

https://dspace.org/

Web-based diagnosis of misconceptions in rational numbers

Layton, Roger David

2016

http://hdl.handle.net/10539/20782

Layton, Roger David (2016) Web-based diagnosis of misconceptions in rational numbers, University of Witwatersrand, Johannesburg, <http://wiredspace.wits.ac.za/handle/10539/20782>

Downloaded from a DSpace repository. https://dspace.org/

Copyright Notice

The copyright of this thesis vests in the University of Witwatersrand, Johannesburg, in accordance with the University's Intellectual Property Policy.

No portion of the thesis may be reproduced in any form or by any means, including analogue and digital media, without prior written permission from the University. Extracts or quotations from this thesis may, however, be made in terms of Sections 12 and 13 of the South African Copyright Act No. 98 of 1978 (as amended), for non-commercial or educational purposes. Full acknowledgement of this thesis is available on the Library Web page (www.wits.ac.za/library) under "Research Resources".

For permission requests, please contact the University Legal Officer or the University Research Office (www.wits.ac.za).

WEB-BASED DIAGNOSIS OF MISCONCEPTIONS IN RATIONAL NUMBERS

Roger David Layton

A thesis submitted to the Wits School of Education, Faculty of Humanities, University of the Witwatersrand in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, 2016

Abstract and Keywords

This study explores the potential for Web-based diagnostic assessments in the classroom, with specific focus on certain common challenges experienced by learners in the development of their rational number knowledge. Two schools were used in this study, both having adequate facilities for this study, comprising a well-equipped computer room with one-computer-per-learner and a fast, reliable broadband connection.

Prior research on misconceptions in the rational numbers has been surveyed to identify a small set of problem types with proven effectiveness in eliciting evidence of misconceptions in learners. In addition to the problem types found from prior studies, other problem types have been included to examine how the approach can be extended. For each problem type a small item bank was created and these items were presented to the learners in test batteries of between four and ten questions. A multiple-choice format was used, with distractor choices included to elicit misconceptions, including those previously reported in prior research. The test batteries were presented in dedicated lessons to learners over four consecutive weeks to Grade 7 (school one) and Grade 8 (school two) classes from the participating schools. A number of test batteries were presented in each weekly session and, following the learners' completion of each battery, feedback was provided to the learner with notes to help them reflect on their performance.

The focus of this study has been on diagnosis alone, rather than remediation, with the intention of building a base for producing valid evidence of the fine-grained thinking of learners. This evidence can serve a variety of purposes, most significantly to inform the teacher on each learners' stage of development in the specific *micro-domains*. Each micro-domain is a fine-grained area of knowledge that is the basis for lesson-sized teaching and learning, and which is highly suited to diagnostic assessment.

A fine-grained theory of constructivist learning is introduced for positioning learners at a development stage in each micro-domain. This theory of development stages is the foundation I have used to explore the role of diagnostic assessment as it may be used in future classroom activity. To achieve successful implementation into time-constrained mathematics classrooms requires that diagnostic assessments are conducted as effectively and efficiently as possible. To meet this requirement, the following elements of diagnostic assessments were investigated: (1) Why are some questions better than others for diagnostic purposes? (2) How many questions need to be asked to produce valid conclusions? (3) To what extent is learner self-knowledge of item difficulty useful to identify learner thinking?

A Rasch modeling approach was used for analyzing the data, and this was applied in a novel way by measuring the construct of the learners' propensity to select a distractor for a misconception, as distinct from the common application of Rasch to measure learner ability. To accommodate multiple possible misconceptions used by a learner, parallel Rasch analyses were performed to determine the likely causes of learner mistakes. These analyses were used to then identify which questions appeared to be better for diagnosis.

The results produced clear evidence that some questions are far better diagnostic discriminators than others for specific misconceptions, but failed to identify the detailed rules which govern this behavior, with the conclusion that to determine these would require a far larger research population. The results also determined that the number of such good diagnostic questions needed is often surprisingly low, and in some cases a single question and response is sufficient to infer learner thinking. The results show promise for a future in which Web-based diagnostic assessments are a daily part of classroom practice. However, there appears to be no

additional benefit in gathering subjective self-knowledge from the learners, over using the objective test item results alone.

Keywords: diagnostic assessment; rational numbers; common fractions; decimal numbers; decimal fractions; misconceptions; Rasch models; World-Wide Web; Web-based assessment; computer-based assessments; formative assessment; development stages; learning trajectories.

Declaration

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Roger Layton
Student Number: 8176055
Ethics Protocol Number: 2007ECE108

Signed this 7th day of June 2016

To Nina. For your total support.

Publications and Presentations Emanating from this Research

Layton, R. (2005). On the automation of diagnostic assessment of mathematical skill. *1st Africa Regional Congress of the International Commission on Mathematical Instruction (ICMI)*. University of the Witwatersrand, Johannesburg, Gauteng, 22-25 June 2005.

Layton, R. (2008). A web-based approach to diagnostic assessment of mathematics misconceptions. *AMESA Conference*. Nelson Mandela Metropolitan University, Port Elizabeth, Eastern Cape, 30 June – 4 May 2008.

Layton, R. (2010). Fine-grained cognitive diagnostic assessment of rational number misconceptions using the World-Wide-Web. *SAARMSTE Conference 2011*. University of KwaZulu-Natal, Durban, KwaZulu-Natal. 18-21 January 2010.

Layton, R. (2011). Using Revised Bloom and other lenses to model systematic errors. *SAARMSTE Conference 2012*, University of the North-West, Mafikeng, North-West Province, South Africa. 18-21 January 2011.

Acknowledgements

I would like to acknowledge and thank my supervisor, Prof. Margot Berger, for her continual guidance and feedback throughout the period of my study and writing, including those times when external factors delayed my progress. Also, thanks to Dr Delia Layton for reviewing and editing.

I acknowledge the entire staff and post-graduate students of the Marang Centre for Maths and Science education for their constructive engagements during my presentations at the PhD Weekends, where I gained valuable feedback on my work which helped to direct my efforts.

Special thanks to the National Research Foundation, for making funds available to assist me during the course of this study.

Finally, I would like to acknowledge my wife, Nina Layton, for her patience in tolerating my turning our entire home into my research library.

Table of Contents – High-Level

CHAPTER 1 : INTRODUCTION 1
CHAPTER 2 : LITERATURE SURVEY 29
CHAPTER 3 : THEORETICAL MODEL 74
CHAPTER 4 : RESEARCH DESIGN AND METHODOLOGY 98
CHAPTER 5 : DATA ANALYSIS AND RESULTS - PRETEST 141
CHAPTER 6 : DATA ANALYSIS - ONLINE DATA 172
CHAPTER 7 : RESEARCH RESULTS AND FINDINGS 252
CHAPTER 8 : CONCLUSIONS AND FUTURES 271
REFERENCES 286
APPENDIX A : RESULTS OF DATA GATHERING 295
APPENDIX B : PRETEST 298
APPENDIX C : WEB-BASED IMPLEMENTATION 305
APPENDIX D : THE ITEM BANK FOR ONLINE TESTS 310
APPENDIX E : ONLINE LESSON STRUCTURE 311
APPENDIX F : RASCH ANALYSIS USING WINSTEPS 320

Table of Contents – Low-Level

CHAPTER 1 : INTRODUCTION 1
1.1 Background to this Study 2
1.2 Outlining South African School Mathematics 5
1.3 South Africa and the TIMSS Studies 7
1.4 Diagnostic Assessment for Learning 13
1.5 The Right Questions 15
1.6 The Role of Self-Knowledge 16
1.7 Diagnostic Assessment 17
1.8 From Scores to Measures 18
1.9 Automating Diagnosis 22
1.10 Research Problem 26
1.11 Research Questions 26
1.12 Outline of Remainder of Thesis 28
CHAPTER 2 : LITERATURE SURVEY 29
2.1 Introduction 29
2.2 Constructivism in Learning and Teaching 30
2.3 Misconceptions 36
2.4 Issues concerning Rational Number Misconceptions 38
2.5 Misconceptions by Type of Rational Number 44
2.6 Educational Assessment and Measurement 53
2.7 Construct Validity 68
2.8 Domains and Learning Trajectories 69
2.9 Web-Based Assessments 70
2.10 Learner Self-Knowledge 72
2.11 Conclusions from the Literature Survey 73
CHAPTER 3 : THEORETICAL MODEL 74
3.1 Introduction 74
3.2 Modeling Diagnostic Assessment 75
3.3 A Constructivist Theory of Teaching and Learning 79
3.4 An Approach to Assessment 81
3.5 From Constructivism to Theory 83
3.6 A Fine-Grained Theory of Learning 85
3.7 Individual Learning Trajectories and the Points A and B 86
3.8 Micro-Domains 88
3.9 Item Difficulty Reconceptualized 89
3.10 Zones and Development Stages 92
3.11 Self-Assessment 96
3.12 Commentary on the Theoretical Framework 96
CHAPTER 4 : RESEARCH DESIGN AND METHODOLOGY 98
4.1 Introduction 98
4.2 Population 99
4.3 Sample 100
4.4 Method 101
4.5 Tools for Data Collection 103
4.6 Self-Knowledge of Item Difficulty 124

