-order thinking skills. Other concerns relate to the inflation of scores due to guessing, the lack of qualitative and statistical information that is gathered when students respond to MC items, and the perception that males achieve better results than females with the MC format. In summary, the aim of this research was to provide evidence of the success of processes that can be used to alleviate these concerns. These processes can promote greater support for the use of MC items as well as more accurate achievement information and fairer outcomes for students.

10.1 Research activity

Before providing further discussion of the outcomes of this study, a summary of all research activities relevant to the addressing of these concerns is provided. Using the findings from earlier research studies, items were created to assess higher-order thinking in the Year 8 test and the success of constructing such items was evident from the scale of item difficulty that was produced. The impact of guessing in MC items was examined by comparing the results of analysing both tailored and non-tailored responses from the Year 8 test and the NAPLAN numeracy tests. In the Year 8 test, the use of the two designs, the adaptive and the non-adaptive designs, provided responses with different levels of guessing and it was possible to compare the outcomes from these two designs.
The context chosen for the study of MC items was student proficiency in the field of proportional reasoning. Because of my previous teaching and research experience and interest, I chose to study the achievement of students in early secondary schooling. Proportional reasoning provided a challenging and interesting context for the study because of its importance for people of all ages and its difficulty for students in the early secondary age group. The limited amount of information that can be gathered from a MC item that tests the skills of proportional reasoning relates to the scoring as correct of just one of the options selected by a student. One of the aims of this study was to increase the capacity of a MC item to capture more student understanding than just the ability to identify a correct response. To determine how to increase the information about student understanding from an item, distractors that address partial knowledge were needed for the study.

MC items with distractors which address partial knowledge are not widely implemented and it was necessary to create such items. This involved identifying what could be described as partial knowledge, and in particular, what is considered as partial knowledge of skills associated with the development of proportional reasoning. These skills were identified by using findings from studies reported in the literature and from the results of a test which was conducted with Year 8 students. Further insight into the skills associated with partial knowledge of proportional reasoning was determined from the analysis of the responses to both CR and MC items. These responses, which had been generated from other assessment programs, were provided to the author by researchers in the field. The identification of partial knowledge skills provided concepts and ideas for the creation of distractors deserving partial credit, thus allowing more information about student achievement to be gathered from each MC item than from dichotomous scoring alone.

The final topic addressed in this study was the perceived advantage for males when items, which are used to assess mathematical understanding, are formatted as MC items. To gather evidence to confirm this perceived advantage, the effects of reducing guessing and of awarding partial credit, on the proficiency estimates of the male and female students were compared. The different rates of item omission for males and females were also compared as these were previously linked to the superior performance of males. The responses to the MC items, in the data which were analysed, were coded for male and female, and thus it was possible to examine any gender differences.
Three research questions were formulated to address the concerns that were described at the beginning of this chapter. The aim of the first question was to determine how to design and create MC items with distractors which contain information about student achievement. The aim of the other two questions was to determine the extent to which the proficiency scale for achievement in proportional reasoning is affected, first, when distractors with information about student learning are awarded partial credit, and second, when guessing is reduced in MC items. These research questions are considered in the following discussion of the outcomes of investigating the concerns.

10.2 MC items to address higher-order thinking

The perception that MC items can only assess basic knowledge and skills may be linked to the more common practice of only writing items that test low levels of thinking (Martinez, 1999). However, this perception cannot be attributed to the MC format, as the results of this study have shown that it is possible to identify and to construct MC items that address higher-order thinking. It follows then, as suggested by Zoumboulis (personal communication, July 8, 2015), that creators of MC items are capable of constructing items that assess complex reasoning and higher-order thinking. Higher-order thinking in mathematics can be indicated when students analyse data, generalise results, justify decisions, create knowledge, follow or evaluate arguments, analyse assumptions, use abstract reasoning, perform complex calculations, or interpret unfamiliar situations (Bloom et al., 1956, McCurry, 2008).

Items 24, 25, and 26 were located in the most difficult block of items and, based on the estimates of item difficulty, they were three of the four most difficult items in the test created for the Year 8 students. Item 24 was described earlier, and it involved division by 112%, which is a complex calculation for Year 8 students, and one with which they were unlikely to be familiar. In Item 25, the students could have used abstract reasoning to recognise that an increase of a quarter of the mass of a chocolate bar would be needed to return the bar, which had been reduced by a fifth of its size, to the original mass. The correct answer could have also been determined more readily, again using higher-order thinking, if the student had trialled a theoretical value to identify their response and to justify their decision. In Item 26, the ability to follow and evaluate a mathematical argument would have helped students to be successful. These three items that assessed higher-order thinking were just some of the items which provided evidence of the possibility of creating such items.
Further evidence of the success of the deliberate construction of items prompting high levels of thinking was seen in the mean item difficulty estimates for the blocks of items in the Year 8 test. Regardless of test design, the means for the two blocks which had been designed to be the most difficult, were greater than the means for any other blocks. Furthermore, the three blocks designed for the lowest levels of difficulty (the standard level) had lower mean estimates than those of the other three blocks. These two observations applied regardless of the inclusion of guessed responses in the analysis. The order of difficulty did not follow exactly as planned for construction, but these observations indicated that the attempt to construct items that tested higher order thinking was successful. The fact that the items, which were designed to test high levels of thinking of the Year 7 curriculum, were more difficult than those designed to test low levels of thinking of Year 8 content warrants further investigation. Such an investigation could provide evidence to determine if items are more difficult because of the levels of thinking, or because of the challenging nature of the content.

Estimates of item difficulty from the Year 8 test and the NAPLAN numeracy tests were across a range of values, indicating that there were items that assessed a variety of levels of mathematical thinking. Having easy items in the test may not have provided much statistical information about the measurement of achievement, but it has indicated some level of learning and it has provided the very weakest students with opportunities for success and the incentive to continue with the assessment. The more difficult test items provided opportunities for the more proficient students to demonstrate their skills and hence be assessed as having achieved higher levels of understanding. This is an important outcome to continue to achieve if MC items are to be more widely accepted as valid for the measurement of a wide range of abilities.

When MC items assess high levels of thinking for the students tested, the possibility of creating items which can assess partial knowledge is enhanced. It is more feasible to create distractors which can be scored with partial credit in items that address high levels of thinking, than in items which only address basic knowledge and skills. The distractors for partial credit can contain some aspects of the correct response at lower levels of thinking than are required to select the correct response. The difficulty of these items is relative to the proficiency of the students and what might be easy for a Year 8 student could constitute partial credit for a Year 6 student. Hence, it is acknowledged that this discussion of the role of the distractors in the collection of information about student learning, relates to the students participating in this test and may not apply for other populations of students.
Providing MC items where the selection of different options has indicated different levels of thinking has added to the information collected about student learning and has increased the efficiency with which MC items functioned. The construction of MC items to test higher levels of mathematical thinking is one way to design and create MC items which have distractors which contain information about student achievement in mathematics: one answer to the first research question posed for this study.

10.3 Controlling for guessing in MC items

It is essential to minimise the impact of guessing in MC items, not only because of the inflation of the scores particularly for the weaker students, but also because of the effect that guessing has on the students who do not guess or who guess to a much lower extent. When guessed responses are included in the data, items appear easier than they really are. As a result, the more able students who answer the more difficult items correctly at a greater rate than if they had guessed, do not get the benefit of being successful on these more difficult items. Furthermore, this inflation has recently been found to also affect the estimates of difficulty for the CR items which are included in tests with MC items (Andrich & Marais, 2018). Such a finding reinforces the need to control for guessing.

To reduce the level of guessing in the Year 8 test, several aspects relating to the design and construction of the items were considered. The use of key randomisation was considered but this was problematic, in that it was difficult to achieve unless the order of the items was flexible. Because each block had items which addressed the variety of concepts associated with proportional reasoning, and each block was set at a particular level of difficulty, key randomisation was not practical. Some of the methods to control for guessing which were described previously, for example, penalty scoring, made extra demands on the test participants so these were not considered for this study. The recommendation by Haladyna et al. (2002), to reduce guessing by keeping options of a similar length and structure, was adopted for the construction of items in the Year 8 test.

As well as attending to item design to reduce the impact of guessing, attention was given to the analyses of responses to the MC items and to the design of the Year 8 test. Post-hoc analyses of the responses in the Year 8 test and the NAPLAN numeracy tests involved tailoring the students’ responses. Tailoring the responses involved the removal of the responses of the students who had been identified as likely to have guessed their answers.
While the initial rationale for the use of the adaptive design was to provide alternate test pathways for students of varying abilities, it was also evident that such a design should reduce guessing. As expected, there was less guessing by the students who were administered the adaptive design of the Year 8 test.

Evidence of the increased item difficulty as a result of tailoring, especially for the more difficult items, was seen in the change in the relative difficulties of the blocks (Table 9.3) and in the graph showing the relationship between the item estimates determined before and after tailoring (Figure 9.1). The increase of the item difficulty range at the upper end, from a tailored analysis provided further evidence. These all indicated that tailoring had a greater effect on item difficulty for the more difficult items and supports the assumption that the more difficult the item, the more guessing is occurring on the item. The histograms representing the distribution of person proficiency (Figures 9.2 and 9.3), show clearly that tailoring results in an increased number of persons with lower proficiency estimates, as well as an increased number of persons with higher proficiency estimates. These results confirm the work of Andrich et al. (2015).

Having a non-adaptive and an adaptive design for the Year 8 test provided an opportunity to compare the levels of guessing and to examine the effects of removing guessing on the estimates of item difficulty and person proficiency. As expected, there was much less guessing by the students doing the adaptive design because they were offered items commensurate with their levels of achievement and with their understanding of the mathematical concepts in the test. This was confirmed when the item difficulty estimates from the adaptive design were much closer in value to those produced from tailoring the responses from the non-adaptive design than those produced from the untailored responses in the non-adaptive design.

The introduction of this adaptive design has reduced the advantage from guessing for the weaker students. However, these proficiency estimates were determined with the guessed responses excluded and this would not be acceptable when reporting student achievement to the students and their parents. It seems reasonable that reporting the numbers of students reaching certain benchmarks, the improvements in numeracy, and the student progress across the years could be done using the proficiency estimates which were generated without the guessed responses. It is not known which estimates are used for such reporting and further study of what difference this might make could be enlightening.
The results of this study indicate that using an adaptive design is preferable to using a non-adaptive design for tests of mathematical understanding, as it has resulted in a reduction of the impact of guessing. At the same time, a test was available, and this test was not too hard for the weaker students and it provided challenging questions for the more able students. It is worth noting that since the conceptualisation of this research, adaptive testing has been developed and trialled for the NAPLAN numeracy tests (ACARA, 2017b). It is not known if this change in practice is due to an effort to reduce guessing, but Martin and Lazendic (2018) have reported that computer-adaptive testing in numeracy has the potential to increase measurement precision.

Some aspects which have not been addressed in this part of the study of MC items, include the nature of the guessing, and the level of confidence that a student might have in their selection of an option. Our research into guessing would be enhanced if we could identify the nature and degree of guessing and the student behaviours associated with the selection of a response. This would at least require us to distinguish between totally random guessing and systematic guessing, which was described previously as the ability to eliminate one or more distractors with guessing between the remaining options. With all the different degrees of guessing, and the different guessing behaviours, it has not been practical, nor the focus of this study, to compare outcomes based on the various guessing behaviours.

When the responses were tailored, it was assumed that guessing is the same for each item of similar difficulty regardless of content and design. Given that the distractor curves can indicate where guessing is occurring, it is possible that they could also be used to indicate more specifically in which items guessing is occurring. Then, rather than remove responses from all items based on the probability of guessing, a subset of items could instead be targeted. It would also be interesting to investigate the link between the nature and extent of the guessing and any partial knowledge that might be assessed in the items.

The above discussion is considered in the context of guessing as a function of the difficulty of an item relative to the proficiency of a person. A student who has a high proficiency relative to the range of item difficulty will find an item easy and is unlikely to guess. In comparison, a student with a very low proficiency relative to item difficulty will find the same item quite challenging and will be unable to distinguish between any of the options provided. Thus, they will guess at random between each of the options. In between these extremes, there are many different behaviours of the kind mentioned above. Guessing, therefore, is a matter of degree and depends on both the property of the item and the behaviour and proficiency of the person.
10.4 Informative distractors

Several researchers have stressed the positive relationship between the quality of the distractors and the quality of the MC items, but little research has been located to indicate that this quality is a widespread concern. Many reports have suggested that distractors should be plausible and achieving this involved using common errors, misconceptions, or predictions of what students might think if the item invited an open response, as plausible options in the MC item. Distractors such as these may be classified as being informative because they provide insight into student learning and in this study, informative distractors, or distractors with information, were defined as those that added to measurement information, that is, these distractors allowed the determination of more precise estimates of student proficiency.

Informative distractors in MC items have been linked to the behaviour of the distractor curves and this study has provided further evidence that the shapes of these curves can be used to predict the presence of an informative distractor. In some research studies, the attempts to purposefully create what would be considered as informative distractors have involved the use of cognitive diagnostic models and the different stages of concept development (Briggs et al., 2006; de la Torre, 2009). For the Year 8 test of skills associated with the development of proportional reasoning, there was too little information available in the research studies, to use the different stages of learning a concept to create item distractors. Instead, the use of part-knowledge of a concept, as explained in the next section, was the basis for the successful creation of distractors with information. The behaviour of the distractor curves was used to identify, and to confirm the presence of, informative distractors in both the Year 8 test and the NAPLAN numeracy tests.

Dysfunctional distractors, which were previously defined as those selected by less than 5% of the test cohort, are believed to contribute little to the measurement information for the cohort tested. In the Year 8 test, there were only four items with dysfunctional distractors and while such items detract from the difficulty and the discriminatory power of the items, the presence of some easy items provides support for weaker students. The ability to effectively ignore a distractor because it is easily recognised makes an item easier for the students. Two of the four items with dysfunctional distractors were in the common block that was offered to all students sitting the test. It is possible that the presence of these easy items among the first 10 items provided some confidence to continue with the assessment. Such practice should be available in all tests to allow students to show some aspects of their learning.
10.5 Partial knowledge

Describing partial knowledge

Having partial knowledge of a concept should indicate being on the way to developing full knowledge; more than knowing nothing, but not a complete understanding. This definition is an attempt to describe partial knowledge which has been given many connotations in the research literature. These definitions include the idea of partial knowledge as a rational error, an indicator of progress, a misconception that develops during conceptual development, an error that the student makes as their understanding increases, or the use of a correct process in the wrong situation. These errors, while still considered as part-knowledge of a concept have their “roots in productive and effective knowledge” (Smith et al., 1993, p. 24).