4.7 Instrumentation 126
4.8 Piloting of the Instruments 129
4.9 Planning Meetings with the Schools 131
4.10 Administering the Tests 131
4.11 Data Collection 134
4.12 Data Capturing 134
4.13 Errors and Limitations in the Research Data Gathering 135
4.14 Rasch Analysis 135
4.15 General Approach to Data Analysis 138
CHAPTER 5 : DATA ANALYSIS AND RESULTS - PRETEST 141
5.1 Pretest Results and Question-Level Analysis 141
5.2 Summary of the Pretest Results 169
CHAPTER 6 : DATA ANALYSIS - ONLINE DATA 172
6.1 Introduction 172
6.2 The Data 172
6.3 Micro-Domains Summary 175
6.4 Micro-Domain PV - Place-Value Knowledge 177
6.5 Micro-Domain DO - Decimal Number Ordering 196
6.6 Micro-Domain CR - Common Fraction Representation 214
6.7 Micro-Domain NL - Number Line for Common Fractions 219
6.8 Micro-Domain CG - Common Fraction Graphics 224
6.9 Micro-Domain CO - Common Fraction Ordering 227
6.10 Micro-Domain CE - Common Fraction Estimation 236
6.11 Micro-Domain CA - Common Fraction Addition 245
6.12 Conclusions 250
CHAPTER 7 : RESEARCH RESULTS AND FINDINGS 252
7.1 Discussion of Results by Research Question 252
7.2 Key Findings from this Study 262
CHAPTER 8 : CONCLUSIONS AND FUTURES 271
8.1 Contributions of this Study 271
8.2 Relating Findings to Theory and Literature 275
8.3 Surprising Results 278
8.4 The Future: Implementation and Research 279
8.5 The Prospects for Web-Based Diagnostic Assessment in the Classroom 282
8.6 Final Words 285
REFERENCES 286
APPENDIX A : RESULTS OF DATA GATHERING 295
APPENDIX B : PRETEST 298
APPENDIX C : WEB-BASED IMPLEMENTATION 305
APPENDIX D : THE ITEM BANK FOR ONLINE TESTS 310
APPENDIX E : ONLINE LESSON STRUCTURE 311
Structure of the Individual Online Tests 318
APPENDIX F : RASCH ANALYSIS USING WINSTEPS 320

List of Figures

Figure 1. Fine-Grained Model of Proficiency and Learning 85
Figure 2. PV1 – Place-value sample test item - select digit at named position 106
Figure 3. PV2 - Place value sample test item - select position from digit 107
Figure 4. Decimal ordering sample test item: two choices 109
Figure 5. Decimal ordering sample test item with five choices 110
Figure 6. Common fraction word representation sample test item 112
Figure 7. Fraction representation sample test item 113
Figure 8. Number line sample test item 115
Figure 9. Graphical fraction sample test item 117
Figure 10. Common fraction ordering sample test item 118
Figure 11. Sample CO test item 119
Figure 12. Sample CO test item 119
Figure 13. Fraction estimation sample: CE1 121
Figure 14. Fraction estimation sample: CE2 121
Figure 15. Fraction addition sample test item 123
Figure 16. Pretest Question 1 - compare common fractions 142
Figure 17. Pretest Question 2 - common fraction estimation 145
Figure 18. Pretest Question 3 - decimal fraction place-value 146
Figure 19. Pretest Question 4 – decimal fraction ordering 148
Figure 20. Pretest Question 5 - decimal number representation 150
Figure 21. Pretest Question 6 - common fractions to percentages 152
Figure 22. Pretest Question 7 – equivalent fractions 153
Figure 23. Pretest Question 8 - results 155
Figure 24. Pretest Question 10b – inserting a decimal point 158
Figure 25. Pretest Question 11 - decimal addition estimation 160
Figure 26. Item 10084 173
Figure 27. Place-Value : Learners by Stage 196
Figure 28. Item 10029 197
Figure 29. Item 10049 198
Figure 30. Decimal Ordering : Learners by Stage 213
Figure 31. Number line example test item 219
Figure 32. Common Fraction Ordering : Learners by Stage 236
Figure 33. Item 10114 237
Figure 34. Item 10112 237
Figure 35. Item 10122 237
Figure 36. Item 10115 238
Figure 37. Common Fraction Estimation CE1 : Learners by stage 239
Figure 38. Item 10126 241
Figure 39. Item 10128 244
Figure 40. Common Fraction Estimation CE2 : Learners by Stage 245
Figure 41. Item 10158 249
Figure 42. Common Fraction Addition : Learners by Stage 249
Figure 43. Example of Web mathematics requirement 307
Figure 44. Example of markup of a test item for the Assessment Markup Language 309
Figure 45. Example of the Information Pages during Web-based assessment 311
Figure 46. Sample Results Page in the Web-based assessment 313

List of Tables

Table 1: Items with Good Diagnostic Value 90
Table 2. Learner Development Stages in micro-domains 95
Table 3. Misconceptions in the PV micro-domain 108
Table 4. Steinle's (2004a) Classification of decimal ways of thinking 111
Table 5. Misconceptions in the CR micro-domain 114
Table 6. Misconceptions in the NL micro-domain 116
Table 7. Misconceptions in the DO micro-domain 120
Table 8. Misconceptions in the CE micro-domain 122
Table 9. Misconceptions in the CA micro-domain 123
Table 10. Decision matrix on learner confidence vs. response type 125
Table 11. Pretest Question 1 - difficulty levels 143
Table 12. Pretest Question 3 – difficulty levels 147
Table 13. Pretest Question 5 – difficulty levels 150
Table 14. Pretest Question 9 – decimal subtraction 156
Table 15. Pretest Question 9 – selected difficulty levels 157
Table 16. Pretest Question 12 – common fraction equivalence 161
Table 17. Pretest Question 13 – common fraction density 162
Table 18. Pretest Question 14 – common fraction density 163
Table 19. Pretest Question 15 – decimal ordering results 166
Table 20. Count of response records by micro-domain 174
Table 21. Place-Value: Counts by test item/response 178
Table 22. Place-Value: Test item correlation 181
Table 23. Place-Value: Item correlations after iteration 4 182
Table 24. Place-Value: Learner measures for ability 183
Table 25. Place-Value: WHOLE misconception correlation 186
Table 26. Place-Value: WHOLE misconception iteration 4 187
Table 27. Place-Value: WHOLE misconception 188
Table 28. Place-Value: EMERGENT+ABSENT learners 191
Table 29. Decimal Ordering: Results of Rasch analysis 203
Table 30. Decimal Ordering: Learner measures 207
Table 31. Common Fraction Representation: Counts by test item/response 215
Table 32. Number Line: Counts by test items/response 220
Table 33. Fraction Diagram: Counts by test item/response 225
Table 34. Common Fraction: Test items by difficulty and duration 228
Table 35. Common Fraction Comparison: Map of learners and items 231
Table 36. Common Fraction Ordering: NUMERATOR misconception 232
Table 37. Common Fraction Ordering: DENOMINATOR misconception results 233
Table 38. Common Fraction Estimation: CE2 Counts by test item/response 243
Table 39. Common Fraction Addition: Counts by test item/response 246
Table 40. Common Fraction Addition: ADDITION misconception results 248
Table 41. Learner A67 Difficulty Index Analysis 261
Table 42. Database fields for online test responses 295
Table 43. Summary of Item Bank by Item Type 310
Table 44. Test Items used by Test 318
Table 45. Example of Control File for Rasch analysis 322

Glossary of Terms and Abbreviations

This glossary covers the words, terms, and abbreviations that have special meaning in the context of this thesis. Some are derived from prior work and others are introduced in the context of this study. Some words are used for elements of the Development Stage model, introduced in Chapter 3, and these words are written with capital letters to highlight their particular meaning in the context of this study—for example ABSENT.

ABSENT	The development stage of a learner in a micro-domain in which the learner has insufficient knowledge to make sense of any question posed, and where guessing is the only possible response. This is defined as "the learner *does not know* this micro-domain".
abundant number	A whole number for which the sum of its factors is larger than its value. For example 12 for which the factors 1, 2, 3, 4, and 6 sum to 16.
ACTIVE	The development stage of a learner in a micro-domain in which there is active learning taking place as evidenced by the learner's use of known misconceptions. This is defined as "the learner is *getting to know* this micro-domain".
AI	Artificial Intelligence.
AMESA	Association of Mathematics Educators of South Africa.
assessment	The process of discovering information about the state of knowledge of a learner in a specific domain. There are various forms of assessment which are used for different purposes.
assessment *for* learning	A process which includes continuous assessment to provide evidence of learners' progress towards proficiency, to inform instructional practices. This term is often used interchangeably with the term "formative assessment".
Average Scale Score	The scoring scale used by TIMSS for the longitudinal evaluation of countries and for the comparison between countries.
CAPS	Curriculum Assessment and Performance Standards.
(a/the) choice	In a multiple-choice test item, this indicates a particular choice option from the range of options for this test item.
conceptual model	The internal knowledge held by a learner at a point in time, which is deemed to consist of a collection of schemas that are used both to understand mathematical problems and to solve such problems.
(a/the) construct	When used as a noun in the context of a learner's conceptual model, this term is the same as "schema".