In this study, the existence of partial knowledge has been confirmed for both CR and MC items. In the CR format, the students supplied the part-knowledge of understanding and in the MC format, the students had opportunities to select options which were classified as indicating partial knowledge of the concepts being tested.

Recognising partial knowledge from item content

Various items of information have assisted with the recognition of what constitutes partial knowledge of concepts associated with skills relating to proportional reasoning. These include the descriptions of the behaviours associated with proportional reasoning (Table 4.1) and the developmental pathways described in Chapter 4, as well as the studies reported in the research literature. By using my own mathematical expertise, my experience of teaching lower secondary students, as well as studying the research literature, I have been able to interpret the curriculum documents and predict likely student errors and misconceptions. The ability to recognise and describe partial knowledge requires considerable research and expertise.

Many examples of the partial knowledge of concepts relating to proportional reasoning have been described in this thesis. From the analysis of student responses to test items, and from the examples described in the research literature, five different types of partial knowledge were identified: additive thinking, proportion ignored, incomplete solution, reasonable estimate, and incorrect method. The results of this study have confirmed the existence of these five types of partial knowledge, but they have also provided evidence to indicate that it is not straightforward to classify some incorrect responses as partial knowledge, nor as particular types of partial knowledge. Further research is needed to identify in which situations, and for which students, some incorrect responses, which appear to belong to one of these five types of partial knowledge, are confirmed demonstrations of partial knowledge.
It is beyond the scope of this study to provide details of all the concepts associated with the partial knowledge of proportional reasoning concepts that have been identified in this research. The many examples described throughout this thesis, supported by the information provided in the Appendices, could be used to provide such detail.

Recognising partial knowledge from the analysis of responses

There are two processes by which partial knowledge has been identified or confirmed as a result of the analyses. First, the behaviour of the distractor curves indicated the potential for partial knowledge, and second, the ordering of thresholds with polytomous scoring confirmed the hypothesised partial knowledge in some items. The recognition of partial knowledge from just the distractor curves was not established as being reliable and it was imperative to consider item content, as well as threshold ordering. Further research into the behaviour of the distractor curves might provide more detail about distractor behaviour, and hence provide a reliable process for identifying the presence or absence of partial knowledge in MC items in the future.

In this study, threshold ordering was used to confirm the theoretical partial knowledge created for each item. The distractors had been created to allow students to select one option in each item which was associated with some partial knowledge behaviour described previously, for example, the recognition of common errors or misconceptions. Hence, one could claim that the student had partial knowledge of the tested concept. From the results of this study, it is suggested that claiming the existence of partial knowledge can only be justified when supported by both ordered thresholds and the qualitative evidence in the item content.

Creating distractors with partial knowledge

This study has shown that it is possible to write items to detect partial knowledge and it addresses the claim by Andrich and Styles (2011) that creating distractors with information is plausible but challenging. For the Year 8 test, 60 items were constructed with one partial knowledge option in each item, and 18 of the items satisfied the analytical requirements for partial knowledge. Considering that this is a novel approach to the construction of MC items, and it has been described as a challenging activity, this undertaking has been quite successful. The 18 items with partial knowledge have allowed students to show more of their understanding of content, by their selection of the informative distractors. This increased insight into student levels of conceptual understanding is particularly useful for teachers in their planning of learning activities.
10.6 Partial credit for partial knowledge

It was possible to provide evidence that both MC and CR items allowed students to demonstrate their partial knowledge even when the items had not been written to achieve that outcome. In the CR items, partial knowledge was confirmed when the students gave answers which were actually incorrect, but which had aspects of the correct response. Such responses were allocated a score, which is in contrast with the standard scoring in mathematics, whereby the students receive a score when they supply a correct response for part of the item and when their response does not contain any incorrect information. In the MC items, partial scores were given when incorrect options of a particular nature were selected. For the three sources of items that were studied; the ICCAMS tests, the NAPLAN numeracy tests, and the Year 8 test, the proportion of items facilitating partial knowledge was between 30% and 45%.

When partial credit was given for partially correct responses, it was mostly the relatively more difficult items where partial knowledge was confirmed and for the nine easiest items in the Year 8 test, the theoretical partial knowledge was not confirmed. The ability to diagnose this possibility is necessarily a function of the proficiency of the group relative to the easiness of the items. Providing partial credit for the items resulted in different effects for the various items and this included a range of threshold distances. These distances indicated the extent to which it was more difficult to demonstrate full understanding of the concept than to demonstrate partial knowledge of the concept.

With the increase in the number of thresholds when partial credit was scored, there were more opportunities for students to gain credit and the resulting estimates of proficiency were more precise. It was usually the students in the middle of the proficiency range who gained the most from the extra scoring. This was as expected, as the weaker students were unable to demonstrate minimum levels of knowledge to reach the first thresholds of many items, and the stronger students were unable to gain from the partial credit because their responses were fully correct.

For any one item, between 25% and 74% of the students in the test cohort benefited from the allocation of partial credit in the three different tests which were analysed. This deserved recognition of knowledge and understanding would be welcomed by the students. Being aware of the skills associated with the partial knowledge demonstrated by the students would provide vital information to teachers for their planning of lesson activities to address any misconceptions or errors, or to progress to the next stage of the conceptual learning.
10.7 Perceived gender bias

For both the NAPLAN numeracy tests and the Year 8 test, the performance of males was significantly greater than that of females and this occurred regardless of any attempts to control for guessing, and regardless of the award of partial credit. The advantage for males was absent only in the adaptive design of the Year 8 test. Similar superior performances by males have been reported in the research literature, where there have also been a few studies reporting some superior performances by females, and others where the difference in achievement is not significant.

The allocation of partial credit did not make any significant difference to the gender gap in performance in the NAPLAN numeracy tests, and there was a small and insignificant reduction in the gap when partial credit was awarded in the Year 8 test. Similar effects were noted when the responses were tailored. It would be good to be able to close the gap in the performance of males and females, but this study has shown that neither controlling for guessing, nor rewarding partial knowledge, appear to be highly effective for reducing the difference in the performances of the males and females. The perception that males are less likely to omit items has not been confirmed, either for the Year 8 test, or in the studies reported in the literature.

10.8 Further outcomes

Development of learning

There are several terms which are used to describe the order in which students develop conceptual knowledge and understanding, including learning metrics, learning progressions, developmental hierarchies, and learning pathways. These descriptions of learning, which are themselves constantly evolving, have been at times associated with the development of strategies and errors as well as skills. They may also include a form of ranking, an overview of the skills that can be demonstrated, and a nominated level of achievement. Some of the descriptions of learning, which are associated with the conceptual field of proportional reasoning, have been outlined in Chapter 4.

The findings of this study, and in particular the identification of partial knowledge, can add to the increasing awareness of the probable order in which students learn new skills as they develop their proportional reasoning. Knowing how much more difficult it is to deal with non-integral scales rather than integral scales, or with inverse rather than direct proportion,
can assist teachers in their decisions for timing the introduction of more difficult concepts. It is beyond the scope of this study to organise and present this extra detail, and to conduct any further research to refine the descriptions of learning, but it would be a useful development for teachers of mathematics of early secondary students. The Item Map presented in Chapter 9, which shows the order of difficulty of the items and the placement of the concepts associated with partial knowledge, would provide valuable guidance for the further development of descriptions of learning for the skills associated with proportional reasoning.

**Threshold distances**

One of the fascinating areas for future study is the relevance of the threshold distance in MC items. These distances represent the disparity of difficulty between the selection of the option designated as partial knowledge, and the selection of a correct response. For the reporting of the successful scoring of partial knowledge in the NAPLAN items, a positive threshold distance was the basis for defining success, but in the Year 8 test, a threshold distance of 0.4 logits was used. It would be interesting to investigate the impact of different threshold distances on the amount of statistical information generated, and to determine at what value of the threshold distance, partial credit scoring is warranted. For example, if the threshold distance is only 0.1 logits, perhaps dichotomous scoring would be just as effective in gathering information as polytomous scoring, but perhaps this, too, depends on item content.

**Analytical processes**

Several outcomes of this study relate to the processes used to detect, confirm, and score the existence of partial knowledge in MC items. Other outcomes relate to the processes which have been used to detect and counter the effects of guessing in MC items. Further to these, there are the processes which were used to investigate the impact of awarding partial credit, and of reducing the amount of guessing, on the estimates of both item difficulty and person proficiency. These processes have been sufficiently described in this thesis to show how they can be implemented for a much broader set of studies involving MC items. The analytical activities which have been described could be used in the study of student achievement in many discipline areas and for students of all ages.

The processes for investigating MC items included the examination of the behaviour of the distractor curves, which were used to identify the presence of guessing and the options for partial credit scoring. There were also processes which described the identification of item difficulty, the implementation of an adaptive design, the tailoring of responses to MC items, and the creation of a scale of item difficulty. Many of the processes have built on the work of
Andrich et al. (2012), Andrich and Styles (2011), and Van Wyke (2003). It appears from this investigation that there are two areas relating to the analytical processes, where further research would be beneficial for improving the function of MC items. The first suggested area for study is the use of MC items with partial credit scoring as link items. Such a study might indicate that fewer link items would be needed, if only those items with partial credit scoring are used to link tests, when not all students respond to all items. The second suggested area for further research relates to the formation of the scale of item difficulty when there are two test designs. In the study described in this thesis, the scale was developed and applied, and its use was justified in part, by logical argument. It would be beneficial to further test this process of creating a scale and to investigate the possibility of justifying such a scale using statistical measures.

10.9 Concluding remarks

The focus of this research has been to investigate ways to improve the function of MC items in response to the many concerns about their usage, particularly for the assessment of mathematical skills and understanding for lower secondary students. There were three specific research questions addressed in this investigation. In answering each of these questions, information has been provided to support the improved function of MC items by attending to the design and creation of the items, and by identifying the impact on the proficiency scale, when partial knowledge is scored and when guessing in student responses is reduced.

Given the number of students providing responses to items testing mathematical understanding, and the fit of the data to the Rasch measurement model, we can be confident that the results described in this report are reliable and valid. The evidence provided has shown how MC items can be improved by attending to their construction and analysis, and this suggests that such action can be applied in a greater variety of educational settings. This has the potential, not only to allay the concerns of educators, but also to provide improved measures of student achievement and to gain greater insight into student learning.

The outcomes of the assessment process have a large impact on the education of students, especially in regard to their future learning and employment. For the benefit of the students and their development, it is important to continue to improve the quality of MC items as they become more prominent in the assessment of student learning. While MC items remain an economical and efficient item format, they will continue to be used. It may be a challenge to continue to extrapolate more useful information from MC items, and to provide an improved role for such items in the suite of assessments, but the benefits justify the research efforts.
REFERENCES


Boud, D. (2015, April). *Is our assessment up to standard?* Symposium conducted at the University of Notre Dame. Fremantle, WA.


Modestou, M., & Gagatsis, A. (2007). Students’ improper proportional reasoning: A result of the epistemological obstacle of “linearity”. *Educational Psychology, 27*, 75–92. [http://dx.doi.org/10.1080/01443410601061462](http://dx.doi.org/10.1080/01443410601061462)


Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., … Vos, P. (2016). *Assessment in Mathematics education: Large-scale assessment and classroom assessment*. Springer Open. [http://dx.doi.org/10.1007/978-3-319-32394-7](http://dx.doi.org/10.1007/978-3-319-32394-7)


http://dx.doi.org/10.1177/0146621603254799
APPENDICES

Table of contents

2.1 Overview of automatic feedback from SMART tests ............................................. 2
2.2 The cognitive domain of the original Bloom’s taxonomy .................................. 3
4.1 Western Australian Curriculum: Mathematics - Content descriptions .................. 4
4.2 Proficiencies in the Western Australian Curriculum: Mathematics K-10 .................. 5
4.3 Key activities in a learning progression for proportional reasoning ....................... 6
7.1 Tests of ratios and fractions with CR items ....................................................... 7
7.2 Partial knowledge (partly correct response) indicated in item content .................... 13
7.3 Summary statistics for analyses with and without partial credit ............................ 14
7.4 Threshold ordering for CR items with polytomous scoring ................................ 15
7.5 Ordered estimates of item difficulty for CR items ............................................. 16
7.6 Statistics generated by comparing analyses of MC items .................................... 17
7.7 Person-item threshold distributions with and without guessing ............................ 18
7.8 Distractor curves showing informative distractors .............................................. 20
7.9 Partial knowledge of proportional reasoning in NAPLAN MC items ..................... 23
8.1 Block structure for student test ........................................................................ 25
8.2 Items for students doing the non-adapted and adapted tests ................................ 27
8.3 Partial knowledge skills in student test ............................................................. 45
8.4 Peer review of items ....................................................................................... 48
8.5 Review of items to identify partial knowledge .................................................... 48
8.6 Outcomes from cognitive interviews .................................................................. 49
8.7 Participation Information Form ....................................................................... 50
9.1 Identifying the scale for the Year 8 test ........................................................... 52
Appendix 2.1 Sample of feedback from SMART tests

Achievement of student

- Problem TYPE I: Stage 3
- Problem TYPE II: Stage 2
- Problem TYPE III: Stage 1

Mathematical focus and overview of developmental stages

Knowledge of working with percentages involves three types of percentage problems:

- TYPE I Finding PART: problems that require a part of the whole be found,
- TYPE II Finding PERCENT: problems that require that a part be expressed as a percentage of the whole,
- TYPE III Finding WHOLE: problems that require the whole to be calculated.

A separate diagnosis is given for each calculation type.

- **Stage 1** Students can solve this type of question with easy numbers and also with numbers of medium difficulty.
- **Stage 3** Students can solve this type of question with any percentages and any type of number including decimals.

Detailed explanations of developmental stages and common misconceptions

- Stage 0 students are below stage 1. These students are as yet unable to solve this type of problem.
- Stage 1 students can solve this type of problem with numbers that are simple multiples of 100.
- Stage 2 students can solve this type of problem with familiar percentages and numbers that are simple multiples of 10.
- Stage 3 students can solve this type of problem with any percentages and any type of number including decimals.

Appendix 2.2 The cognitive domain of the original Bloom’s taxonomy

(Bloom, 1956)

1.0 Knowledge
   1.10 Knowledge of specifics
       1.11 Knowledge of terminology
       1.12 Knowledge of specific facts
   1.20 Knowledge of ways and means of dealing with specifics
       1.21 Knowledge of conventions
       1.22 Knowledge of trends and sequences
       1.23 Knowledge of classifications and categories
       1.24 Knowledge of criteria
       1.25 Knowledge of methodology
   1.30 Knowledge of universals and abstractions in a field
       1.31 Knowledge of principles and generalizations
       1.32 Knowledge of theories and structures

2.0 Comprehension
   2.1 Translation
   2.2 Interpretation
   2.3 Extrapolation

3.0 Application

4.0 Analysis
   4.1 Analysis of elements
   4.2 Analysis of relationships
   4.3 Analysis of organizational principles

5.0 Synthesis
   5.1 Production of a unique communication
   5.2 Production of a plan, or proposed set of operations
   5.3 Derivation of a set of abstract relations

6.0 Evaluation
   6.1 Judgments in terms of internal evidence
   6.2 Judgments in terms of external criteria
<table>
<thead>
<tr>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Compare fractions with related denominators and locate and represent them on a number line. Solving problems involving addition and subtraction of fractions with the same or related denominators. Finding a simple fraction of a quantity where the result is a whole number, with and without digital technologies. Adding and subtracting decimals, with and without digital technologies, and using estimation and rounding to check the reasonableness of answers. Multiplying decimals by whole numbers and performing divisions by non-zero whole numbers where the results are terminating decimals, with and without digital technologies. Multiplying and dividing decimals by powers of 10. Making connections between equivalent fractions, decimals and percentages. Investigating and calculating percentage discounts of 10%, 25% and 50% on sale items, with and without digital technologies. Continuing and creating sequences involving whole numbers, fractions and decimals. Describing the rule used to create the sequence.</td>
</tr>
<tr>
<td>7</td>
<td>Comparing fractions using equivalence. Locating and representing positive and negative fractions and mixed numbers on a number line. Solving problems involving addition and subtraction of fractions, including those with unrelated denominators. Multiplying and dividing fractions and decimals using efficient written strategies and digital technologies. Expressing one quantity as a fraction of another, with and without the use of digital technologies. Rounding decimals to a specified number of decimal places. Connecting fractions, decimals and percentages and carrying out simple conversions. Finding percentages of quantities and expressing one quantity as a percentage of another, with and without digital technologies. Recognising and solving problems involving simple ratios. Investigating and calculating 'best buys', with and without digital technologies.</td>
</tr>
<tr>
<td>8</td>
<td>Carrying out the four operations with rational numbers and integers, using efficient mental and written strategies and appropriate digital technologies. Solving problems involving the use of percentages, including percentage increases and decreases, with and without digital technologies. Solving a range of problems involving rates and ratios, with and without digital technologies. Solving problems involving profit and loss, with and without digital technologies. Given coordinates, plotting points on the Cartesian plane and finding coordinates for a given point. Solving simple linear equations. Investigating, interpreting and analysing graphs from authentic data.</td>
</tr>
</tbody>
</table>
Appendix 4.2 Proficiencies in the Australian Curriculum: Mathematics K-10

The following descriptions of the proficiencies have been copied from the ACARA website: https://www.australiancurriculum.edu.au/f-10-curriculum/mathematics/key-ideas/

The proficiency strands describe the actions in which students can engage when learning and using the content. While not all proficiency strands apply to every content description, they indicate the breadth of mathematical actions that teachers can emphasise.

UNDERSTANDING

Students build a robust knowledge of adaptable and transferable mathematical concepts. They make connections between related concepts and progressively apply the familiar to develop new ideas. They develop an understanding of the relationship between the 'why' and the 'how' of mathematics. Students build understanding when they connect related ideas, when they represent concepts in different ways, when they identify commonalities and differences between aspects of content, when they describe their thinking mathematically and when they interpret mathematical information.

FLUENCY

Students develop skills in choosing appropriate procedures, carrying out procedures flexibly, accurately, efficiently and appropriately, and recalling factual knowledge and concepts readily. Students are fluent when they calculate answers efficiently, when they recognise robust ways of answering questions, when they choose appropriate methods and approximations, when they recall definitions and regularly use facts, and when they can manipulate expressions and equations to find solutions.

PROBLEM SOLVING

Students develop the ability to make choices, interpret, formulate, model and investigate problem situations, and communicate solutions effectively. Students formulate and solve problems when they use mathematics to represent unfamiliar or meaningful situations, when they design investigations and plan their approaches, when they apply their existing strategies to seek solutions, and when they verify that their answers are reasonable.

REASONING

Students develop an increasingly sophisticated capacity for logical thought and actions, such as analysing, proving, evaluating, explaining, inferring, justifying and generalising. Students are reasoning mathematically when they explain their thinking, when they deduce and justify strategies used and conclusions reached, when they adapt the known to the unknown, when they transfer learning from one context to another, when they prove that something is true or false and when they compare and contrast related ideas and explain their choices.
### Appendix 4.3  
**Key activities in a learning progression for proportional reasoning**  
(Briggs & Peck, 2015)

<table>
<thead>
<tr>
<th>Level</th>
<th>Grades</th>
<th>What students know and can do</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>K-2</td>
<td>Determine the size of one share when a collection can be shared equally e.g. sharing 20 marbles between 5 children</td>
</tr>
<tr>
<td>2</td>
<td>1-3</td>
<td>Determine the size of one share when a continuous object is shared by a number of people e.g. sharing an orange between 4 people</td>
</tr>
<tr>
<td>3</td>
<td>1-5</td>
<td>Determine the whole amount when given a single share e.g. find the size of a whole cake when you know how much each of the 10 people receive</td>
</tr>
<tr>
<td>4</td>
<td>3-5</td>
<td>Determine the effect of changing the number of people sharing a collection e.g. increasing the number of people sharing from 4 to 6 if there are 24 balloons in the collection</td>
</tr>
<tr>
<td>5</td>
<td>3-5</td>
<td>Determine the size of one share when multiple wholes are shared by a number of people such that the number of sharers is not a factor of the number of wholes e.g. sharing 3 cakes between 4 people</td>
</tr>
<tr>
<td>6</td>
<td>4-7</td>
<td>Fair sharing by equipartitioning which includes other types of proportional reasoning including comparing two ratios or finding missing values when given equivalent ratios e.g. if dogs weighing 8 kg get 20 treats each week, how many treats should the dog get each week if it weighs 12 kg</td>
</tr>
<tr>
<td>7</td>
<td>6-8</td>
<td>Solve problems in which two quantities change together when the change in one quantity is associated with a proportional change in the other, and there is a starting amount, e.g. If the taxi charge has a flag fee of $13.50 and a rate of 120c per km, determine the number of km that the taxi can go for $67.50.</td>
</tr>
</tbody>
</table>
Appendix 7.1 Tests of ratios and fractions with CR items

This test was provided by Jeremy Hodgen. Questions from this test were used in the Concepts in Secondary Mathematics and Science (CSMS) project and the ICCAMS study with King’s College London.

1. Onion Soup Recipe for 8 persons

8 onions
2 pints water
4 chicken soup cubes
2 dessertspoons butter
½ pint cream

a) I am cooking onion soup for 4 people
   How much water do I need?
   How many chicken soup cubes do I need?

b) I am cooking onion soup for 6 people
   How much water do I need?
   How many chicken soup cubes do I need?
   How much cream do I need?

2. a) There are 3 eels, A, B and C, in the tank at the zoo.

   15 cm long
   ________A

   10 cm long
   ________B

   5 cm long
   ________C

   The eels are fed sprats, the number depending on their length.
   If C is fed two sprats, how many sprats should B and A be fed to match?
   If B eats 12 sprats, how many sprats should A be fed to match?
   If A gets 9 sprats, how many sprats should B get to match?
b) Three other eels, X, Y and Z, are fed with fishfingers, the length of the fishfinger depending on the length of the eel.

`25 cm long----------Z`

`15 cm long---------Y`

`10 cm long--------X`

If X has a fishfinger 2 cm long, how long should the fishfinger given to Z be?
If Y has a fishfinger 9 cm long, how long should the fishfinger given to Z be?
If Z has a fishfinger 10 cm long, how long should the fishfingers given to X and Y be?

3. In a doctors’ surgery Dr Ahmed comes in to work 2 days a week.
Dr Brown comes in to work 4 days a week.
Dr Cartier comes in to work 6 days a week.
The bill for lighting the surgery for these doctors is 240p.
How much should each pay for it to be fair?

4. a) Finish drawing the diagram below so that it is the same shape but bigger than this diagram.

`|`  

`------------------A`

`------------------A'`
b) Work out how long the missing line should be if this diagram is to be the same shape but bigger than the one above.

5. You can see the height of Mr Short measured in triangles.

Mr Short has a friend Mr Tall.

When we measure their height with matchsticks,

Mr Short’s height is four matchsticks,

Mr Tall’s height is six matchsticks.

How many triangles are needed for Mr Tall’s height?
6. In a particular metal alloy there are
1 part mercury to 5 parts copper
3 parts tin to 10 parts copper
8 parts zinc to 15 parts copper

You would need how many parts mercury to how many parts tin?
You would need how many parts zinc to how many parts tin?

7. These two letters are the same shape. One is larger than the other.
AC is 8 units. RT is 12 units.

The curve AB is 9 units. How long is the curve RS?
The curve UV is 18 units. How long is the curve DE?

8. % means **per cent** or **per 100**, so 3% is 3 out of every 100.
   a) 4 children out of the hundred on the school trip forgot to bring their lunch.
      What percentage is this?
   b) 6% of children have free dinners. There are 250 children in the school.
      How many children have free dinners?
   c) The newspaper says that 24 out of 800 Avenger cars have a faulty engine.
      What percentage is this?
   d) The price of a coat is £20. In the sale it is reduced by 5%.
      How much does it cost now?
F2 Shade in two-thirds of this shape.

F9 Fill in the missing number in each case.

a) \( \frac{1}{3} = \frac{2}{\phantom{0}} \)

b) \( \frac{6}{8} = \frac{3}{\phantom{0}} \)

c) \( \frac{5}{10} = \frac{\phantom{0}}{30} \)

d) \( \frac{2}{3} = \frac{\phantom{0}}{15} \)

e) \( \frac{4}{12} = \frac{1}{\phantom{0}} \)

F15 I am putting tiles on the floor. The tiles are grey. What fraction of the floor has been tiled?

F16 Javed pays \( \frac{3}{5} \) of his wages in tax. What fraction of his wages is left?
He also pays \( \frac{1}{10} \) of his wages on rent. What fraction of his total wages does he have left after tax and rent have been paid?

F18 How many fractions lie between \( \frac{1}{4} \) and \( \frac{1}{2} \)?

F19 A relay race is run in stages of \( \frac{1}{8} \) km each. Each runner runs one stage. How many runners would be required to run a total distance of \( \frac{3}{4} \) km?

F21

Area = 10 square centimetres

Width = …………………
F22

Area = \( \frac{1}{3} \) square centimetre

Length = ................

F24  What fraction of the square metre is coloured grey?

End of test
## Appendix 7.2  Partial knowledge (partly correct response) indicated in item content

<table>
<thead>
<tr>
<th>Test</th>
<th>Item</th>
<th>Type</th>
<th>Item description</th>
<th>Partial knowledge (PK)</th>
<th>Type PK</th>
</tr>
</thead>
<tbody>
<tr>
<td>R.1a)_1</td>
<td>1</td>
<td>D</td>
<td>Given 2 for 8, how much for 4?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R.1a)_2</td>
<td>2</td>
<td>D</td>
<td>Given 4 for 8, how many for 4?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R.1b)_1</td>
<td>3</td>
<td>D</td>
<td>Given 2 for 8, how much for 6?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R.1b)_2</td>
<td>4</td>
<td>D</td>
<td>Given 4 for 8, how many for 6?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R.1b)_3</td>
<td>5</td>
<td>P</td>
<td>Given ½ for 8, how much for 6?</td>
<td>Fraction less than 3/8</td>
<td>RE</td>
</tr>
<tr>
<td>R.2a)_1</td>
<td>6</td>
<td>P</td>
<td>2 for 5 = ? for 10</td>
<td></td>
<td>AT</td>
</tr>
<tr>
<td>R.2a)_2</td>
<td>7</td>
<td>P</td>
<td>2 for 5 = ? for 15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R.2a)_3</td>
<td>8</td>
<td>P</td>
<td>12 for 10 = ? for 15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R.2a)_4</td>
<td>9</td>
<td>P</td>
<td>9 for 15 = ? for 10</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>R.2b)_1</td>
<td>10</td>
<td>P</td>
<td>2 for 10 = ? for 25</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>R.2b)_2</td>
<td>11</td>
<td>P</td>
<td>9 for 15 = ? for 25</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>R.2b)_3</td>
<td>12</td>
<td>P</td>
<td>10 for 25 = ? for 10</td>
<td>Positive number &lt; 4</td>
<td>RE</td>
</tr>
<tr>
<td>R.2b)_4</td>
<td>13</td>
<td>P</td>
<td>10 for 25 = ? for 15</td>
<td>7, 8, or 9</td>
<td>RE</td>
</tr>
<tr>
<td>R.3</td>
<td>14</td>
<td>P</td>
<td>Share 240 in ratio 2:4:6</td>
<td>60, 80, 100</td>
<td>PI</td>
</tr>
<tr>
<td>R.4a</td>
<td>15</td>
<td>P</td>
<td>Draw similar diagram</td>
<td>Scaled line only</td>
<td>IS</td>
</tr>
<tr>
<td>R.4b</td>
<td>16</td>
<td>P</td>
<td>Identify length - enlarged arm</td>
<td></td>
<td>AT</td>
</tr>
<tr>
<td>R.5</td>
<td>17</td>
<td>P</td>
<td>6 clips: 4 sticks, X : 6 sticks?</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>R.6_1</td>
<td>18</td>
<td>P</td>
<td>Given 3 ratios identify another</td>
<td>1:3, 2:6</td>
<td>PI</td>
</tr>
<tr>
<td>R.6_2</td>
<td>19</td>
<td>P</td>
<td>Given 3 ratios identify another</td>
<td>8:3, 16:6</td>
<td>PI</td>
</tr>
<tr>
<td>R.7_1</td>
<td>20</td>
<td>P</td>
<td>8:12 = 9:?</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>R.7_2</td>
<td>21</td>
<td>P</td>
<td>8:12 = 18 : ?</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>R.8a</td>
<td>22</td>
<td>P</td>
<td>4 out of 100 as %</td>
<td>4 out of 100, 4/100</td>
<td>IS</td>
</tr>
<tr>
<td>R.8b</td>
<td>23</td>
<td>P</td>
<td>6% of 250</td>
<td>6 or 6%</td>
<td>PI</td>
</tr>
<tr>
<td>R.8c</td>
<td>24</td>
<td>P</td>
<td>24 of 800 = 7%</td>
<td>Absolute value: 24</td>
<td>PI</td>
</tr>
<tr>
<td>R.8d</td>
<td>25</td>
<td>P</td>
<td>Reduce £20 by 5%</td>
<td>£1, £15</td>
<td>IS</td>
</tr>
<tr>
<td>F2</td>
<td>26</td>
<td>D</td>
<td>Shade two thirds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F9a</td>
<td>27</td>
<td>P</td>
<td>Fraction: 1/3 = 2/?</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>F9b</td>
<td>28</td>
<td>P</td>
<td>Fraction: 6/8 = 3/?</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>F9c</td>
<td>29</td>
<td>P</td>
<td>Fraction: 5/10 = ?/30</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>F9d</td>
<td>30</td>
<td>P</td>
<td>Fraction: 2/3 = ?/15</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>F9e</td>
<td>31</td>
<td>P</td>
<td>Fraction: 4/12 = 1/?</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>F15</td>
<td>32</td>
<td>D</td>
<td>Counting tiles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F16</td>
<td>33</td>
<td>P</td>
<td>1 - 3/5 - 1/10</td>
<td>2/5</td>
<td>IS</td>
</tr>
<tr>
<td>F18</td>
<td>34</td>
<td>P</td>
<td>Number of fractions in interval</td>
<td>Many</td>
<td>IS</td>
</tr>
<tr>
<td>F19</td>
<td>35</td>
<td>D</td>
<td>No. eighths in three quarters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F21</td>
<td>36</td>
<td>D</td>
<td>Area reverse 10 ÷ 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F22</td>
<td>37</td>
<td>P</td>
<td>Divide fractions</td>
<td>Operation given</td>
<td>IS</td>
</tr>
<tr>
<td>F24</td>
<td>38</td>
<td>P</td>
<td>Multiply fractions</td>
<td>Operation given</td>
<td>IS</td>
</tr>
</tbody>
</table>