CTT	Classical Test Theory.
decimal comma	A comma used as a decimal mark.
decimal fraction	The fractional part of a decimal number after the decimal mark.
decimal mark	The mark used to separate the whole number part from the decimal fraction part of a decimal number. This is either a decimal comma or a decimal point.
decimal number	The entire decimal number consists of a whole number part, a decimal mark, and a decimal fraction part.
decimal point	When the point (.) is used as the decimal mark.
decision matrix	A grid of rows and columns in which the cells represent decisions or actions to be taken when the conditions associated with a particular row and column are met.
(development) stage	This term is used in the context of my model of Development Stages to identify the stage of knowledge of a learner, within the context of a specific micro-domain, at a specific point in time.
Development Stage (model)	When used with initial capitals, this refers to the model of learning in a micro-domain which is introduced in Chapter 3, Table 2 on page 95.
diagnostic assessment	A form of assessment that detects conceptual obstacles to learning, such as schemas that are not fit for their purpose.
Difficult	This term is used to indicate a learner's identification of the level of difficulty of a test item presented to them in a diagnostic assessment as being one of: Easy, Just Right, or Difficult.
distractor	In a multiple-choice question format, a distractor is any choice presented which is not a correct choice. Distractors are divided into rich distractors and random distractors.
(level of) difficulty	The selection, by the learners, of one of the options of Easy, Just Right, or Difficult when asked how difficult they found a particular test item.
DBE	Department of Basic Education (National), Republic of South Africa. This was formerly part of the combined Department of Education (DOE).
DOE	Department of Education (National), Republic of South Africa. The former name of the Department of Basic Education.
D/K	Don't Know.
Easy	This term is used to indicate a learner's identification of the level of difficulty of a test item presented to them in a diagnostic assessment as being one of: Easy, Just Right, or Difficult.

EMERGENT	The development stage of a learner in a micro-domain, defined as "the learner is *just starting to know* this micro-domain". This is considered as the novice stage of conceptual development.
FET	Further Education and Training. The final phase in the South African school structure, which comprises Grades 10-12.
formative assessment	This term is commonly used as a synonym of "assessment *for* learning". Formative assessment is a process and a tool used in the classroom as outlined by Wiliam (2011b).
GDE	Gauteng Department of Education.
GDP	Gross Domestic Product.
IMMINENT	The development stage of a learner in a micro-domain in which the constructs are sufficiently mature to handle most problem situations. This is defined as "the learner *almost knows* the micro-domain".
INFIT	A statistic for misfitting items which identifies an item's fit to the Rasch model by giving more weight to learners whose ability measure is close to the item measure.
IRT	Item Response Theory.
item	Same as "test item".
Item	When used with an initial capital, Item refers to an identified test item from the Item Bank as provided in Appendix D.
Just Right	This term is used to indicate a learner's identification of the level of difficulty of a test item presented to them in a diagnostic assessment as being one of: Easy, Just Right, or Difficult.
late-stage misconception	This is a misconception which is used by IMMINENT stage learners, thus occurring late in the development stages.
learner	This is the term in common usage within the South African education system to represent the individual engaged in learning, as distinct from the teacher or educator. The terms pupil, scholar, and student are in less common usage in South Africa.
Lesson	In the context of this study, this means the specific school lessons which were used to conduct the online diagnostic tests.
Mathematics	When used with an initial capital letter, this refers to the school subject of Mathematics as in the South African curriculum (DBE, 2011a). When used without an initial capital this refers to the general discipline of mathematics.

MCQ	Multiple-choice question. This is a test item presented to a learner in which the response is required to be a selection of one and only one of a set of choices presented.
MERGA	Mathematics Education Research Group of Australia
micro-domain	This is a small subset of a larger domain of mathematics knowledge, which is typically used as the basis for a single lesson or a single type of problem. Micro-domains are used in this study as the target for diagnostic assessment practices.
(my) model	My Development Stage Model of learning as introduced in Chapter 3.
NCS	National Curriculum Standards, as published by the Department of Education, Republic of South Africa.
N/S	Not selected.
OECD	Organization for Economic Co-operation and Development
OUTFIT	A statistic for misfitting items which identifies an item's fit to the Rasch model by giving equal weight to learner's measures and which is more sensitive to outliers such as guesses and slips.
point-measure correlation	This is a correlation between the responses for the items with the person measures. This is shown in the PTMEASURE column of the WinSteps outputs.
PISA	Programme for International Student Assessment
random distractor	A choice in a multiple-choice question which is neither a correct choice nor a rich distractor.
rich distractor	A choice in a multiple-choice question which is designed to elicit evidence of one or more misconceptions.
RMT	Rasch measurement theory
RQ1	Research Question 1
RQ2	Research Question 2
RQ3	Research Question 3
schema	A schema is a theoretical unit of knowledge in a learner's conceptual model. A schema is conceived as unit of the conceptual model which is constructed and used by learners during problem-solving activities in mathematics with the assumption that schemas are called upon as required.
School A **School B**	The two schools used in this study.

self-knowledge	Knowledge held by a learner concerning the scope and limitations of their own knowledge, in terms of what they know and what they do not know.
slip	A mistake made by an otherwise proficient learner.
STABLE	The development stage of a learner in a micro-domain in which the constructs are stable and mature and where the learner is able to correctly answer every problem presented. This is the expert stage of knowledge and is defined as "the learner *knows* this micro-domain".
(development) stage	A stage of development of a learner in the context of a micro-domain, as used in the Development Stage model. These stages are ABSENT, EMERGENT, ACTIVE, IMMINENT, and STABLE, in terms of increasing maturity of cognitive development.
summative assessment	A form of assessment with the goal of assessing proficiency at the end of a course of study, such as is the case with final examinations. This is also referred to as "assessment *of* learning" and has the goal of positioning learners on a scale of achievement or grading.
test item	A problem presented to the learner in an assessment setting or as a part of a teaching and learning process.
TIMSS	Trends in Mathematics and Science Survey.
Web	The World-Wide Web
zone of competence	Within a micro-domain, this zone is a position where the learners can be located if they demonstrate proficiency in the micro-domain. This encompasses the development stages of STABLE and IMMINENT.
zone of incompetence	Within a micro-domain, this zone represents learners who show no understanding of the micro-domain and no evidence of having developed conceptual models or schemas. This covers the ABSENT development stage.
zone of learning	Within a micro-domain, this zone represents learners who show evidence of schema development, but for which the schemas are insufficient for full proficiency. This covers the development stages of EMERGENT and ACTIVE.
ZPD	Zone of Proximal Development.

CHAPTER 1 : INTRODUCTION

> *"…assessment is a, perhaps the, central process in effective instruction" Wiliam (2011a).*

Humans are tool builders. Tools allow us to control our world better and to do things which we cannot do without these tools. The tools we build are both physical, such as the wheel, the motor car, the mobile phone; and they are also conceptual, such as models, theories, languages, money. By far the most important conceptual tool ever built by humans is mathematics—both in its pure form as a set of common representations and techniques; and in its applied form to address the problems of humanity. By using mathematics we can understand our world better, and from this we continually develop our knowledge and practices in almost every discipline of human activity. We also use mathematics as the basis for developing new tools, both physical and conceptual.

Mathematics, like all conceptual tools, exists primarily in the human mind, and it is from the mind that it is developed, accessed and used. The application of conceptual knowledge may be direct, by the individuals themselves, or may be delegated, such as to computer programs or to intelligent robots. Even though mathematics has been developed and taught for thousands of years, the evolutionary process has not hard-coded mathematics into our DNA and it is rather our capacity to learn and to develop the conceptual models of mathematics which is a distinguishing feature of humans. Given the increasing societal demand for mathematical proficiency, our educational systems emphasize the development of mathematical knowledge as a core element of a successful education. Research continues to explore how learners come to know and to apply mathematics, and how such learning can be improved.

Given that we are not born with innate mathematical knowledge, we must learn by constructing our own mathematics as a personalized conceptual toolset. This process of knowledge construction, which does appear to be an innate function of the human, involves a continuous process of firstly developing a conceptual model, then testing this model on our experiences and observations, and finally refining this model until it is in line with our observations. As we construct and adapt our models, they will continue to

improve as reflections of the world of our observations, but there will remain gaps between this inner knowledge and the outer, observed experiences, and the gap between our internal conceptual model and the external observations is a dynamic point of conflict. Where our conceptual model fails to fit the observations, an innate process of equilibrium will induce our conceptual model to adapt. In this context it could be considered that we are all always in a state of learning—if learning is seen as this process of conceptual development. As a learner develops personal conceptual models of mathematics they will fail when their models are insufficiently developed to match the observations of the world. This occurs, for example, when attempting an unseen type of mathematical problem. Teachers gain useful evidence from their learners' attempts, since the nature of the mistakes can lead the teachers to better understand their learners' conceptual models and to provide suitable instruction to guide the learners towards conceptual models which are a better fit for their purpose.

My study explores how we can better understand the conceptual models developed and used by learners as they make sense of their mathematical experiences; where this information can help teachers to provide the right mix of observations to accelerate the natural process of adaptation which drives conceptual development. Mathematics is a very large body of knowledge and my specific focus is on the rational numbers, which are introduced in the Senior Phase of the South African curriculum, in Grades 7-9 (DBE, 2011a).

1.1 Background to this Study

Learners make mistakes as a part of their ongoing learning process. However mistakes are essential to the learning process and are not inherently negative (Nesher, 1987; Smith, diSessa & Roschelle, 1993). Within the domain of the rational numbers there has been extensive inquiry, over a long period, on the nature and extent of learner mistakes and the nature of effective remediation (De Morgan 1831/1898/2013; Monroe, 1917; Robertson, 1924; Buswell & John, 1926; Brueckner, 1928a; Brueckner, 1928b; Neal & Foster, 1928; Cooke, 1931; Cooke, 1932; Guiler, 1945; Vinner, Hershkowitz & Bruckheimer, 1981; Hiebert & Wearne, 1985). This long-term programme of inquiry has led to an improved understanding of the nature and causes of learner mistakes, and has identified that many types of mistakes are of a systematic and predictable nature; arising as a commonly-

experienced step in the conceptual development of learners (Radatz, 1979; Sackur-Grisvard & Leonard, 1985; Movshovitz, Zaslavky & Inbar, 1987; Resnick et al., 1989; Steinle, 2004b). These systematic and predictable errors are referred to as misconceptions (Smith, diSessa & Roschelle, 1993) and are conceptual models which are not suited for the task to which they are applied—being incomplete or incorrect. Misconceptions are distinct from stable conceptual models which have evolved to the point where they are sufficient to address a specific type of problem. Such stable conceptual models form the core of proficiency, and are labelled as "stable" because they are less subject to change, having adapted to the point of suitability for general use in a specific class of problems.

Diagnosis is an important element of teaching practice, assisting with the identification of the causes of learner errors. True diagnosis should not merely identify the occurrence and form of errors (Olander, 1933; Sprague, 1939; Bejar, 1984; Wylie & Wiliam, 2006), but is best structured as an integral element of a formative assessment practice (Black & Wiliam, 1998a; Black & Wiliam, 1998b; Wiliam, 2011a; Wiliam, 2011b; Stacey, 2013; Stacey, Price & Steinle, 2012). My study explores the diagnostic assessment of misconceptions in the rational numbers, specifically concerning the effectiveness and efficiency of such diagnostic work, as part of a larger vision of improving success in the learning of mathematics through automating and replicating such diagnostic processes using the Web.

On a personal note, I have spent more than 40 years as a part-time mathematics tutor and during this period I have observed that every generation of learners moves through similar paths of conceptual understanding as they tackle the rational numbers. I believe that much can be gained from analyzing the mistakes made by learners and that this analysis is greatly improved when the right problems are posed to the right learner at the right time. Such right problems maximize the ability to obtain evidence of the fine-grained learner thinking which accounts for their responses to problems. In my tutor role, through gaining access to my students' fine-grained knowledge within specific contexts, I have improved my effectiveness in tutoring, leading to a gain in learner understanding that would not have been achieved without this knowledge. Rather than tutoring from a rote lesson structure, my modus operandi has been to identify those gaps in the learners' knowledge which are obstacles to understanding and to address these as a priority.

Thus, one of my insights from these tutoring activities is that there are some questions which are good for diagnosis of learner misconceptions, and are both effective and efficient in eliciting evidence of thinking. The contrary insight is that the majority of questions add little or nothing to our understanding of learner thinking, and I see this distinction as important in order to develop our knowledge to provide effective cognitive diagnosis.

My personal motivation to embark on this inquiry was to better understand and to formalize diagnostic assessment in the classroom, by positioning such assessment as an integral element of daily teaching and learning, and from this to provide real benefits to the learners within the limited time scheduled for teaching the rational numbers. This motivation was extended to consider how diagnostic practices could be understood sufficiently to be replicated using computer systems, and made available to learners throughout the country using the World-Wide Web (the "Web").

Diagnostic assessment is a broad area of educational measurement, and my focus was on the rational numbers, which were identified as an important element of the shift in knowledge from the relative simplicity of the whole numbers to the complexity of the real numbers. This shift is positioned within the "transition years" of Grades 7-10 which Usiskin (2005) identifies as the period of greatest need for learners of mathematics, and where schools should be encouraged to assign their best mathematics teachers.

From this study I hoped to uncover the qualities of mathematics questions which render them effective assessment instruments for cognitive and conceptual diagnosis. My plan was to undertake an empirical evaluation of various diagnostic questions, using a purely statistical approach to identify the diagnostic value of a particular question. However, this notion of "diagnostic value" is not well defined, and part of my study was to clarify this notion as a theoretical construct on which to base the remainder of this study. I also explore, in selected cases, the qualitative nature of such questions—to determine the conditions under which one mathematical question may be better than another in a given situation from external observation and without statistical evidence. The challenge is that questions which appear similar on the surface may have vastly different behaviours when used for diagnostic purposes. I thus clarify the distinction between two purposes of testing: firstly, to establish "ability" within a domain of

knowledge; and secondly, to diagnose particular conceptual models which do not fit their purpose, and which require attention.

I have used the word "question" to describe particular mathematical problems posed to a leaner within the scope of this introduction, but my preference is to use the terms "test item", or just "item", to identify a mathematical question which is being applied within an assessment context. These terms are adopted consistently within the literature on educational measurement.

I continue this chapter by reflecting on the national need for improvement in the state of school mathematics in South Africa with a targeted focus on the domain of the rational numbers, which are a consistent challenge for all generations of learners and are also a fundamental building block to further mathematics. This chapter provides the rationale for this study and leads to the research questions which this study is designed to answer.

1.2 Outlining South African School Mathematics

Mathematical proficiency is heavily weighted as a component of the entry requirements for most higher-education disciplines. For example, at the University of the Witwatersrand, the admission requirements for scientific disciplines includes Mathematics at Levels 6 (70-79%) or 7 (80-100%), and for most other disciplines there is a minimum requirement for Level 4 (50-59%) or 5 (60-69%)[1]. Thus, without a good pass mark in the final Grade 12 Mathematics summative examination there are limited prospects for entry into tertiary studies in many disciplines.

Mathematics is used throughout life and work, and proficiency in mathematics is considered sufficiently important that the South African curriculum requirements for the Further Education and Training (FET) phase, comprising Grades 10-12, requires a compulsory selection of either Mathematics, focusing on pure mathematics; or Mathematical Literacy, with a focus on applied, real-world mathematics. Both Mathematics and Mathematical Literacy draw heavily on the theoretical and practical knowledge introduced in the Senior Phase of Grades 7-9. Many universities adopt

[1] http://www.wits.ac.za/prospective/undergraduate/admissionrequirements/11644/admission_requirements_nsc.html. Retrieved on 18 May 2015.

admission requirements which deny access to learners who have taken Mathematical Literacy in Grades 10-12, accepting only a qualification which includes Mathematics. Thus when exiting Grade 9, the 14-15 year-old learners are required to make life decisions on subject choices which may severely limit their future prospects.

The South African school system is administered by the national Department of Basic Education (DBE) which is responsible for the national curriculum and for educational policy, and this is then administered by the Department of Education at the Provincial Government level. The DBE is responsible for the Grade 12 school-leaving summative assessment as well as the Annual National Assessments (ANAs), which are used to monitor systemic education performance in numeracy and literacy. The school years are structured into four phases being the Foundation Phase (Grades R-3), Intermediate Phase (Grades 4-6), Senior Phase (Grades 7-9), and the FET phase (Grades 10-12). For each school subject the DBE provides curriculum statements, containing subject matter content and guidelines for scheduling and pacing of the classroom instruction for each of these four phases.

Rational numbers are introduced in the earliest years of schooling—in the Foundation and Intermediate phases—through the concepts of sharing and equipartitioning, which lead to the initial representations of simple fractions in both words and notations such as "one half", $\frac{1}{2}$; and "three-quarters", $\frac{3}{4}$. The rational numbers are given full treatment in the Senior Phase for which the content includes different representations including common fractions, decimal numbers, percentages, and proportional representations. This phase also covers the operations and calculations involving these various representations of rational numbers including conversions between different representations. Within the FET phase no further specific instruction on rational numbers is provided and yet this important class of numbers are used throughout the final three years of school mathematics in both the Mathematics and Mathematical Literacy curricula.

The South African curriculum policy statement is called "CAPS" (Curriculum Assessment and Performance Standards), and was introduced following a number of years of outcomes-based education (OBE), which was in force at the time that this study commenced. Reference is made to both the OBE and CAPS curriculum statements within this thesis.

The CAPS curriculum statement includes a schedule of instructional activities which teachers are expected to follow, indicating when in the school year each topic should be covered and how much time is devoted to the topic. Common fractions and decimal numbers are each given nine hours in Term 2 of the four-term school schedule for Grade 7; then seven and six hours respectively in Term 3 in Grade 8; followed by 4.5 hours each in Term 1 in Grade 9. These are the only time allocations within the entire Senior Phase schedule for these numbers—in effect a total of 40 hours' instructional time over a three-year period of the learners' school life. Given the richness of the intermediate conceptions and misconceptions which are developed as the learners increase their proficiency in rational numbers, I expect that this limited period of instructional time is insufficient for such an important area of mathematics.

1.3 South Africa and the TIMSS Studies

South Africa spends a considerable portion of its annual budget on education, being R165.1 billion for the 2010/11 national revised budget, which is 19.9% of the national government expenditure of R 829.6 billion (South African Treasury, 2010, p. 118), and which is also 6.1% of the GDP of R 2 699.9 billion.

South Africa has a considerable desire to succeed in basic education, as outlined in the curriculum statements (DBE, 2011a). However, even with the large allocation of funds for education, and the will to succeed from the educational policy, South African mathematics education has been assessed consistently as being within the lowest group of countries on international surveys conducted in the past 20 years. The Trends in Mathematics and Science Survey (TIMMS) study is conducted every four years at the Grade 4 and Grade 8 levels, and South Africa has participated in a number of the Grade 8 studies, but is consistently placed in the bottom three places. This is a serious situation which has not improved in terms of the relative ranking of South Africa when compared to other countries over this period.