**KEY**  
D | Dichotomous scoring (8) | Proportion ignored | PI  
P | Polytomous scoring (30) | Incomplete solution | IS  
RE | Reasonable estimate |
Appendix 7.3  Summary statistics for analyses with and without partial credit

<table>
<thead>
<tr>
<th>Statistic</th>
<th>No partial credit</th>
<th>Partial credit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WITH EXTREMES</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean item difficulty</td>
<td>0.0</td>
<td>−0.4121</td>
</tr>
<tr>
<td>Mean Fit Residual (items)</td>
<td>0.1424</td>
<td>0.1254</td>
</tr>
<tr>
<td>Mean person proficiency</td>
<td>−0.7653</td>
<td>−0.7709</td>
</tr>
<tr>
<td>Mean Fit Residual (persons)</td>
<td>−0.3545</td>
<td>−0.3035</td>
</tr>
<tr>
<td>Skewness: persons location</td>
<td>0.1035</td>
<td>−0.0636</td>
</tr>
<tr>
<td>Kurtosis: persons location</td>
<td>−0.0410</td>
<td>0.3752</td>
</tr>
<tr>
<td>PSI</td>
<td>0.87740</td>
<td>0.87821</td>
</tr>
<tr>
<td><strong>NO EXTREMES</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean item difficulty</td>
<td>0.0</td>
<td>−0.4121</td>
</tr>
<tr>
<td>Mean Fit Residual (items)</td>
<td>0.1424</td>
<td>0.1254</td>
</tr>
<tr>
<td>Mean person proficiency</td>
<td>−0.7069</td>
<td>−0.7294</td>
</tr>
<tr>
<td>Mean Fit Residual (persons)</td>
<td>−0.3545</td>
<td>−0.3035</td>
</tr>
<tr>
<td>PSI</td>
<td>0.88336</td>
<td>0.88697</td>
</tr>
<tr>
<td><strong>WITH OR WITHOUT EXTREMES</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item separation index</td>
<td>0.99936</td>
<td>0.99918</td>
</tr>
<tr>
<td>Item separation variance</td>
<td>4.66939</td>
<td>2.947</td>
</tr>
<tr>
<td>Item separation error</td>
<td>0.00298</td>
<td>0.00242</td>
</tr>
<tr>
<td>Person/Item separation index</td>
<td>0.88473</td>
<td>0.88837</td>
</tr>
<tr>
<td>Person/Item separation variance</td>
<td>15058.96</td>
<td>11056.42</td>
</tr>
<tr>
<td>Person/Item separation error</td>
<td>1735.949</td>
<td>1234.205</td>
</tr>
<tr>
<td>Power of analysis</td>
<td>Excellent</td>
<td>Excellent</td>
</tr>
</tbody>
</table>
Appendix 7.4       Threshold ordering for CR items with polytomous scoring

<table>
<thead>
<tr>
<th>Item</th>
<th>Type</th>
<th>Item description</th>
<th>Partial knowledge (PK)</th>
<th>Type PK</th>
<th>Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>P</td>
<td>Given ½ for 8, how much for 6?</td>
<td>Fraction less than 3/8</td>
<td>RE</td>
<td>ordered</td>
</tr>
<tr>
<td>6</td>
<td>P</td>
<td>2 for 5 = ? for 10</td>
<td>7</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>7</td>
<td>P</td>
<td>2 for 5 = ? for 15</td>
<td>12</td>
<td>AT</td>
<td>disordered</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>12 for 10 = ? for 15</td>
<td>17</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>9</td>
<td>P</td>
<td>9 for 15 = ? for 10</td>
<td>4</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>2 for 10 = ? for 25</td>
<td>17</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>11</td>
<td>P</td>
<td>9 for 15 = ? for 25</td>
<td>19</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>12</td>
<td>P</td>
<td>10 for 25 = ? for 10</td>
<td>Positive number &lt; 4</td>
<td>RE</td>
<td>disordered</td>
</tr>
<tr>
<td>13</td>
<td>P</td>
<td>10 for 25 = ? for 15</td>
<td>7, 8, or 9</td>
<td>RE</td>
<td>disordered</td>
</tr>
<tr>
<td>14</td>
<td>P</td>
<td>Share 240 in ratio 2:4:6</td>
<td>60, 80, 100</td>
<td>PI</td>
<td>disordered</td>
</tr>
<tr>
<td>15</td>
<td>P</td>
<td>Draw similar diagram</td>
<td>Scaled line only</td>
<td>IS</td>
<td>ordered</td>
</tr>
<tr>
<td>16</td>
<td>P</td>
<td>Identify length - enlarged arm</td>
<td>4</td>
<td>AT</td>
<td>ordered</td>
</tr>
<tr>
<td>17</td>
<td>P</td>
<td>6 clips: 4 sticks, X : 6 sticks?</td>
<td>8</td>
<td>AT</td>
<td>ordered</td>
</tr>
<tr>
<td>18</td>
<td>P</td>
<td>Given 3 ratios identify another</td>
<td>1:3, 2:3</td>
<td>PI</td>
<td>ordered</td>
</tr>
<tr>
<td>19</td>
<td>P</td>
<td>Given 3 ratios identify another</td>
<td>8:3, 16:6</td>
<td>PI</td>
<td>ordered</td>
</tr>
<tr>
<td>20</td>
<td>P</td>
<td>8:12 = 9:?</td>
<td>13</td>
<td>AT</td>
<td>ordered</td>
</tr>
<tr>
<td>21</td>
<td>P</td>
<td>8:12 = 18 : ?</td>
<td>14</td>
<td>AT</td>
<td>ordered</td>
</tr>
<tr>
<td>22</td>
<td>P</td>
<td>4 out of 100 as %</td>
<td>4 out of 100, 4/100</td>
<td>IS</td>
<td>disordered</td>
</tr>
<tr>
<td>23</td>
<td>P</td>
<td>6% of 250</td>
<td>6 or 6%</td>
<td>PI</td>
<td>disordered</td>
</tr>
<tr>
<td>24</td>
<td>P</td>
<td>24 of 800 = ?%</td>
<td>Absolute value: 24</td>
<td>PI</td>
<td>disordered</td>
</tr>
<tr>
<td>25</td>
<td>P</td>
<td>Reduce £20 by 5%</td>
<td>£1, £15</td>
<td>IS</td>
<td>ordered</td>
</tr>
<tr>
<td>27</td>
<td>P</td>
<td>Fraction: 1/3 = 2/?</td>
<td>4</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>28</td>
<td>P</td>
<td>Fraction: 6/8 = 3/?</td>
<td>5</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>29</td>
<td>P</td>
<td>Fraction: 5/10 = ?/30</td>
<td>25</td>
<td>AT</td>
<td>disordered</td>
</tr>
<tr>
<td>30</td>
<td>P</td>
<td>Fraction: 2/3 = ?/15</td>
<td>14</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>31</td>
<td>P</td>
<td>Fraction: 4/12 = 1/?</td>
<td>9</td>
<td></td>
<td>disordered</td>
</tr>
<tr>
<td>33</td>
<td>P</td>
<td>1 - 3/5 - 1/10</td>
<td>2/5</td>
<td>IS</td>
<td>ordered</td>
</tr>
<tr>
<td>34</td>
<td>P</td>
<td>Number of fractions in interval</td>
<td>Many</td>
<td>IS</td>
<td>disordered</td>
</tr>
<tr>
<td>37</td>
<td>P</td>
<td>Divide fractions</td>
<td>Operation given</td>
<td>IS</td>
<td>disordered</td>
</tr>
<tr>
<td>38</td>
<td>P</td>
<td>Multiply fractions</td>
<td>Operation given</td>
<td>IS</td>
<td>disordered</td>
</tr>
</tbody>
</table>

**KEY**  
P: Polytomous scoring (30)  
AT: Additive thinking  
RE: Reasonable estimate  
PI: Proportion ignored  
IS: Incomplete solution

• Initially ordered (2)  
• Finally ordered (8)  
• Initially disordered (20)
Appendix 7.5

<table>
<thead>
<tr>
<th>Item</th>
<th>Qn</th>
<th>Item description</th>
<th>Threshold</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>F21</td>
<td>Shade two thirds of rectangle divided into 3 parts</td>
<td>26</td>
<td>-4.28444</td>
</tr>
<tr>
<td>2</td>
<td>1a2</td>
<td>Half of 4 cubes</td>
<td>2</td>
<td>-3.38447</td>
</tr>
<tr>
<td>1</td>
<td>1a1</td>
<td>Half of 2 pints</td>
<td>1</td>
<td>-3.27237</td>
</tr>
<tr>
<td>22</td>
<td>8a1</td>
<td>What percentage is 4 out of 100?</td>
<td>22</td>
<td>-2.79395</td>
</tr>
<tr>
<td>29</td>
<td>F9c</td>
<td>Equivalent fractions: five tenths is # thirtieths</td>
<td>29</td>
<td>-2.6509</td>
</tr>
<tr>
<td>27</td>
<td>F9a</td>
<td>Equivalent fractions: one third is 2/#</td>
<td>27</td>
<td>-2.58509</td>
</tr>
<tr>
<td>28</td>
<td>F9b</td>
<td>Equivalent fractions: six eighths is 3/#</td>
<td>28</td>
<td>-2.36626</td>
</tr>
<tr>
<td>6</td>
<td>2a1</td>
<td>2 sprats for 5 cm = # for 10 cm</td>
<td>6</td>
<td>-2.09737</td>
</tr>
<tr>
<td>33</td>
<td>F16</td>
<td>Subtract 3 fifths from 1</td>
<td>33.1</td>
<td>-1.96009</td>
</tr>
<tr>
<td>3</td>
<td>1b1</td>
<td>Given 2 pints for 8, how much for 6?</td>
<td>3</td>
<td>-1.9</td>
</tr>
<tr>
<td>30</td>
<td>F9d</td>
<td>Equivalent fractions: two thirds is # fifteenths</td>
<td>30</td>
<td>-1.80738</td>
</tr>
<tr>
<td>17</td>
<td>5a1</td>
<td>AS: Adding the difference - Mr Tall</td>
<td>17.1</td>
<td>-1.71853</td>
</tr>
<tr>
<td>4</td>
<td>1b2</td>
<td>Given 4 cubes for 8, how many for 6?</td>
<td>4</td>
<td>-1.59037</td>
</tr>
<tr>
<td>7</td>
<td>2a2</td>
<td>2 sprats for 5 cm = # for 15 cm</td>
<td>7</td>
<td>-1.46749</td>
</tr>
<tr>
<td>20</td>
<td>7a1</td>
<td>AS: Adding the difference for ratio - curly K (8:12)</td>
<td>20.1</td>
<td>-1.08962</td>
</tr>
<tr>
<td>5</td>
<td>1b3</td>
<td>Identify answer is less than a half</td>
<td>5.1</td>
<td>-1.07315</td>
</tr>
<tr>
<td>31</td>
<td>F9e</td>
<td>Equivalent fractions: four twelfths is 1/#</td>
<td>31</td>
<td>-1.0621</td>
</tr>
<tr>
<td>18</td>
<td>6a1</td>
<td>Absolute values in ratios rather than proportion (1:3)</td>
<td>18.1</td>
<td>-0.91985</td>
</tr>
<tr>
<td>21</td>
<td>7a2</td>
<td>AS: Subtracting the difference for ratio - curly K (12:8)</td>
<td>21.1</td>
<td>-0.5049</td>
</tr>
<tr>
<td>16</td>
<td>4b1</td>
<td>AS: Adding the difference for scaling 2-D diagram</td>
<td>16.1</td>
<td>-0.32062</td>
</tr>
<tr>
<td>32</td>
<td>F15</td>
<td>Expressing shaded triangular tiles as fraction</td>
<td>32</td>
<td>-0.22301</td>
</tr>
<tr>
<td>19</td>
<td>6a2</td>
<td>Absolute values in ratios rather than proportion (8:3)</td>
<td>19.1</td>
<td>-0.14199</td>
</tr>
<tr>
<td>9</td>
<td>2a4</td>
<td>9 sprats for 15 cm = # for 10 cm</td>
<td>9</td>
<td>-0.004</td>
</tr>
<tr>
<td>18</td>
<td>6a1</td>
<td>Given 3 ratios with common factor identify another</td>
<td>18.2</td>
<td>0.054159</td>
</tr>
<tr>
<td>35</td>
<td>F19</td>
<td>Number of eighths in three quarters</td>
<td>35</td>
<td>0.063833</td>
</tr>
<tr>
<td>10</td>
<td>2b1</td>
<td>2 cm fish for 10 cm = # fish for 25 cm</td>
<td>10</td>
<td>0.091525</td>
</tr>
<tr>
<td>8</td>
<td>2a3</td>
<td>12 sprats for 10 cm = ? for 15</td>
<td>8</td>
<td>0.134666</td>
</tr>
<tr>
<td>25</td>
<td>8d1</td>
<td>Reduce £20 by 5%</td>
<td>25</td>
<td>0.153455</td>
</tr>
<tr>
<td>23</td>
<td>8b1</td>
<td>Determine 6% of 250</td>
<td>23</td>
<td>0.168706</td>
</tr>
<tr>
<td>24</td>
<td>8c1</td>
<td>What percentage is 24 out of 800?</td>
<td>24</td>
<td>0.199493</td>
</tr>
<tr>
<td>36</td>
<td>F21</td>
<td>Area =10, length = 4, determine width</td>
<td>36</td>
<td>0.215514</td>
</tr>
<tr>
<td>17</td>
<td>5a1</td>
<td>6 clips for 4 matchsticks is # clips for 6 matchsticks</td>
<td>17.2</td>
<td>0.527643</td>
</tr>
<tr>
<td>15</td>
<td>4a1</td>
<td>Scaled diagram - lines (not the gaps)</td>
<td>15.1</td>
<td>0.59204</td>
</tr>
<tr>
<td>14</td>
<td>3a1</td>
<td>Share 240 in the ratio of 2:4:6</td>
<td>14</td>
<td>0.60201</td>
</tr>
<tr>
<td>13</td>
<td>2b4</td>
<td>10 cm fish for 25 cm = # fish for 15 cm</td>
<td>13</td>
<td>0.607068</td>
</tr>
<tr>
<td>33</td>
<td>F16</td>
<td>Subtract one tenth from two fifths</td>
<td>33.2</td>
<td>0.755352</td>
</tr>
<tr>
<td>12</td>
<td>2b3</td>
<td>10 cm fish for 25 cm = # fish for 10 cm</td>
<td>12</td>
<td>0.761343</td>
</tr>
<tr>
<td>11</td>
<td>2b2</td>
<td>9 cm fish for 15 cm = # fish for 25 cm</td>
<td>11</td>
<td>1.001884</td>
</tr>
<tr>
<td>21</td>
<td>7a2</td>
<td># units to 18 is 8 units to 12 units</td>
<td>21.2</td>
<td>1.078432</td>
</tr>
<tr>
<td>19</td>
<td>6a2</td>
<td>Given 3 ratios with common factor identify another</td>
<td>19.2</td>
<td>1.138103</td>
</tr>
<tr>
<td>5</td>
<td>1b3</td>
<td>Given half for 8, how many for 6?</td>
<td>5.2</td>
<td>1.40645</td>
</tr>
<tr>
<td>15</td>
<td>4a1</td>
<td>Scaled diagram considering all aspects</td>
<td>15.2</td>
<td>1.439134</td>
</tr>
<tr>
<td>20</td>
<td>7a1</td>
<td>8 units to 12 units is 9 units to #</td>
<td>20.2</td>
<td>2.030208</td>
</tr>
<tr>
<td>38</td>
<td>F24</td>
<td>Interpret problem and multiply fractions</td>
<td>38</td>
<td>2.440932</td>
</tr>
<tr>
<td>16</td>
<td>4b1</td>
<td>Identify length of enlarged arm using proportions</td>
<td>16.2</td>
<td>2.775335</td>
</tr>
<tr>
<td>37</td>
<td>F22</td>
<td>Divide fractions: area / width</td>
<td>37</td>
<td>3.436628</td>
</tr>
<tr>
<td>34</td>
<td>F18</td>
<td>Identify number of fractions between two fractions</td>
<td>34</td>
<td>3.91812</td>
</tr>
</tbody>
</table>
## Appendix 7.6  Statistics generated by comparing analyses of MC items

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Year 7 2013</th>
<th>Year 9 2013</th>
<th>Year 7 2014</th>
<th>Year 9 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No partial</td>
<td>Partial</td>
<td>No partial</td>
<td>Partial</td>
</tr>
<tr>
<td>Mean item location</td>
<td>0</td>
<td>-0.116</td>
<td>0</td>
<td>-0.1258</td>
</tr>
<tr>
<td>Mean item FR</td>
<td>0.223</td>
<td>0.2747</td>
<td>0.104</td>
<td>0.2747</td>
</tr>
<tr>
<td>SD: Item location</td>
<td>1.4706</td>
<td>1.475</td>
<td>1.5008</td>
<td>1.475</td>
</tr>
<tr>
<td>Mean person proficiency</td>
<td>0.0264</td>
<td>0.0262</td>
<td>-0.0101</td>
<td>0.0167</td>
</tr>
<tr>
<td>Mean person FR</td>
<td>-0.167</td>
<td>-0.169</td>
<td>-0.1736</td>
<td>-0.1691</td>
</tr>
<tr>
<td>Mean male proficiency</td>
<td>0.147</td>
<td>0.145</td>
<td>0.102</td>
<td>0.136</td>
</tr>
<tr>
<td>Mean female proficiency</td>
<td>-0.102</td>
<td>-0.1</td>
<td>-0.129</td>
<td>-0.11</td>
</tr>
<tr>
<td>SD: persons</td>
<td>1.2501</td>
<td>1.1614</td>
<td>1.34</td>
<td>1.1614</td>
</tr>
<tr>
<td>SD: males</td>
<td>1.3</td>
<td>1.2</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>SD: females</td>
<td>1.19</td>
<td>1.1</td>
<td>1.26</td>
<td>1.1</td>
</tr>
<tr>
<td>skewness persons</td>
<td>0.6915</td>
<td>0.8798</td>
<td>0.3308</td>
<td>0.8798</td>
</tr>
<tr>
<td>kurtosis persons</td>
<td>0.6624</td>
<td>1.072</td>
<td>0.7445</td>
<td>1.072</td>
</tr>
<tr>
<td>PSI</td>
<td>0.9289</td>
<td>0.9297</td>
<td>0.9265</td>
<td>0.9297</td>
</tr>
<tr>
<td>Maximum Information function &amp;</td>
<td>11.5</td>
<td>13.99</td>
<td>11.34</td>
<td>14</td>
</tr>
<tr>
<td>Power of analysis</td>
<td>Excellent</td>
<td>Excellent</td>
<td>Excellent</td>
<td>Excellent</td>
</tr>
</tbody>
</table>
Appendix 7.7  Person-item threshold distributions: Guessed responses included

Year 7 2013 No partial credit

Year 9 2013 No partial credit

Year 7 2013 Partial credit

Year 9 2013 Partial credit

Year 7 2014 No partial credit

Year 9 2014 No partial credit

Year 7 2014 Partial credit

Year 9 2014 Partial credit
Appendix 7.7 (cont’d)  Person-item threshold distributions: Guessed responses excluded

Year 7 2013 No partial credit (no guessing)

Year 9 2013 No partial credit (no guessing)

Year 7 2013 Partial credit (no guessing)

Year 9 2013 Partial credit (no guessing)

Year 7 2014 No partial credit (no guessing)

Year 9 2014 No partial credit (no guessing)

Year 7 2014 Partial credit (no guessing)

Year 9 2014 Partial credit (no guessing)
Appendix 7.8  Distractor curves showing informative distractors

Item 55: Threshold difference = 4.017

Item 24: Threshold difference = 2.514

Item 24: Threshold difference = 1.9805

Item 23: Threshold difference = 1.9187

Item 23: Threshold difference = 1.9162

Item 24: Threshold difference = 1.8686
Appendix 7.8 (cont’d)  Distractor curves showing informative distractors

Item 54: Threshold difference = 1.475

Item 6: Threshold difference = 1.2657

Item 52: Threshold difference = 1.1923

Item 51: Threshold difference = 1.1001

Item 57: Threshold difference = 0.7572

Item 54: Threshold difference = 0.6304
Appendix 7.8 (cont’d)  
Distractor curves showing informative distractors

Item 51: Threshold difference = 0.61

Item 18: Threshold difference = 0.4919

Item 50: Threshold difference = 0.4251

Item 47: Threshold difference = 0.4127

Item 16: Threshold difference = 0.3343

Item 17: Threshold difference = 0.3095
**Appendix 7.9 Partial knowledge of proportional reasoning in NAPLAN MC items**

<table>
<thead>
<tr>
<th>Threshold difference</th>
<th>Test</th>
<th>Item</th>
<th>Skill for partial knowledge</th>
<th>Type</th>
<th>AC: content</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Y7</td>
<td>23</td>
<td>adds absolute value rather than proportion</td>
<td>AT</td>
<td>linear</td>
</tr>
<tr>
<td>High</td>
<td>Y7</td>
<td>52</td>
<td>1200 m is 200 cm (cf. 20 cm) for scale 1:6000</td>
<td>RE</td>
<td>ratio</td>
</tr>
<tr>
<td>High</td>
<td>Y7</td>
<td>53</td>
<td>adds numerators, denominators to add fraction</td>
<td>PI</td>
<td>fractions</td>
</tr>
<tr>
<td>High</td>
<td>Y9</td>
<td>6</td>
<td>5% growth is by 5 cm</td>
<td>PI</td>
<td>percentages</td>
</tr>
<tr>
<td>High</td>
<td>Y9</td>
<td>24</td>
<td>used rate of change but not initial value</td>
<td>IS</td>
<td>linear</td>
</tr>
<tr>
<td>High</td>
<td>Y7</td>
<td>24</td>
<td>added denominators to add fraction</td>
<td>RE</td>
<td>fractions</td>
</tr>
<tr>
<td>High</td>
<td>Y7</td>
<td>55</td>
<td>fraction halfway between 1/3 and 3/5 is 2/4</td>
<td>PI</td>
<td>fractions</td>
</tr>
<tr>
<td>High</td>
<td>Y9</td>
<td>23</td>
<td>added denominators to add fraction</td>
<td>RE</td>
<td>fractions</td>
</tr>
<tr>
<td>High</td>
<td>Y9</td>
<td>24</td>
<td>applies linear scale factor to area</td>
<td>PI</td>
<td>ratios</td>
</tr>
<tr>
<td>High</td>
<td>Y9</td>
<td>54</td>
<td>applies first proportion but not second</td>
<td>IS</td>
<td>decimals</td>
</tr>
<tr>
<td>Medium</td>
<td>Y7</td>
<td>11</td>
<td>correct operation (division) but numbers reversed</td>
<td>IM</td>
<td>formula</td>
</tr>
<tr>
<td>Medium</td>
<td>Y7</td>
<td>38</td>
<td>selects most black balloons cf. highest proportion</td>
<td>PI</td>
<td>chance</td>
</tr>
<tr>
<td>Medium</td>
<td>Y7</td>
<td>20</td>
<td>represents 4 cm with scale 1:20000 as 80 m (cf. 800 m)</td>
<td>IS</td>
<td>ratio</td>
</tr>
<tr>
<td>Medium</td>
<td>Y9</td>
<td>51</td>
<td>counts number of sections cf. identifying area</td>
<td>PI</td>
<td>fractions</td>
</tr>
<tr>
<td>Medium</td>
<td>Y9</td>
<td>57</td>
<td>selects total cf. proportion as the denominator</td>
<td>PI</td>
<td>fractions</td>
</tr>
</tbody>
</table>

**KEY:**

<table>
<thead>
<tr>
<th>AT</th>
<th>Additive thinking</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>Proportion ignored</td>
</tr>
<tr>
<td>IS</td>
<td>Incomplete solution</td>
</tr>
<tr>
<td>RE</td>
<td>Reasonable estimate</td>
</tr>
<tr>
<td>IM</td>
<td>Incorrect method</td>
</tr>
</tbody>
</table>
Appendix 7.9 (cont’d)

Partial knowledge of proportional reasoning in NAPLAN MC items

<table>
<thead>
<tr>
<th>Threshold difference</th>
<th>Test</th>
<th>Item</th>
<th>Skill for partial knowledge</th>
<th>Type</th>
<th>AC: content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Y7 2013</td>
<td>48</td>
<td>estimate third of 23,958 as 7000 rather than 8000</td>
<td>RE</td>
<td>rounding</td>
</tr>
<tr>
<td>Low</td>
<td>Y7 2013</td>
<td>50</td>
<td>350 million ÷ 79 million ~ 4 rather than 4</td>
<td>RE</td>
<td>division</td>
</tr>
<tr>
<td>Low</td>
<td>Y7 2013</td>
<td>14</td>
<td>number line: from $\frac{2}{3}$ to $\frac{1}{3}$ is $\frac{1}{3}$</td>
<td>IS</td>
<td>fractions</td>
</tr>
<tr>
<td>Low</td>
<td>Y9 2013</td>
<td>17</td>
<td>adds absolute value rather than proportion</td>
<td>AT</td>
<td>linear</td>
</tr>
<tr>
<td>Low</td>
<td>Y9 2013</td>
<td>18</td>
<td>uses absolute value cf. proportion to identify percent</td>
<td>PI</td>
<td>percentages</td>
</tr>
<tr>
<td>Low</td>
<td>Y9 2013</td>
<td>19</td>
<td>identifies 40 out of 200 as 0.4</td>
<td>PI</td>
<td>decimals</td>
</tr>
<tr>
<td>Low</td>
<td>Y9 2013</td>
<td>53</td>
<td>applies difference on graph to convert miles to km</td>
<td>AT</td>
<td>graph</td>
</tr>
<tr>
<td>Low</td>
<td>Y7 2014</td>
<td>51</td>
<td>identifies only one of the correct proportions</td>
<td>IS</td>
<td>graph</td>
</tr>
<tr>
<td>Low</td>
<td>Y7 2014</td>
<td>54</td>
<td>counts number of sections cf. identifying area</td>
<td>PI</td>
<td>fractions</td>
</tr>
<tr>
<td>Low</td>
<td>Y7 2014</td>
<td>19</td>
<td>reading a scale as 0.1 units cf. 0.2 as a unit</td>
<td>PI</td>
<td>decimals</td>
</tr>
<tr>
<td>Low</td>
<td>Y7 2014</td>
<td>47</td>
<td>add decimals: 0.3 + 0.99 = 1.02</td>
<td>IM</td>
<td>decimals</td>
</tr>
<tr>
<td>Low</td>
<td>Y9 2014</td>
<td>16</td>
<td>reading a scale as 0.1 units cf. 0.2 as a unit</td>
<td>PI</td>
<td>decimals</td>
</tr>
<tr>
<td>Low</td>
<td>Y9 2014</td>
<td>17</td>
<td>represents 4 cm with scale 1:20000 as 80 m (cf. 800 m)</td>
<td>IS</td>
<td>ratio</td>
</tr>
<tr>
<td>Low</td>
<td>Y9 2014</td>
<td>43</td>
<td>add decimals: 0.3 + 0.99 = 1.02</td>
<td>IM</td>
<td>decimals</td>
</tr>
</tbody>
</table>