Within the TIMSS 2011 Grade 8 mathematics study (Mullis, Martin, Foy & Arora, 2011), South Africa was assessed using Grade 9 learners rather than Grade 8 learners—this was a choice available to countries participating in TIMSS if it was expected that the Grade 8 test items would be too difficult for the country's Grade 8 learners. Even with this advantage of older Grade 9 learners, South Africa was in the

third-last position with an Average Scale Score of 352. This Average Scale Score is central to the TIMSS measurement system, and is fixed at 500 as the benchmark average with each 100 points on the scale indicating one standard deviation over the entire set of country scores. This scale is structured to be invariant across assessments and is a reliable measure, enabling each country to measure its progress between the assessment years, and also to position itself against other countries on the international scale.

The TIMSS 2011 Grade 8 mathematics study used a bank of 217 items which are divided among four content domains (Number 30%, Algebra 30%, Geometry 20%, Data and Chance 20%), and also divided on the dimension of cognitive domains (Knowing 35%, Applying 40%, Reasoning 25%). The items are split equally between multiple-choice questions (MCQ), where the learners select from a limited list of choices, and constructed-response questions, in which the learners are required to write their answers.

South Africa did not participate in the 1995 or 2007 TIMSS studies, but participated, and was placed last, in both the 1999 and 2003 studies. In 1999 South Africa achieved an Average Scale Score of 275 and in 2003 the score remained essentially the same at 269. For the 1999 study South Africa was also reported as spending 8.0% of the Gross National Product on education—the third highest among the set of 38 participating countries (Mullis et al., 2000, p.25).

For the 2011 TIMSS mathematics study the Average Scale Score for South Africa jumped up to 352, from the bottom up to third-last place (Mullis, Martin, Foy, & Arora, 2011). This is encouraging and can be viewed as relatively significant, since the lowest performing countries were far below other under-performing countries and thus to move up and out of the bottom place requires a substantial national effort.

These international studies are important resources to help countries understand their position in a global context and as a catalyst for improvement. Such improvements are needed continuously throughout the entire educational system of the country since it is the totality of the elements of educational system which result in the learner responses as the observable symptoms of the national status of mathematics education.

The TIMSS 2011 study (Mullis et al., 2011) reports that South African learners are significantly under-achieving on rational number items, and for some of the test items the results are so low that they indicate a systematic basis in the errors, which I now analyze in further detail. I specifically focus my attention on the items classified under

the topic of “Fractions and Decimals” within the “Number” content domain, which was the selected mathematical area for my study. Within TIMSS 2011, 25 out of the total item bank of 217 items fall within this topic. Of these 25 items, 11 are included within the Released Item Set (IEA, 2013), which are made available for further study by TIMSS. These 11 items are split over the cognitive domains of Knowing, Applying, and Reasoning in the ratio 6:4:1. Thus the assessment of the proficiency in common fractions and decimal numbers is measured mostly in terms of conceptual knowledge rather than in the application of this knowledge or with problems which require reasoning. I now dive into the reported results for 4 of these 11 test items from the 2011 TIMSS Grade 8 mathematical study, to see to what extent these items from international studies support my study on the quality of test items for diagnostic purposes. These items are M02_01, M02_04, M05_02 and M07_02.

Item M02_01

Item M02_01 asks the learner to select the correct decimal representation of the common fraction $\frac{3}{5}$ when given four choices (A) 0.8 (B) 0.6 (C) 0.53 (D) 0.35. The correct choice is (B) and the other choices are distractors which may elicit a response from learners who lack a sufficient understanding of the relationship between common fractions and decimal numbers. I note that each of these distractors is not randomly introduced, but has been carefully designed to provide a plausible alternative—based upon various ways of thinking in common fractions and decimal fractions. The question statement contains the common fraction $\frac{3}{5}$ which contains the digits 3 and 5, and these digits also feature in three of the choices provided: Choice (A) uses the digit 8 which is 3+5; Choices (C) and (D) use 3 and 5 within the structure of the decimal numbers 0.53 and 0.35. The only choice which does not use digits 3 or 5 is the correct choice (B). Thus, whereas this test item is used for systemic assessment, it is designed as a diagnostic question for which incorrect responses point to alternative ways of thinking. The results show that 61% of South African learners obtained the correct response, which was below the international benchmark average of 68%. Given that South Africa used Grade 9 learners for this study, the performance may be impacted by the relative curriculum scheduling of the different countries, since the knowledge of rational number conversion would be introduced in different grades in different countries. Thus learners in some

countries may have been exposed to this kind of problem in advance of the TIMSS assessment, whereas others may not have had such exposure in the mathematics classes.

Item M02_04

Item M02_04 concerns the selection of the correct method for subtracting two common fractions and is classified within the "Applying" cognitive domain. South Africa ranks in joint second-worst place with Chile on a score of 12%. This item is in MCQ format with four choices, and thus a pure random response from a single learner has a 25% probability of being correct. When this is averaged over the entire learner sample for a single TIMSS country assessment, the expected country score is also 25% if all responses were purely random. Inferences can be drawn from this score, since a score which is significantly higher than 25% indicates better knowledge among some of the learner samples, and a score significantly lower than 25% points to a non-random pattern of responses—in effect a systematic, as well as systemic, pattern of selecting incorrect choices.

On the basis of this reasoning, a low score of 12% for South Africa points to the existence of systematic errors which were commonly made by the learners, and this provides evidence of some level of conceptual development in the learners, in their attempt to make sense of the problem using their conceptual models. Thus this apparently very low score of 12% is more likely to arise from the consistent use of misconceptions than to result from random guessing. I view this as an important issue, concerning the general problem of measuring low performing learners or groups of learners, which is significant for my work in exploring how to identify the cognitive causes of such low results.

As a contrived example to further illustrate this point, consider a group of 100 learners who are each presented with a single MCQ with four choices. Consider also that 20 of the learners select the correct response on the basis of their proficiency and the remaining 80 have no proficiency and select a choice on a purely random basis. I am not initially considering learners with partially developed conceptual models, or misconceptions, so this is an artificial situation in which the learners either know or they randomly guess. I consider random guessing as a response from a learner which does not result from using either stable conceptions, and also does not result from any other rule, such as always selecting the first choice presented. In such random guessing, there is an

equal chance of each of the choices being selected. For this example, there are 20% correct from the 20 learners who selected correct choice, and another 25% of the remaining 80 learners (being 20 learners) who selected the correct choice by random guessing, making up 40% (20+20) of the 100 learners who selected the correct choice. In the alternate situation, in which none of the learners are proficient and all engage in random guessing, 25% of the learners are expected to obtain the correct answer purely through this random guessing. On the basis of this argument 25% is then the minimum possible score, since if there are some learners who are sufficiently proficient to answer the question successfully with the remainder of the learners guessing randomly, then the expected success rate will be larger than 25%. Thus, success rates of less than 25%, such as the 12% determined from the TIMSS study for this item, point to the inference that the learners who do not select the correct response through their level of proficiency are not guessing randomly. Rather these learners are applying some intermediate conception or misconception to answer the question, and this is being done consistently across the entire sample of learners to cause such a low score.

I infer that the very low scores on rational number items are likely attributable to misconception usage rather than random guessing. Whereas the TIMSS results consider the impact of guessing, by their usage of the Item Response Theory (IRT) three-parameter model when calculating the Average Scale Score, it is not evident that this consideration for guessing is also used when determining the success rates of the individual questions. Thus my argument for the potential evidence of nation-wide misconceptions from the low TIMSS success rates points a way to further inquiry. My argument is that the methods we use to determine proficiency, such as using the fraction of learners who achieve success on an item, are not suited for measuring these intermediate conceptions and misconceptions, which are exhibited as systematic errors made by learners, and which thus can result in scores which are lower than the expected scores arising from random guessing. Thus a different method is needed for this measurement which is more suited for low-performing learners.

Item M05_02

Item M05_02 asks the learners to position a number “N” on a number line. This is categorized under the “Reasoning” cognitive domain; the most complex of the range of cognitive categories. The question stem presents a number line on which the positions

of 0, 1, and 2 are marked with ticks and labelled with these numbers. Between the 0 and 1 ticks there are two named points, P and Q, with P<Q. The question asks where the value of N should be placed, where N = PxQ. Four choices are presented for the position of N: (1) between the 1 and 2, (2) between the Q and 1, (3) between the P and Q, and (4) between the 0 and P. The success rates achieved by the best performing countries are in the region of 50%, while South Africa scored among the lowest with 10% which, as for Item M02_04 above, is far below the expected value of 25% which would have resulted if the learners had been guessing randomly. Using the argument which I previously used for accounting for such low success rates, I again conclude that such a low success rate is not expected to result from pure guessing and this points to the existence of systematic selection of the non-correct distractor choices, with this being common throughout the learner sample. My argument does not imply that there is only one such systematically incorrect approach, but rather that the combination of such alternative choices as selected by the learners has a higher frequency of response than for the correct choice. For this test item there is a well-known misconception, in which multiplication will always result in a larger number while division will always result in a smaller number (Bell, Swan & Taylor, 1981). This is a plausible explanation for choices (1) and (3), both of which are larger than Q, which itself is larger than P on the original number line. Thus this test item, like many others within the TIMSS item bank, is also diagnostic in nature—by its inclusion of choices which point to known or predicted misconceptions. Would the success rate for South African learners have been so low if the distractors were not explicitly linked to misconceptions but were rather random choices? It is likely that in such a case the results may have been 25% or more, since there would be no plausible explanations for the various choices. However, this particular item cannot be easily reformulated to avoid misconception in its distractor choices.

Item M07_02

Item M07_02 asks the learners to convert the common fraction $3\frac{5}{6}$ to a decimal number, rounding the answer to two decimal places and is categorized under the "Knowing" cognitive domain. This is a constructed-response question, and only four countries achieved a success rate higher than 50%; with South Africa at 7% being in a group of countries achieving less than 10%. Given the open-ended nature of constructed-

response questions, it is more challenging to interpret the results and to identify misconceptions. Whereas these constructed-response questions are suitable for formative and summative assessments, such open-ended questions require considerable analysis to address each of the possible and unusual outcomes and thus they do not appear to be suited for automated diagnostic assessments. Further research, using the techniques of artificial intelligence, will help with automating the interpretation of such open responses, however this is beyond the scope of this study.

In conclusion, there is evidence, from the analysis of the TIMSS 2011 results and the four cited items, that South African learners are using some intermediate conceptions or misconceptions as the basis for their responses. Thus the measured success rate—in terms of the fraction of learners who achieved success—does not accurately reflect the level of conceptual development of the learners, and appears to consider that the learners either are proficient or are guessing, since the scoring method cannot adequately address the partial conceptual development which gives rise to informed, but incorrect, responses. Arising from this initial analysis of the TIMSS published results, I largely reject the use of raw success rates for identifying low-performing learners and I identify the need for an improved model which can position learners on a scale of conceptual development. Whereas I have used TIMSS to help set the context, I do recognize that TIMSS is a systemic study for ranking of country-level performance, and does not have as its core goal the diagnosis of individual learners.

1.4 Diagnostic Assessment for Learning

I now turn my attention to how effectiveness of assessment can be improved, to better inform teaching and to consequently improve learner proficiency, so that learners will succeed more often on mathematics problems and specifically on the types of problems posed in class tests, in the annual summative assessments, and in international studies such as TIMSS.

Effective diagnostic assessment means that learner conceptions can be measured accurately and can guide effective instruction. Without a fine-grained window into learner thinking, instruction cannot target the needs of the individual learners (Wiliam, 2011b). This targeted approach to instruction assumes that learning progresses in a constructivist process through the learners' invention, use, and self-reflective evaluation of conceptual

models, as they adapt to increase their success and to avoid failure. There is a continuous process of learning in which learners develop conceptual models to respond to the external world—which for this study consists of mathematical problems in the rational numbers. My position is that learner knowledge consists of a connected structure of such conceptual models and by succeeding and failing on various problems the learners will refine these conceptual models to meet the need to create the equilibrium between the internal, mental world of the learner and their observations of the external world (Piaget, 1964/2003; Piaget, 1985). I use the term "conceptual model" to represent a collection of "schemas", each of which is applicable for a specific pattern which a learner uses to address mathematical problems, and which are constructed and changed through the learning process. I thus see these collections of schemas as the totality of a learner's knowledge, and I see learning as the progression of changes in the collection of schemas.

Teachers cannot directly access their learners' minds to identify, observe, and modify their learners' schemas—there is no equivalent to brain surgery available to the teacher to diagnose and to fix these conceptual problems. For the teacher, the knowledge of their learners is limited to what they can observe indirectly from the learners' responses to mathematical problems as well as through learner utterances and engagement during instruction and tutoring. To discover what is wrong with learners who are making mistakes is a diagnostic task and is a process of exploring the possible explanations in the form of a lack of conceptions or as misconceptions with the follow-on remedial work. These diagnostic tasks involve asking a question and analyzing the results to help to infer the schemas that can give rise to the response. Such diagnostic practices are more effective when the right questions are asked, since the better the questions, the more that can be inferred about the learners' conceptual models. The more that can be inferred, the better the teacher can respond by planning targeted instructional activities. In the absence of such information about individual learner conceptions, teachers cannot direct their instruction to meet the specific needs of the learners, since these needs will be unknown. In such cases teaching is not optimal, and in worst case scenarios all teaching may be ineffective. Thus my concern is with the improvement of our knowledge of learner conceptions, and specifically to position each learner's state of knowledge onto some spectrum of increasing conceptual development. My concern is how to determine the specific schemas a learner uses for his or her responses; including pre-conceptions, prior

knowledge, intermediate conceptions, misconceptions, or stable conceptions. There are often situations in which learners have a complete lack of schemas for a given problem and thus can only resort to random guessing.

I contend that not all questions are equally suited for diagnostic purposes and that, even with the wealth of knowledge which has been accumulated in understanding and modeling the trajectories of learning in the development of proficiency in the rational numbers, there remains an open question of how to use this diagnostic knowledge for improving teaching and to stimulate learning. Whereas this may be seen as merely a practical application of existing knowledge, I argue that there is no clear and well understood path that takes us from a theoretical understanding of misconceptions to the practical application of this wealth of knowledge. Rather, given a range of test items, within a particular domain of knowledge, I postulate the existence of a spectrum on which each item can be placed in terms of its diagnostic effectiveness.

This falls into the practices of diagnostic assessment, positioned within a framework of formative assessment, or "assessment *for* learning", in which assessments inform teaching activities rather than being used solely to assess learners. It is within the context of formative assessment that diagnostic assessment is best positioned to make a difference to improving learning. Diagnostic assessments conducted outside of a formative assessment framework are likely to be piecemeal and not directed towards the goals of mathematics education.

1.5 The Right Questions

My study is concerned with identifying questions which are suited, and perhaps optimal, for eliciting evidence of learner thinking, and which can be used for diagnostic purposes as an integral element of formative assessment practices, and which can be called "*diagnostic* assessment *for* learning". Such questions will support more effective diagnostic processes and practices, in terms of the validity and reliability of inferences drawn from the learner responses, and to ensure efficient application by being conducted with the least effort and in the shortest time. I see this as important for the design of diagnostic assessments given that school time is a highly constrained resource, with only a short amount of time allocated to each topic in the mathematics curriculum. As indicated earlier, only 18, 13, and 9 hours are allocated to decimal and common fractions in Grades

7-9 respectively, and this very short time allocation must be used as effectively and efficiently as possible to ensure that the learners develop maximum proficiency in the rational numbers.

To achieve this economy of practice, I have identified two key issues which need to be addressed. The first is to ask questions which provide valid evidence of learner thinking, and the second concerns the fact that a number of such questions may be required to infer a learner's stage of conceptual development in the area of concern. The first issue concerns the effectiveness of diagnostic assessment—are we correctly inferring the existence of particular ways of thinking? The second issue addresses the efficiency of the diagnostic process—are we making these inferences using the least amount of time and effort?

1.6 The Role of Self-Knowledge

Learners possess knowledge, in the form of schemas, which are used to address the mathematical problems and questions as presented. When a new problem is presented, the learner selects schemas, the response is determined, and the results are given. However, in addition to these schemas, learners also have self-knowledge about the scope and limitations of their own knowledge, in terms of what they know and what they don't know.

Whereas assessments are objective practices of gathering data from observations and making inferences from the learners' responses, it is also possible to inquire into the learners' self-knowledge by asking whether they know how to solve particular problems. This self-knowledge can be expressed verbally by the learners using statements such as "I know how to solve this problem" or "I don't know how to solve this problem". Thus, whereas we can rely solely on observed responses to assessment questions, there is an opportunity to use such learner self-knowledge as an alternative source of diagnostic information. This opens the door to an inquiry on the value of such self-knowledge in improving the effectiveness and efficiency of diagnostic practices.

1.7 Diagnostic Assessment

Diagnostic assessment, as outlined in the South African curriculum statements, is a broader activity than I outline above, and is defined as follows:

> It is not intended for promotion purposes but to inform the teacher about the learner's Mathematics problem areas that have the potential to hinder performance. Two broad areas form the basis of diagnostic assessment: content-related challenges where learners find certain difficulties to comprehend, and psychosocial factors such as negative attitudes, Mathematics anxiety, poor study habits, poor problem-solving behaviour, etc. Appropriate interventions should be implemented to assist learners in overcoming these challenges early in their school careers. (DBE 2011a, Senior Phase, p.154).

My concern is with the first of these areas, being the content-related challenges, which focus on conceptual limitations which hinder mathematical performance and constrain the advancement of learning. However, mathematical proficiency is non-linear and multi-dimensional in nature, and such conceptual limitations are also complex in nature. This is evident in the work of Kilpatrick, Swafford and Findell (2001) who structure mathematical proficiency as five interleaved and interwoven strands:

- *conceptual understanding*—comprehension of mathematical concepts, operations, and relations
- *procedural fluency*—skill in carrying out procedures flexibly, accurately, efficiently, and appropriately
- *strategic competence*—ability to formulate, represent, and solve mathematical problems
- *adaptive reasoning*—capacity for logical thought, reflection, explanation, and justification
- *productive disposition*—habitual inclination to see mathematics as sensible, useful, and worthwhile, coupled

> with a belief in diligence and one's own efficacy. [italics in original] (p. 116)

Content-related challenges may be present in any of these strands but it is "conceptual understanding" which I primarily address in this study, given that this is a necessary basis for proficiency in the other strands. Without conceptual understanding it is not possible to carry out operations, not possible to decide on methods to use, and not possible to reason mathematically. Concepts form the language of mathematics and this language must be known in order to "do" mathematics. In many ways this language is mathematics. For the rational numbers this language includes the verbal and notational representations of common fractions and decimal numbers, the operations on these mathematical objects, and how these are related to visual and more tangible representations such as graphical diagrams and number lines. Concept-related challenges may be present in any of these language elements of the rational numbers, and may also exist in other strands of proficiency. Whereas the above DBE definition of diagnostic assessment highlights the nature of content-related challenges, and states the need for intervention, it does not indicate how such content-related challenges should be discovered, and also does not indicate the nature of the possible interventions. The CAPS document suggests that "Assessment for learning has the purpose of continuously collecting information on learner performance that can be used to improve their learning" (DBE 2011a, Senior Phase, p. 155), which includes a range of informal, non-documented, tasks which are conducted on a daily basis in the classroom. Within this study I address the usage of diagnostic tests to provide evidence for the discovery of content-related challenges, and I also explore the potential for automation of this task. With my approach being the application of targeted diagnostic tests, I thus need to know how I make inferences from such tests, and ideally as few test items as possible to meet the requirement for efficiency, in order to derive some measure which could be useful and meaningful to the teacher.