**KEY:**
- **AT** Additive thinking
- **PI** Proportion ignored
- **IS** Incomplete solution
- **RE** Reasonable estimate
- **IM** Incorrect method
### Appendix 8.1  Block structure for student test (Blocks 1–3)

<table>
<thead>
<tr>
<th>Block</th>
<th>Item</th>
<th>Content</th>
<th>Year</th>
<th>Level</th>
<th>Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>multiplicative thinking</td>
<td>8</td>
<td>low</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>fractions</td>
<td></td>
<td>low</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>multiplicative thinking</td>
<td></td>
<td>low</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>rates</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>linear relations</td>
<td></td>
<td>low</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>percentages</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>percentages</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>spatial reasoning</td>
<td></td>
<td>low</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>percentages</td>
<td></td>
<td>low</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>ratios</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>percentages</td>
<td>8</td>
<td>medium</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>percentages</td>
<td></td>
<td>medium</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>multiplicative thinking</td>
<td></td>
<td>medium</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>linear relations</td>
<td></td>
<td>medium</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>spatial reasoning</td>
<td></td>
<td>medium</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>rates</td>
<td></td>
<td>medium</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>percentages</td>
<td></td>
<td>medium</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>fractions</td>
<td></td>
<td>medium</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>multiplicative thinking</td>
<td></td>
<td>medium</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>ratios</td>
<td></td>
<td>medium</td>
<td>problem solving</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>linear relations</td>
<td>8</td>
<td>high</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>multiplicative thinking</td>
<td></td>
<td>high</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>rates</td>
<td></td>
<td>high</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>percentages</td>
<td></td>
<td>high</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>fractions</td>
<td></td>
<td>high</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>percentages</td>
<td></td>
<td>high</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>spatial reasoning</td>
<td></td>
<td>high</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>ratios</td>
<td></td>
<td>high</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>rates</td>
<td></td>
<td>high</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>multiplicative thinking</td>
<td></td>
<td>high</td>
<td>reasoning</td>
</tr>
</tbody>
</table>
## Appendix 8.1 (cont’d)  Block structure for student test (Blocks 4–6)

<table>
<thead>
<tr>
<th>Block</th>
<th>Item</th>
<th>Content</th>
<th>Year</th>
<th>Level</th>
<th>Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>31</td>
<td>spatial reasoning</td>
<td>7</td>
<td>low</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>fractions</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>ratios</td>
<td></td>
<td>low</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>fractions</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>ratios</td>
<td></td>
<td>low</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>percentages</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>percentages</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>multiplicative thinking</td>
<td></td>
<td>low</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td>multiplicative thinking</td>
<td></td>
<td>low</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>rates</td>
<td></td>
<td>low</td>
<td>reasoning</td>
</tr>
<tr>
<td>5</td>
<td>41</td>
<td>decimals</td>
<td>7</td>
<td>medium</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>multiplicative thinking</td>
<td></td>
<td>medium</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>spatial reasoning</td>
<td></td>
<td>medium</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>percentages</td>
<td></td>
<td>medium</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>ratios</td>
<td></td>
<td>medium</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>fractions</td>
<td></td>
<td>high</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>47</td>
<td>fractions</td>
<td></td>
<td>high</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>multiplicative thinking</td>
<td></td>
<td>medium</td>
<td>reasoning</td>
</tr>
<tr>
<td></td>
<td>49</td>
<td>rates</td>
<td></td>
<td>high</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>percentages</td>
<td></td>
<td>medium</td>
<td>understanding</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
<td>fractions</td>
<td>6</td>
<td>low</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>percentages</td>
<td></td>
<td>low</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>53</td>
<td>fractions</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>54</td>
<td>rates</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>percentages</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>56</td>
<td>fractions</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>57</td>
<td>number</td>
<td></td>
<td>low</td>
<td>fluency</td>
</tr>
<tr>
<td></td>
<td>58</td>
<td>percentages</td>
<td></td>
<td>low</td>
<td>understanding</td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>fractions</td>
<td></td>
<td>low</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>decimals</td>
<td></td>
<td>low</td>
<td>problem solving</td>
</tr>
</tbody>
</table>
Appendix 8.2 Items for students doing the non-adapted and adapted tests

(Item number) ** key * hypothesised informative distractor

MULTIPLE CHOICE ITEMS – BLOCK 1

1. (1) If the number of people sharing the cost of building a Cat Refuge were to quadruple (multiply by 4), then the amount of money that each person needs to give will

a. ** reduce to a quarter of the original amount
b. reduce to a half of the original amount
c. * increase to four times the original amount
d. increase to double the original amount

2. (2) These two circles have the same area. Three quarters of the first circle is pink.

What fraction of the yellow area (whole second circle) is \( \frac{1}{6} \) (one sixth) of the pink area?

a. * \( \frac{1}{6} \)
b. \( \frac{1}{4} \)
c. ** \( \frac{1}{8} \)
d. \( \frac{1}{24} \)

3. (3) The table shows the carbohydrate content of 100 g of bread and falafel.

<table>
<thead>
<tr>
<th>Food</th>
<th>Carbohydrate (g / 100 g of food)</th>
</tr>
</thead>
<tbody>
<tr>
<td>falafel</td>
<td>25</td>
</tr>
<tr>
<td>bread</td>
<td>40</td>
</tr>
</tbody>
</table>

If you eat 100 g of falafel and 200 g of bread then you take in

a. * 65 g of carbohydrate
b. 75 g of carbohydrate
c. 95 g of carbohydrate
d. ** 105 g of carbohydrate
4. (4)
Milly drives her delivery truck from the farm to the depot at an average of 100 km per hour. This journey takes 90 minutes.

What is the distance from the farm to the depot?

a. * 90 km  
b. 100 km  
c. ** 150 km  
d. 175 km

5. (5)
The table shows the number of kilograms of oranges and their respective costs.

<table>
<thead>
<tr>
<th>number of kilograms</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>cost in dollars</td>
<td>$2.50</td>
<td>$5.00</td>
<td>$7.50</td>
<td>$10.00</td>
<td>$12.50</td>
</tr>
</tbody>
</table>

The relationship between the cost and the number of kilograms is

a. ** cost = number of kilograms \times$2.50  
b. cost = number of kilograms \div$2.50  
c. number of kilograms = cost +$2.50  
d. * add$2.50 to the cost

6. (6)
Which of the following expressions does NOT calculate 65% of $450?

a.   65 \times 450 \div 100  
b. ** 65 \times 100 \div 450  
c. * 0.65 \times 450  
d. 65 \div 100 \times 450

7. (7)
Oil production was forecast to be 1.2 million barrels per day.

Instead, it reached 1.6 million barrels per day.

This increase in what was forecast is closest to

a. 4%  
b. * 25%  
c. ** 30%  
d. 40%
8. (8)
For the diagram provided
\[
\frac{\text{radius of small circle}}{\text{radius of large circle}} = \frac{1}{3}
\]
The radius of the large circle is 60 cm.
What is the radius of the small circle?

a. 10 cm  
b.** 20 cm  
c.* 30 cm  
d. 40 cm

9. (9)
The correct answer in a student’s homework was $744.
The question could have been

a.** Increase $600 by 24%  
b.* Increase $700 by 44%  
c. Decrease $700 by $44  
d. Decrease $800 by $166

10. (10)
Read the following problem and study the solution shown.
For every 3 kg of plastic recycled, there are 5 kg of cardboard recycled.
What weight of recycled plastic would you expect for 8 kg of recycled cardboard?
Four lines of working are shown.
What is the first line on which an error is made?

\[
\frac{3}{5} = \frac{x}{8} \quad 1.
\]
\[
\frac{3}{8} = \frac{x}{5} \quad 2.
\]
\[
24 = 5x \quad 3.
\]
\[
x = \frac{5}{24} \quad 4.
\]

a. 1  
b.** 2  
c.* 3  
d. 4
MULTIPLE CHOICE ITEMS – BLOCK 2

1. (11)
The percentage of carbohydrate in a packet of 9 oatcakes is 72%.
If you share these oatcakes evenly between 3 people, then the percentage of carbohydrate in each person’s share of biscuits is

a. 8%
b. 9%
c. * 24%
d. ** 72%

2. (12)
Frida started with $400 in the bank. She wanted to have a total of $800 in the bank.
In the first month she added 25% more to her starting amount.
What percentage of the total was still needed?

a. * 75%
b. ** 37.5%
c. 50%
d. 25%

3. (13)
There are 1500 kilojoules in 100 g of drink mix.
How many kilojoules are there in 40 g of this drink mix?

a. ** 600
b. * 900
c. 15
d. 60

4. (14) Which of these graphs represents the highest cost per kilogram?

a. * A
b. B
c. ** C
d. D
5. (15)

Two square gym floors need to be polished.

The time estimate for the larger floor is 8 hours.

If the floors are polished at the same rate, then the time needed for the smaller floor is

a. 16 hours
b. 8 hours
c.* 4 hours
d.** 2 hours

6. (16)
Returning home Jim drove for a quarter of an hour at an average speed of 40 km per hour.

After that, he drove for three quarters of an hour at an average speed of 80 km per hour.

For the total journey, Jim’s average speed was

a. 40 km per hour
b.* 60 km per hour
c.** 70 km per hour
d. 80 km per hour

7. (17)

For her two Maths tests, Annie scored \( \frac{7}{10} \) (70%) and \( \frac{12}{15} \) (80%).

For the combined test scores, Annie’s percentage achievement is

a.* 75%
b.** 76%
c. 78%
d. 80%
The area of this rectangular paddock (not drawn to scale) is \( \frac{1}{3} \text{ km}^2 \).

The length of one side is \( \frac{3}{5} \) km.

To determine the length of the other side (km), the correct calculation is

a. \( \frac{1}{3} + \frac{3}{5} \)

b. \( \frac{1}{3} - \frac{3}{5} \)

c. * \( \frac{1}{3} \times \frac{3}{5} \)

d. ** \( \frac{1}{3} \div \frac{3}{5} \)

9. (19)
An 800 g packet of Quavers is priced at $2.40.
Jack divides 2.40 by 8.
His answer is 0.3.
This value of 0.3 represents

a. The cost in cents for 100 g of Quavers

b. ** The cost in dollars for 100 g of Quavers

c. * The mass in grams that you can buy for $1

d. The mass in grams that you can buy for 1c

10. (20)
In a fruitcake recipe the mass of some dried fruits occur in the following ratios:

\[
\text{cherries to sultanas} = 1:6 \\
\text{raisins to sultanas} = 2:9
\]

The ratio of the mass of cherries to the mass of raisins is

a. 3:15

b. * 1:2

c. ** 3:4

d. 2:3
MULTIPLE CHOICE ITEMS – BLOCK 3

1. (21)
The graphs represent the journeys of 4 different vehicles A, B, C and D.

![Graphs](image)

It is TRUE to say that, over the 5-hour period
a.* Vehicle A is the only vehicle that travels as far as 300 km
b. Vehicle B has not gone as far as 200 km after 5 hours
c. Vehicle C goes further than any other vehicle
d.** Vehicle D goes faster than any other vehicle

2. (22)
For the cost of 1 burger I can buy 2 loaves of bread.
For the cost of 4 loaves of bread I can buy 6 cans of spaghetti.
The best estimate for the number of cans of spaghetti that I can buy for the same cost as 10 burgers is:
a. 10
b. 20
c.* 25
d.** 30

3. (23)
A toy car travels 60 cm in 2.4 seconds.
One student divided 2.4 by 60 and wrote the answer as 0.04.
Select the correct statement for this situation.
a.** The car takes 0.04 seconds to go 1 cm.
b. Every minute the car goes 0.04 cm.
c. The student made an error because 2.4 ÷ 60 = 25.
d.* The speed of the car is 0.04 cm per second.

4. (24)
After a pair of shoes is discounted by 12%, the price is $35.20.
The original price before the discount was given can be calculated by
a. $35.20 \times 88\%
b.* $35.20 \times 112\%
c.** $35.20 ÷ 88\%
d. $35.20 ÷ 112\%
5. (25)
AULD bars of chocolate were reduced in size by \( \frac{1}{5} \) of their mass and the bars formed were called NOO bars.
To increase the NOO bar back to the mass of the AULD bar then the fraction of the mass of the NOO bar that it must be increased by is

a. \( \frac{1}{5} \)

b. \( \frac{1}{4} \)

c. \( \frac{4}{5} \)

d. \( \frac{5}{4} \)

6. (26)
Kara scored 60 marks out of 100 in her first maths test.
In her second maths test her mark was 20% higher than the first mark.
Kara’s third test mark was 25% higher than the second test mark.
Her percentage increase from the first test to the third test was

a. over 100%

b. ** 50%

c. * 45%

d. 22.5%

7. (27)

Area of any triangle = half the base x perpendicular height.
Consider the effect on the area of right-angled triangles when the base and height are changed.
If the base is multiplied by \( a \) and the height is multiplied by \( b \) then

a. * new area = original area \( \times (a + b) \)

b. ** new area = original area \( \times (a \times b) \)

c. new area = original area \( \div (a + b) \)

d. new area = original area \( \div (a \times b) \)
8. (28)  
The relationships between three variables are as follows:  
\[ a = 8b \]
\[ a = \frac{c}{5} \]

The ratio \( b : c \) is  
a. \( 5 : 8 \)  
b. \( 8 : 5 \)  
c.* \( 40 : 1 \)  
d.** \( 1 : 40 \)

9. (29)  
Jack cycled to Poppy’s house and walked back along the same route.  
His cycling speed was 9 km per hour and his walking speed was 3 km per hour.  
If his travel time was a total of 8 hours, what distance did he travel?  
a. 96 km  
b.* 48 km  
c.** 36 km  
d. 24 km

10. (30)  
Shane took 4 hours to mow all the grass tennis courts and for the same job Col took 12 hours.  
It is true to say that  
a. Col works three times as fast as Shane.  
b.** If they started at the same time and worked together, it would take 3 hours to mow all the courts.  
c. Shane can mow as many courts in 3 hours as Col can do in 1 hour.  
d.* It would take Col 14 hours to mow as many courts as Shane can do in 6 hours.
MULTIPLE CHOICE ITEMS – BLOCK 4

1. (31)
Jane’s drawing of an insect is 10 cm long.
Her drawing is 15 times the insect’s actual length.
The correct calculation to determine the length of the insect (in centimetres) is

a. ** 10 ÷ 15
b. * 15 ÷ 10
c. 10 x 15
d. 15 - 10

2. (32)
The fraction \( \frac{48}{64} \) in its simplest form is

a. * \( \frac{24}{32} \)
b. \( \frac{4}{6} \)
c. \( \frac{2}{3} \)
d. ** \( \frac{3}{4} \)

3. (33)
A rug is made from 2 yellow, 4 red, 5 black and 9 blue balls of wool.