1.8 From Scores to Measures

Tests are instruments which gather data from learners for various assessment purposes. Tests are the "ruler" which we use to determine the extent of a particular construct, such

as ability in a specific domain, or more importantly for this study, evidence of a content-related challenge.

Consider a test which is comprised of five items, from which a score in the range 0-5 is possible from each of the learners in the class, on the basis of one score point for each successful response. However, what does this score mean? Whereas test results are used as though they are measures—with the score 0-5 being used as evidence which means something, like the ruler is used for measuring length—this is not actually the case, and this is due to a range of considerations that exist between the nature of a specific test, and the inferences that we can draw from the test results. Foremost of these is the validity of the test items, concerning whether they are appropriate instruments to measure the construct of interest.

Whereas international surveys, such as TIMSS, are designed to be good measures, this requires considerable effort from a large body of experts to develop such suitable instruments, and such resources are not available for informal assessment in the classroom. We have no easy way to determine whether a teacher is providing the right test items to aid their daily assessment practice, and also whether they are using the results to elicit valid evidence about their learners' challenges. To examine this further, it is necessary to dive deeper into the testing process, to identify how this can be improved to provide valid evidence to the teacher. This leads to the specific research questions that I frame at the end of this chapter.

For my purposes I consider a test as consisting of a set of items, where each item is designed as a suitable instrument for the "construct" which the test as a whole is measuring. This construct is an intangible element of the learners' conceptual model. A construct must be measureable so that the test instruments can produce a value—in which different learners will achieve different scores resulting from the items in a test. Higher scores are evidence of more of the construct and lower scores indicate less of the construct. However, not all test items will be equally useful for measuring the construct, and thus using the raw test scores alone is not a fair measure.

Rather, the ideal for educational measurement is to create fundamental measures, such as exist within the physical sciences. For example, there are well-developed international standards for length and for temperature, with corresponding scales of measurement that allow for consistent readings anywhere, at any time, using any

instrument which is aligned to the standards. Educational measurement strives for the same consistent rulers and thermometers which meet this requirement for a fundamental measurement system. With such a fundamental system of measurement, educational tests can produce universally accepted measures of a construct—using different learners, at different places, and using different instruments. This requirement for fundamental measurement is an essential element of modern educational measurement practice (Bond & Fox, 2012).

We use tests as a means to gather evidence from our learners. Each test gathers data indirectly, such as by observing a learner in the process of solving a problem. However, such observations are also a challenge, since many of the processes involved in solving a problem take place in the mind of the learner and are not explicit and thus not directly observable. The learner may read aloud the methods they are using but this may not produce the results we need since many mental process are not at the conscious level of thinking and occur too fast to explain. Thus, it is easier to limit our observations to the final responses given by learners to the problems as presented, as is done with both MCQ and constructed-response items. I am not discounting the potential value of observations other that the learner's final response, but these other observations are not used within this study and are also time-intensive, requiring the teacher to be devoted to one learner at a time. Thus for practical purposes we need to work with the methods which are available to us, being the presentation of a test and the inference we can draw from the test scores.

Given a single test item, structured to measure a particular construct, and a learner's response, the question arises of how good is this test item for measuring the construct of concern. The construct may be broad, such as the learner's overall proficiency in rational numbers, or may be finer-grained, such as the ability to add common fractions with the same denominators. It is essential that test items are valid and reliable instruments for the construct they are measuring, and that they are used appropriately to infer measures which are valid. This is termed "construct validity" and is the degree to which inferences drawn from test results are valid and reliable measures of the construct (Messick, 1989; Messick, 1995). However, in most cases in educational measurements the constructs are not defined in detail, and are rather expressed as a short description of a desired competency within a curriculum statement such as, for example

"the conversion between different forms of rational numbers". This is a statement indicating a proficiency, and misconceptions can also be expressed in similar words, such as "ignoring the decimal point when determining the value of a decimal number". The next step is to find suitable measurement processes to quantity these constructs so as to provide evidence to teachers about the learner's challenges.

Individual test items are unique, and have their own properties. Some test items may be more difficult than others, so that fewer learners succeed on the difficult items in contrast to the easier items. As a result, the items in a test may be all difficult or all easy, and this can bias the results. To illustrate this point, let us imagine that there are two tests S and T to be given to two groups of learners A and B of similar ability. Test S is composed of difficult items and is given to group A, and test T consists of easy items and is given to group B. Is this fair? Will we obtain accurate results from the tests? We would expect, since the ability of the groups A and B are similar, that group A will produce a lower score than group B due to the bias in the item difficulties between the two tests. But the two groups are both taking a test on the same topic and it is thus evident that using the success rates alone—how many learners select the correct responses—does not provide for a fair comparison between these two tests without also considering the difficulty of test items in these tests. For example, for questions in which a learner is required to convert a decimal number to a common fraction, it is likely that the decimal numbers 0.5 and 0.7943 have different levels of difficulty by considering only the individual nature of these two numbers and without using empirical evidence and statistical inference arising from the use of these items in actual tests with learners. It is required to consider the level of difficulty of the items to obtain an unbiased measure and this issue pervades all educational testing.

A second consideration is that some test items may be a better fit to the construct than other items. As a thought experiment, if I include a test item "What is 1+1?" in a test which purports to measure the ability to convert between decimal numbers and common fractions, it is clear that this item does not fit the construct and should be excluded. However, if this item does not fit the construct then how do I determine which items do fit the construct and which do not. Rather than this being simply a case of fit or non-fit there is expected to be a spectrum of fit levels, occurring between the extremities of "does not fit at all" to "fits perfectly", in terms of the extent to which a test item is able to

measure the construct. Whereas “What is 1+1” is an easy question, it does not fit the construct being measured.

Returning to the notion of fundamental measurement as found in the physical sciences, we use rulers to measure length and any ruler will do. When we purchase a ruler it is based on a standard and thus our measurements will not vary if we use different rulers to measure a single length, or we use the same ruler to measure two equal lengths. This property of invariance is an essential element of fundamental measurement and is thus also a necessary requirement for educational measurement. To address this requirement for invariant, and perhaps universal, measurement methods, Georg Rasch developed the method which bears his name (“Rasch Analysis” or “Rasch Measurement Theory (RMT)”) which jointly considers the difficulty of the items as well as the ability of the learners when analyzing the results of a test (Bond & Fox, 2012; Wright, 1997). Rasch’s method has led to a profound change in educational measurement theory and practice. I provide some details of the role of Rasch methods for diagnostic measurement in Chapter 2, and in Chapter 4 I describe how I used Rasch measurement for this study. My study is concerned specifically with test items that measure constructs which are useful in a diagnostic context, and measure misconceptions using a similar approach to the measurement of ability.

The Rasch process requires non-trivial computational effort and if this is to be applicable in a daily assessment context it is necessary to explore how it may be automated.

1.9 Automating Diagnosis

A learner’s conceptual model is a moving target within the context of educational assessment, since these conceptions are changing continuously from day to day and even from minute to minute during instruction in the classroom. These changes may be in fine, continuous increments, or may be coarse-grained shifts in thinking, such as “aha” experiences. From my personal experience in tutoring learners I often observe these changes from one moment to the next as the learners grasp an element of mathematics which previously eluded them. Armed with this new knowledge they succeed on problems which a few moments earlier they were unable to solve. Thus the measurement of a learner’s constructs must consider the dynamic nature of these constructs. Any

measurements taken are only relevant at the time they are taken since the learner's conceptual model consist of continually changing schemas. Thus to implement "assessment *for* learning", as outlined in the CAPS statements (DBE 2011a), diagnostic information should be available immediately after diagnostic tests have been administered to provide the maximum benefit to instructional decisions.

Some misconceptions may be persistent and may thus span different measurement tests, but there will be a point in time when each misconception is first constructed by the learner as a tentative, intermediate schema to address a particular problem presented. Such tentative and incomplete schemas persist when they are not challenged by the right type of problems being given during instruction and through consequent remediation. Thus, identifying a learner's conceptual model requires not only identifying misconceptions arising from current instruction, but also misconceptions which persist from past learnings.

Given that diagnostic information is more useful if it is current, and that it is non-trivial to calculate, this points to the potential role that automated methods can play in calculating and providing this information to the teachers and learners. We are witnessing, year by year, increasing usage of computers in the classroom, including both dedicated computer rooms filled with desktop computers and the modern approach of providing Internet-connected tablet computers to every learner. The trend is towards increased provision of such technology to enhance education, and this is becoming a permanent element of the educational system. This trend is comparable to the introduction of the calculator in the 1980s as a mandatory tool for learners of mathematics and science. However, there are questions of whether the introduction of calculators has reduced learners' mental arithmetic, such as their ability to recall the "times tables", and a similar argument will hold for the introduction of computers in the classroom. There is universal acceptance that computers will play a positive role in the future classroom, and within my study I am concerned with the use of these computers for diagnostic assessment, as part of the daily assessment practices. I thus argue for diagnostic assessment as a good application of computers to enhance the teaching and learning of mathematics. The effectiveness of technology for diagnostic assessment is dependent on a number of pre-conditions:

- firstly, that there are test items which are suited to measuring the constructs identified within domain of mathematics for which diagnosis is being conducted;
- secondly, that these items can be presented to learners efficiently, with learners able to provide responses with minimal effort;
- thirdly, that the results can be gathered and analyzed as quickly as possible;
- fourthly, that the measured results can be communicated to the teachers and/or learners without delay, and;
- finally, that the teachers are able to make use of this information to adapt their instruction practices to what has been inferred about their learners' content-related challenges.