It is INCORRECT to say that;
a. the ratio of yellow to red to black to blue is 2 : 4 : 5 : 9
b. * the ratio of blue to black can also be written as 18 : 10
c. ** the ratio of yellow to red is 2 : 1
d. the ratio of black to yellow is 5 : 2
4. (34)
A batch of scones requires one quarter ($\frac{1}{4}$) of a cup of sugar.
I have 8 cups of sugar.

How many batches of scones can I make?

a.** 32
b. 8
c. $8 \frac{1}{4}$
d.* 2

5. (35)
In a packet of tortillas there are 23 g of fat and 63 g of carbohydrate per 100 g of tortillas.

The ratio of fat to carbohydrate in the packet is closest to;

a.** 1 : 3
b.* 3 : 1
c. 1 : 5
d. 3 : 5

6. (36)
Tim used 600 kL of water in January.

He used 30% more water in February than he did in January.

The increase in water usage was

a. 630 kL
b.** 180 kL
c. 100 kL
d.* 30 kL

7. (37)
Rosa received a discount of $40 on a laptop priced at $800.

What was her percentage discount?

a.* 40%
b. 20%
c. 8%
d.** 5%
8. (38)
Daniel has two dogs; Benson who weighs 10 kg and Shamrock who weighs 15 kg.
Daniel gives them treats according to the ratio of their weights.
If Daniel gives Benson 12 treats, how many should he give to Shamrock?
   a. 24
   b. 18
   c. 17
   d. 12

9. (39)
To make 4 L of GREEN paint, 3 L of yellow paint are mixed with 1 L of blue paint.
To make 36 L of GREEN paint, the number of litres of blue paint needed is
   a. 12
   b. 9
   c. 3
   d. 1

10. (40)
I can buy ice cream in two different tub sizes.
    2 Litres for $2.20
    4 Litres for $4.30
Select the correct statement which explains why the 4 L tub is better value for money.
   a. The cost per litre of ice cream in the 2 L tub is $1.10 and the cost per litre of ice cream
      in the 4 L tub is $1.15.
   b. The cost of two 2 L tubs is less than the cost of a 4 L tub.
   c. The 4 L tub is larger and larger containers give better value for money.
   d. The cost per litre of ice cream in the 4 L tub is less than $1.10 which is the cost per
      litre of ice cream in the 2 L tub.
MULTIPLE CHOICE ITEMS – BLOCK 5

1. (41)
The diagram below shows the heights of two trees in 2011 and 2015.

The diagram is not drawn to scale.

<table>
<thead>
<tr>
<th></th>
<th>Jan’s tree</th>
<th>Michael’s tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>Height = 2 m</td>
<td>Height = 3 m</td>
</tr>
<tr>
<td>2015</td>
<td>Height = 5 m</td>
<td>Height = 6 m</td>
</tr>
</tbody>
</table>

The true statement is

a. ** The height of Jan’s tree in 2011 was 0.4 times its height in 2015.
b. Jan’s tree gained more in height than Michael’s tree.
c. The height of Michael’s tree in 2011 was 0.3 times its height in 2015.
d. * Michael’s tree in 2015 was 0.5 times its height in 2011.

2. (42)
For every 6 banana muffins sold in Vera’s shop, 15 apple muffins are sold.
When 26 banana muffins are sold, how many apple muffins would be sold?

a. 35 
    b. 47 
    c. * 62 
    d. ** 65

3. (43)

Phil has two boxes which are both rectangular prisms.
The larger box is twice as high, three times as wide and twice as long as the smaller box.
How many times greater is the capacity of the larger box?

a. 24 
    b. ** 12 
    c. 14 
    d. * 7
4. (44)
In March, Luke was 135 cm tall.
Between March and October, Luke grew 10 cm in height.
In October, Luke’s height was
a. 90% of his height in March.
b. more than 90% but not more than 100% of his height in March.
c.** at least 100% but not as much as 110% of his height in March.
d.* at least 110% of his height in March.

5. (45)
Nikki and her mother celebrate their birthdays on the same day.
When her mother turns 40 this year, the ratio of Nikki’s age to her mother’s age will be 3:10.
After 8 more years the ratio of Nikki’s age to her mother’s age will be
a. 1:2
b. 3:10
c.* 11:18
d.** 20:48

6. (46)
Jenni donated \(\frac{5}{12}\) (five-twelfths) of her savings to charity and Peter donated \(\frac{3}{8}\) (three-eighths) of his savings to charity.
Identify the true statement.

a. Jenni donated more than Peter because 5 is larger than 3.
b. Jenni donated less than Peter because her fraction is less than a half.
c.* You need to make equivalent fractions before you can form any conclusion.
d.** You need to know the savings of each person before you can form any conclusion.

7. (47)
Consider the number line shown below.

What is the value of \(x\) on the number line?

a.** \(\frac{1}{4}\)
b. \(-\frac{1}{5}\)
c.* \(\frac{1}{8}\)
d. \(\frac{1}{9}\)
8. (48)

Patrick, Nicola, Louise and Alix had all saved up different amounts of money to spend at the Show. The graph shows the amount of money each person spent and the amount of money each person had left over.

The statements below describe the person who spent the highest proportion of the money they had saved.

Select the correct statement.

a. ** Patrick: he has more blue than red in his bar.
   b. Nicola: she has more red than blue in her bar.
   c. * Louise: her blue section reaches further to the right than all the others.
   d. Alix: her red section is shorter than all other red sections.

9. (49)

The table shows four different sized packets of GUBY bars and their costs.

Which packet has the **highest** cost per 100 g?

<table>
<thead>
<tr>
<th>Packet</th>
<th>Number of bars in each packet</th>
<th>Size of each bar</th>
<th>Total mass of bars in the packet</th>
<th>Cost of packet</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>100 g</td>
<td>100 g</td>
<td>$1.10</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>50 g</td>
<td>200 g</td>
<td>$2.40</td>
</tr>
<tr>
<td>C</td>
<td>12</td>
<td>25 g</td>
<td>300 g</td>
<td>$3.70</td>
</tr>
<tr>
<td>D</td>
<td>24</td>
<td>25 g</td>
<td>600 g</td>
<td>$6.80</td>
</tr>
</tbody>
</table>

a. * A
   b. B
   c. ** C
   d. D

10. (50)

Drew received a discount of 8% on the $60 cost of a shirt.

Which of the following discounts is a greater amount of money than Drew’s discount?

a. * 60% discount on an item costing $8
   b. 8% discount on an item costing $50
   c. ** 7% discount on an item costing $70
   d. 6% discount on an item costing $60
MULTIPLE CHOICE ITEMS – BLOCK 6

1. (51)
In which two diagrams is the same fraction of the diagram shaded?

a. B and D  
b. A and D  
c. C and E  
d. C and B

2. (52)
Tina is copying a movie onto her computer.
The darker shading on the bar shows how much of a movie has been copied.

For Tina, 85% of her movie has been copied.

Which bar shows the amount of Tina’s movie that has been copied?

a. 1  
b. 2  
c. 3  
d. 4

3. (53)
Tam and Lou each have a ruler.

Tam’s ruler is one third ($\frac{1}{3}$) of the length of Lou’s ruler.

If Lou’s ruler is 300 mm in length, then the length of Tam’s ruler is

a. 100 mm  
b. 300 mm  
c. 600 mm  
d. 900 mm
4. (54)
If 8 kg of pears cost $48, what will 10 kg of pears cost?

a. $38.40  
b. ** $60  
c. * $80  
d. $148

5. (55)
It costs Jala $80 to make each birthday cake.
Jala determines the price to sell her cakes by adding 10% to her costs.
Jala sells each cake for

a. $10  
b. $80  
c. ** $88  
d. * $90

6. (56)
Jon’s pancake recipe requires \( \frac{3}{4} \) cups of flour.
How much flour will Jon need when he doubles the recipe?

a. ** 3 \( \frac{1}{2} \) cups  
b. 3 cups  
c. * 2 \( \frac{6}{8} \) cups  
d. 2 \( \frac{1}{2} \) cups

7. (57)
Marc bought 18 team shirts which cost $25 each.
Which of the following expressions can be used to calculate the total cost?

a. 20 \( \times \) 18 \( \times \) 5 \( \times \) 18  
b. * 20 \( + \) 5 \( \times \) 8 \( + \) 10  
c. 20 \( \times \) 10 \( + \) 5 \( \times \) 8  
d. ** 20 \( \times \) 18 \( + \) 5 \( \times \) 18
8. (58)
Carla rolled a die 50 times and made a table of her results. The number showing on the top face is in the first column. The number of times this number showed on the top face is in the second column.

<table>
<thead>
<tr>
<th>Number showing on the top face</th>
<th>Number of times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>50</strong></td>
</tr>
</tbody>
</table>

Select the TRUE statement.

a. Exactly 50% of the rolls resulted in a 3 showing on the top face.

b. * Exactly 10% of the rolls resulted in a 1 showing on the top face.

c. Exactly 100% of the rolls resulted in a number less than 5 showing on the top face.

d. ** Exactly 80% of the rolls resulted in a number greater than 1 showing on the top face.

9. (59)
Abe’s art lesson has been going for one hour. The lesson is four-fifths (\(\frac{4}{5}\)) completed.

How much time is left in the lesson?

a. * 12 minutes

b. ** 15 minutes

c. 48 minutes

d. 60 minutes

10. (60)
Four students entered a goal-throwing competition. The table shows the number of successes and the number of attempts.

<table>
<thead>
<tr>
<th>Player</th>
<th>Number of successes</th>
<th>Number of attempts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matt</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Nick</td>
<td>35</td>
<td>50</td>
</tr>
<tr>
<td>Alice</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>Kat</td>
<td>75</td>
<td>90</td>
</tr>
</tbody>
</table>

For which student is the proportion of success equal to 0.75?

a. ** Matt

b. Nick

c. Alice

d. * Kat
Appendix 8.3  Partial knowledge skills in student test

= Partial credit successful [AT: additive thinking, IM: incorrect method, IS: incomplete solution, PI: proportion ignored, RE: reasonable estimate]

<table>
<thead>
<tr>
<th>Block</th>
<th>Item</th>
<th>Skill for full credit</th>
<th>Skill / misconception awarded partial credit</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Increasing the number of people that share a set cost results in a smaller contribution needed per person. (Inverse proportion)</td>
<td>Quadrupling is linked to increasing to four times the original amount.</td>
<td>IM</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>One sixth of three quarters of a circle is one eighth of a whole circle.</td>
<td>One sixth of three quarters of a circle is one sixth of the circle.</td>
<td>PI</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>There is a total of 105 g of carbohydrate in 100 g of falafel and 200 g of bread when the rates are 25 g/100 g and 40 g/100 g respectively.</td>
<td>Components are added without consideration of the proportion.</td>
<td>PI</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>Distance travelled in 90 minutes at 100 km per hour is 150 km.</td>
<td>Distance is calculated using 100 km per hour as 1 km per minute.</td>
<td>RE</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>Rule to determine cost (from a table of values) is number of kilograms multiplied by the cost for 1 kg.</td>
<td>Identification of recursion from a table of values</td>
<td>IS</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>65 x 100 ÷ 450 is not a calculation to determine 65% of $450.</td>
<td>identifies multiplication by 450, but not 65 ÷ 100 as 0.65</td>
<td>IS</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>Increasing from 1.2 million to 1.6 million is about a 30% increase.</td>
<td>Percentage increase is calculated relative to the final amount.</td>
<td>RE</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>If A/B = 1/3 and B = 60, then A = 20.</td>
<td>Identification of 30 cm as one third of 60 (calculation error)</td>
<td>RE</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>$744 is the result of increasing $600 by 24%.</td>
<td>Identifies $44 as 44% increase</td>
<td>PI</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>3:5 = x:8 does not mean that 3/5 = x/8</td>
<td>Interprets the context to form a correct ratio</td>
<td>IS</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>Percentage of components unchanged when amount of product changes.</td>
<td>Sharing 72% between 3 people results in 24% per person.</td>
<td>AT</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>Adding 25% to a starting value of $400 means that 37.5% of the total of $800 is still needed to get to this total.</td>
<td>Recognises 75% as the complement of 25%</td>
<td>PI</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>If there are 1500 kj in 100 g then there are 600 kj in 40 g.</td>
<td>Given a proportion, the relative size of the component is identified.</td>
<td>RE</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>Highest cost per kilogram is represented by the steepest set of points.</td>
<td>The highest cost per kilogram is represented by the highest point.</td>
<td>RE</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>Doubling the length of the side of a square, results in an area which is four times the size of the original area.</td>
<td>Doubling the length of the side of a square, results in an area which is double the size of the original area.</td>
<td>PI</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>Speed is total distance travelled divided by total time taken</td>
<td>Calculates the average of two speeds</td>
<td>PI</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>Calculates overall percentage using total of parts out of possible total</td>
<td>Calculates the average of two percentages</td>
<td>PI</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>Given area and length of rectangle (fractions) determines missing length</td>
<td>Recognises the formula for area of a rectangle</td>
<td>IM</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>Dividing cost, given in dollars, by the number of multiples of 100 g of the mass gives cost, in dollars, per 100 g</td>
<td>Recognises when dividing cost by mass the answer must be a cost</td>
<td>IM</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>Uses common multiple for A:C given A:B = 1:6 and C:B = 2:9</td>
<td>Compares two ratios in terms of the first members of each pair.</td>
<td>PI</td>
</tr>
<tr>
<td>Block Item</td>
<td>Skill for full credit</td>
<td>Skill / misconception awarded partial credit</td>
<td>Type PK</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------</td>
<td>---------------------------------------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Speed is linked to the steepness of a line in a time-distance graph.</td>
<td>Reads point on a line in the context of the question</td>
<td>IS</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Two different proportions require a common multiple for comparison.</td>
<td>Uses the incorrect ratio to solve the problem</td>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Dividing time taken by the number of units of distance travelled gives time taken to travel one unit of distance.</td>
<td>Applies familiar problem algorithm to unfamiliar problem</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Recognises division of discounted price by new percentage is needed to determine original price</td>
<td>Recognises that multiplication by percentages greater than 100% will increase the value of a number</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Increasing a bar by one quarter of its mass (not given) is needed to return to the original mass after it reduction by one-fifth.</td>
<td>Applies same proportional decrease to return to original amount</td>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Multiplies percentages to determine successive percentage increases</td>
<td>Adds percentages to determine successive percentage increases</td>
<td>RE</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Scaling a triangle: area changes by a factor determined by multiplying the scaling factors for height and base</td>
<td>Recognises that scaling by integers increases the area of a triangle</td>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>If (a = 8b) and (a = c/5) then (b:c = 1:40)</td>
<td>Determines and misinterprets relationship between variables</td>
<td>RE</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Travelling different speeds during an 8-hour period requires proportional distribution of time</td>
<td>Averages speeds to determine distance when given time taken</td>
<td>PI</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>If Shane takes 4 hours to do a task and Col takes 12 hours to do the same task, then the task is done in 3 hours if working together</td>
<td>Applies additive thinking to a proportional relationship</td>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Identifies division by 15 is needed to determine the length</td>
<td>Identifies division is needed to determine the length</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>The simplest form of (48/64) is (3/4)</td>
<td>Partly simplifies (48/64) to (24/32)</td>
<td>IS</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Recognise that 2:9 is the same as 18:10</td>
<td>Recognise ratios using 1:1 correspondence with models of objects</td>
<td>IS</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Recognises there are 32 quarters in 8</td>
<td>Uses multiplication instead of division ((8 \times 1/4))</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>23:63 is closest to 1:3</td>
<td>Recognise that 63 is closest to 3 times 23 (numbers reversed)</td>
<td>RE</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>An increase of 30% on 600 kL is 180</td>
<td>Calculates an increase of 30% using 100 kL as the base</td>
<td>PI</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>A discount of $40 on $800 is 5%</td>
<td>Uses $100 as the base to calculate 40% of $800</td>
<td>PI</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Sharing treats according to weights means that 10:15 = 12:18</td>
<td>Uses additive (add on) thinking when applying a ratio in context.</td>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Mixing 1 L blue with 3 L yellow to make a total of 4 L green requires 9 L blue for a total of 36 L green</td>
<td>Applied the ratio to calculate the amount of yellow paint needed rather than the amount of blue paint</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Buying 4 litres for $4.30 is less than $1.10 per litre</td>
<td>Uses common contexts in which the larger the amount purchased, the lower is the unit cost</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>Block Item</td>
<td>Skill for full credit</td>
<td>Skill / misconception theoretically worthy of partial credit</td>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PK</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>2 m is 0.4 of 5 m</td>
<td>Recognises that 3 m is 0.5 of 6 m</td>
<td>RE</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>A ratio of 6:15 is equivalent to 26:65</td>
<td>A ratio of 6:15 is equivalent to 26:62 (x4, add 2: additive thinking)</td>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>Increasing the width, height and length of a box results in a volume which is increased by a product of the scale factors</td>
<td>Increasing the width, height and length of a box results in a volume which is increased by the sum of the scale factors</td>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>Increasing 135 cm by 10 cm --&gt; between 100% and 110% of 135 cm</td>
<td>Interprets 10 cm increase in relation to 1 m as a 10% increase</td>
<td>AT</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>Considers original values before adding time elapsed to each term</td>
<td>Adds the increase over time to each term of the ratio.</td>
<td>PI</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Identifies whole amount needed for fraction comparison</td>
<td>Knows comparison of fractions requires common denominators</td>
<td>IS</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>Determining the size of the partitions on a number line requires consideration of the number of partitions</td>
<td>Uses given fraction denominator to identify scale on a number line</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>Links greater proportions of parts to relative size</td>
<td>Links greater proportions of parts to absolute size (ignoring whole)</td>
<td>PI</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>Divides total cost by number of 100 g units to determine cost per 100 g</td>
<td>Uses well-known algorithm for calculating &quot;best buys&quot;</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>7% of $70 is greater than 8% of $60.</td>
<td>Recognises that 60% is greater than 6%, 7% and 8%</td>
<td>IS</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>Considers size as well as number of partitions of a model</td>
<td>Counts number of partitions in a fraction model</td>
<td>PI</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>85% of a bar is over three-quarters of the bar</td>
<td>Recognises that 85% is greater than half-way</td>
<td>RE</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>Dividing by 3 is needed to determine one third</td>
<td>Associates thirds with multiplication by 3</td>
<td>IM</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>10 kg will cost $60 if 8 kg cost $48</td>
<td>Determines total cost from (incorrect) unit cost</td>
<td>RE</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>Adding 10% to $80 gives $88</td>
<td>Determines 10% as $10, relative to a base of $100 cf $80</td>
<td>PI</td>
<td></td>
</tr>
</tbody>
</table>
| 56         | \[
2 \times \frac{3}{4} - \frac{1}{2} = \frac{6}{8}
\]                                                                                                      | \[
2 \times \frac{3}{4} = \frac{6}{8}
\]                                                                                                                                                                                                                  | IS   |
| 57         | Identifies use of associative law and order of operations                                                                                                  | Identifies associative law only (incorrect order of operations)                                                                                                                                                                                               | IS   |
| 58         | 40 out of 50 is exactly 80%                                                                                                                                  | Identifies number rather than proportion for the percentage                                                                                                                                                                                                      | PI   |
| 59         | If one hour is four fifths, then one fifth is 15 minutes                                                                                                   | Calculates one fifth of one hour                                                                                                                                                                                                                              | RE   |
| 60         | 30 out of 40 is 0.75                                                                                                                                   | Recognises 75 out of 90 as 0.75                                                                                                                                                                                                                              | PI   |
## Appendix 8.4 Peer review of items

Each block of 10 items was reviewed by five teachers: 50 is the maximum possible score for each rating feature.

<table>
<thead>
<tr>
<th>Feature for rating</th>
<th>BLOCKS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>The item can be mapped to the curriculum for that level.</td>
<td>47</td>
</tr>
<tr>
<td>The mathematical expectation of the item is clear</td>
<td>48</td>
</tr>
<tr>
<td>The language used is appropriate for the item and the students.</td>
<td>48</td>
</tr>
<tr>
<td>The item is presented clearly for the students.</td>
<td>48</td>
</tr>
<tr>
<td>The item is <strong>not far too</strong> difficult for students in Year 8.</td>
<td>47</td>
</tr>
</tbody>
</table>

## Appendix 8.5 Review of items to identify partial knowledge

For each item in each block, the distractors worthy of partial credit are identified.

- A: Distractor designed by the **author** to be awarded partial credit
- B: Distractor identified by the **first reviewer** as worth partial credit
- C: Distractor identified by the **second reviewer** as worth partial credit

Note: x indicates that the reviewer could not identify a single distractor to be awarded partial credit.

<table>
<thead>
<tr>
<th>BLOCK 1 Items</th>
<th>BLOCK 2 Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>a</td>
</tr>
<tr>
<td>B</td>
<td>x</td>
</tr>
<tr>
<td>C</td>
<td>b</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BLOCK 3 Items</th>
<th>BLOCK 4 Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>a</td>
</tr>
<tr>
<td>B</td>
<td>a</td>
</tr>
<tr>
<td>C</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BLOCK 5 Items</th>
<th>BLOCK 6 Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>d</td>
</tr>
<tr>
<td>B</td>
<td>d</td>
</tr>
<tr>
<td>C</td>
<td>c</td>
</tr>
</tbody>
</table>

**KEY:**
- Some agreement with author
- Both agreeing with author
Appendix 8.6  Outcomes from cognitive interviews

<table>
<thead>
<tr>
<th>Block of items</th>
<th>1:8L</th>
<th>2:8M</th>
<th>3:8H</th>
<th>4:7L</th>
<th>5:7MH</th>
<th>6:6L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of times reviewed</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Actions of 10 students during cognitive interviews

<table>
<thead>
<tr>
<th>No.</th>
<th>Action</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Read the item aloud accurately</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Paraphrase the item correctly</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Said they could understand the wording</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Could describe item content</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>Said they knew what action was required for the item</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>Described what they thought the item required</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Said they could eliminate one of the options?</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>Identified an incorrect option</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>Said the item was an easy one to understand</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>Said it was easy to work out the correct option for this item</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Said the item did not need changing</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>Said item was easier than the previous one</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>Describe what could be done to make this item easier for you.</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>Knew mathematical terms indicated</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>First thought about item was related to it mathematically</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>Thinking along the mathematical lines needed to respond</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>Identified a correct option.</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>18</td>
<td>Provided a sensible explanation for choosing an option</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>Could explain how the first option differed from the last one</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>Suggested further knowledge needed to identify correct response</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>21</td>
<td>Thought the test might have been a bit hard for Year 8 students</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>22</td>
<td>Identified ways by which test could be improved for students</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
Dear Student,

I am a professor in the Faculty of Education at The University of Western Australia. I am writing to invite you to take part in a research project that is being conducted by Ms Joan Burfitt as part of her Doctor of Philosophy here at UWA. The aim of the research is to improve the way multiple-choice items collect information about student learning.

I am asking for your help with this project because the content of the research is about proportional reasoning which is an important topic for Year 8 students to understand. This request is being extended to several hundred other students in Year 8 in Western Australia, and your school is one of several schools in Western Australia invited to participate.

Your involvement

If you agree to take part you will be asked to complete an online test of the skills necessary to develop sound proportional reasoning, for example questions on fractions, decimals, percentages and ratios. The online test will be conducted in November 2016 and it will consist of 30 multiple-choice items based on the Mathematics curriculum for Years 6-8. It is anticipated that it should be comfortably finished within 40 minutes. You will be asked select your gender, whether you are male or female, and your selections for the thirty multiple-choice items. Calculators will not be permitted.

An optional practice test will be made available to your teacher early Term 4, 2016 and this test will provide an opportunity to practice doing multiple choice items for mathematics in the online testing environment. The data from this practice test will not be analysed.

Giving consent

You are free to choose to participate or not to participate. I will respect your decision whichever choice you make, and I will not question it. Participating in this research will not affect your grades, your relationship with your teacher(s), or with your school. If you say no, but then change your mind and want to take part, contact your teacher who will let you know if you can still join in. If you say yes, but then want to stop participating, then let your teacher know and you can withdraw at any time before the completion of the test.

Your data

The data will be stored securely on the University server and on the researcher’s personal computer; it can only be accessed by the research team. All data storage will be in accordance with the University’s guidelines.

After all the information for the project has been collected and analysed, the findings will be published in a research journal and also will be shared with other teachers of mathematics. When this occurs, your name and the name of your school will not be mentioned or recorded. After the project is complete the data will be destroyed. You can request a copy of the project findings by emailing the research student at joan.burfitt@research.uwa.edu.au. The summary should be available by November, 2018.
Please note that if you write something while you are answering the survey questions and I need to report what you have written because the law requires me to do so, then I will have to do so.

**Approvals**

Approval to conduct this research has been provided by The University of Western Australia, in accordance with its ethics review and approval procedures. This research has also met the policy requirements of the Department of Education.

Please talk about the project with your parents first. Then, if you would like to talk with me or one of the research team more, please contact me by phone on 6488 2394 or by email at peter.merrotsy@uwa.edu.au. If at any time you wish to speak with a person who is not involved in the project about how something was handled, please contact the Human Ethics Office at the University of Western Australia on (08) 6488 3703 or by emailing to humanethics@uwa.edu.au.

**Providing consent**

If you have had all questions about the project answered to your satisfaction, and you have discussed the project with your parents and you want to be part of this research please complete the *Consent to Participate in a Research Project: Year 8 Students* form attached to this letter. It should be returned by Date.

This information letter is for you to keep.

Yours sincerely,

Professor Peter Merrotsy

Date

Approval to conduct this research has been provided by the University of Western Australia, in accordance with its ethics review and approval procedures. Any person considering participation in this research project, or agreeing to participate, may raise any questions or issues with the researchers at any time.

In addition, any person not satisfied with the response of researchers may raise ethics issues or concerns, and may make any complaints about this research project by contacting the Human Ethics Office at the University of Western Australia on (08) 6488 3703 or by emailing to humanethics@uwa.edu.au.

All research participants are entitled to retain a copy of any Participant Information Form and/or Participant Consent Form relating to this research project.
Appendix 9.1  Identifying the scale for the Year 8 test

This section describes in greater detail, the process for identifying the final scale of estimates for the item parameters. These estimates can be used to create a scale on which all persons can be located, regardless of which test design they had been offered.

To create a scale on which to place all persons I need

(a) Estimates of item difficulty for all 42 dichotomous items
(b) Parameter estimates for the 17 partial–credit (PC) items
   (i) second threshold values
   (ii) threshold distance
   (iii) first threshold value

Summary of actions taken and reasons for the actions

<table>
<thead>
<tr>
<th>Actions</th>
<th>Reasons</th>
</tr>
</thead>
</table>
| (a)     | These estimates from the tailored non-adaptive design were less biased than those from the non-tailored non-adaptive and the adaptive designs:  
- Guessing was reduced  
- Scale not as stretched as adaptive  
- Threshold ordering differed between the non-adaptive and adaptive designs, so it was inappropriate to average these two sets of estimates  
- In the adaptive design, there were very few responses to some items, so the estimates were not considered to be valid |
| (b) (i)  | These values represented the estimate of difficulty of getting a score of 1 in the tailored non-adaptive and were considered equivalent to getting a score of 2 if the item was scored for partial credit.  
They are from the same analysis as the estimates for (a); they were considered very comparable |
| (b) (ii) | 1. Could not use (a) because all estimates were dichotomous.  
2. When PC was scored, the estimates for the dichotomous items changed and in trying to keep the estimates from (a) for the dichotomous items it was necessary to investigate other possibilities for identifying the first threshold. |
Located an analysis of similar origin to the one used for the estimates of the dichotomous items and the second threshold of the PC items, that is (a). Calculated the threshold distance using $T_2 - T_1$.

3. With anchoring the estimates of the dichotomous items changed, so it was a matter of identifying a comparable analysis of the same origin.*

4. The threshold distance had been generated by RUMM from an analysis similar to the one from which the dichotomous items were determined i.e. non-adaptive, tailored and not anchored.

5. The threshold distance was not altered when the responses were tailored

<table>
<thead>
<tr>
<th>(b)</th>
<th>(iii)</th>
<th>Standard calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Threshold 2 – Threshold distance</td>
<td></td>
</tr>
</tbody>
</table>

* To determine threshold values for the PC items, I tried anchoring the dichotomous items and giving partial credit. This was problematic with both types of anchoring. With mean-item anchoring, the range of difficulties was reduced and the estimates for the dichotomous items changed considerably. With individual item anchoring, previously ordered threshold became disordered.