The considerations and options available for implementation of such automated diagnostic systems include the possibility of using the Internet effectively to provide the assessment, to process the results, and to communicate these results to those who can use this information. This approach removes the need for the resources and the effort required to install and run computer programs on separate computers or school-level servers. I envision a future where there is central repository of good diagnostic questions which is available on demand for teachers. Such a centralized repository has the potential for creating an environment in which additional benefit can be gained from the data obtained from centralized collection, providing a rich research resource to support further studies into the nature of misconceptions, and on effective methods of measurement to detect these misconceptions.

From a practical perspective, many schools in South Africa do not have access to computer technology and some do not even have electricity. For such situations manual diagnosis is possible using an approach as suggested by Wylie and Wiliam (2006) which only requires that each learner has a set of cards to hold up to show their answer to single questions posed by the teacher. This is a novel and innovative solution for technology-poor schools, and also very easy to apply. However, this approach does not provide access to the fine-grained knowledge of individual learner thinking which comes from a more comprehensive diagnostic process. It also requires careful selection of the questions to use, since ideal questions should be adapted for the specific conditions of the learners and a standard set of items may not be appropriate to elicit the right evidence. However, in

the absence of technology, this approach of Wylie and Wiliam (2006) is far better than the alternative of doing nothing.

I considered the extent to which the automation of diagnostic information, specifically using classroom-based assessment tools and access to web-based server support systems, should have been included into my study. My concern was whether automation of diagnostic assessment would add value to this inquiry, or whether this is merely a practical application that is not a knowledge-creating research activity in its own right. I did see the value of this inclusion, but I also considered the considerable growth in the scope of this study which would have rendered it impractical within the limits of time and resources for a doctoral study. I have argued above that more needs to be known about the effectiveness and efficiency of diagnosis before a technological implementation can be considered as a serious research inquiry or practically employed in the classroom. Thus, my conclusion is that studies on the usage of technology for diagnosis must be preceded by an inquiry into the nature of effective diagnostic practices. I have used the Web as my primary means for collecting diagnostic information from the learners, and something has been learned from this experience, and thus I have identified questions that warrant further inquiry concerning the usage of the Web for diagnostic assessment.

Thus, irrespective of how the results of this study may be used in the future, I need to use technology and I expect that future diagnostic assessment practices cannot be performed without a reasonable level of technological support if they are to be effectively used for dynamic, real-time support in the mathematics classroom. I thus argue that attempting diagnostic assessment without technological support cannot be effective, given the complexity of the conceptual models of the learners, and the wide range of misconceptions that occur, as well as the computational requirements to obtain valid measures.

Thus I have positioned my study as an inquiry into the nature of good diagnostic items that are suited for future automation using the Web, given that without this knowledge any attempt at automation will be subject to the well-known computer limitation of “garbage-in-garbage-out”. Without high quality diagnostic processes and practices, no level of automation and no amount of technology will help to improve teaching and learning. As a result, I have used Web-based diagnosis for this study as a tool rather than as a unit of analysis.

1.10 Research Problem

My rationale, as presented above, has been to establish why diagnostic assessment should be an integral element of formative assessment, and why it is important that the diagnostic practices are both effective and efficient for the teaching and learning of the rational numbers.

My research problem is thus concerned with the effective and efficient diagnosis of conceptual difficulties, or misconceptions, in learner thinking, within the realm of the rational numbers.

My study is not exploring how diagnostic information may improve teaching and learning of mathematics nor how such information is communicated to or used by the teachers. Rather thus study is limited to the problem of detecting and identifying misconceptions which are preventing learners from achieving success in their attack on problems in the rational numbers.

My research problem is stated as:

"How can diagnostic assessment be conducted in an effective and efficient manner to detect learner misconceptions in the rational numbers".

My scope does not include a study of the alternative methods of conducting diagnostic assessments, and is rather focused on the nature of the diagnostic test instruments and the corresponding measurement processes which are fit-for-purpose. My scope implicitly includes how the processes of diagnostic assessment may be replicated through Web-based automation. However, whether these diagnostic instruments are implemented in the classroom through paper-and-pencil tests, or are available through an online, repository-based assessment system is a secondary concern for this study and will form part of follow-on studies.

1.11 Research Questions

My research problem raises a number of issues that were identified in my rationale and motivation, and which now form the core questions that direct the course of this study.

In essence, effectiveness and efficiency in diagnosis has been equated respectively to the selection of the right diagnostic problems and to the determination of how many such problems are required to ensure validity and reliability in the diagnostic inference.

Whereas the posing of test items to learners is only one possible approach to obtaining diagnostic information, it is the only approach considered within the scope of this study. To illustrate an alternate approach, I outline Piaget's clinical interview method which is used to ascertain a learner's conceptions, as described by Posner and Gertzog (1982):

> Its [clinic interview] chief goal is to ascertain the nature and extent of an individual's knowledge about a particular domain by identifying the relevant conceptions he or she holds and the perceived relationships among those conceptions. (p. 195).
> …allowing a skilful researcher both to probe the areas of the knowledge domain of particular interest and to let the subject speak freely, while constantly checking his or her spontaneous remarks for those that will prove genuinely revealing. (p.196).

The clinical interview approach to diagnosis of conceptions requires access to a skilled researcher dedicated to each learner, which is impractical and time-consuming and thus challenges the goal of efficiency. However, the clinical interview may provide additional information arising from physical observations, such as actions and gestures, pauses, and various utterances, which are not available within the scope of formal testing processes. What is needed is a process with the same effectiveness as the clinical interview, but which is also efficient considering the availability of time and resources.

My focus is on the use of diagnostic assessments, conducted as formal tests, to elicit evidence concerning which misconceptions are used by a learner and the extent to which these are used. As a consequence of the previous argument in this chapter I have framed the following research questions:

- RQ1 (EFFECTIVENESS): How can we measure test items in terms of their fitness-for-purpose as good diagnostic instruments?
- RQ2 (EFFICIENCY): Given a particular diagnostic context, how many good diagnostic questions are sufficient to establish valid and reliable evidence?
- RQ3 (SELF-KNOWLEDGE): Does access to learner self-knowledge aid the process of diagnosis, in terms of the additional benefit for the added effort in obtaining this information?

1.12 Outline of Remainder of Thesis

Chapter 2 discusses prior work on formative and diagnostic assessment practices, rational number misconceptions, and educational measurements which collectively impact and inform this study.

Chapter 3 introduces the theoretical framework developed for this study, called the "Development Stage Model" which helps to frame the research questions and the data requirements.

Chapter 4 outlines my research approach and methodology including the nature of the pretests and the online tests used.

Chapter 5 presents my analysis of the data arising from the pretests, and Chapter 6 presents the analysis of data from the online diagnostic assessments conducted using a web-based assessment system.

Chapter 7 provides the results arising from Chapters 5 and 6, with reference to the research questions, and presents the core findings arising from this research.

Finally, Chapter 8 summarizes the outcomes of the study and offers suggestions for follow-on research work as well as on practical implementation of the results of the study.

A number of appendices provide additional details to supplement the content of the thesis.

Appendix A provides additional information on the data gathered, including the database fields used to store the data from the online assessment.

Appendix B provides the details of the pretests conducted.

Appendix C outlines the basis for the Web implementation including the MCQ structure and how user security was managed, as well how these mathematical problems are rendered in a Web browser.

Appendix D provides the item bank as used for the online tests.

Appendix E provides the structure of the four assessment lessons and how these were structured into information pages, diagnostic tests, and results pages.

Appendix F outlines how the WinSteps Rasch analysis program is used for the data analysis chapters in this thesis.

CHAPTER 2 : LITERATURE SURVEY

> *"…the road to a state of expertise is paved with errors and misconceptions" Nesher (1987).*

2.1 Introduction

My study concerns diagnostic assessment, and is focused on its application within the rational numbers. This study is thus located within the broader domain of educational measurement—addressing the need to measure a learner's ability, or some other construct of interest, within the target domain.

Diagnostic assessment is a special kind of assessment, requiring a more fine-grained approach to measurement, and to position my study I explore prior work in a number of related fields:

- constructivism—as a theory of learning which accounts for learner errors and misconceptions, where such errors are an integral step in the progression of learning
- rational numbers—with a focus on decimal fractions, common fractions and the number line, including the conceptual development of rational number knowledge and the identification and classification of common misconceptions in the rational numbers
- educational assessment—including summative, formative and diagnostic assessment, and with a comparison to the modern practice of Cognitive Diagnostic Assessment
- educational measurement—and particularly the Rasch method, concerning how measures are obtained from raw data
- trajectories of learning—and how these are applied within micro-domains of mathematical knowledge
- computer-based and web-based diagnostic assessment practices

For consistency in this thesis, I adopt the following definitions: "decimal numbers" are numbers which contain a decimal mark, being a "decimal point" or a "decimal comma". The decimal comma is used by countries which have adopted this element of the metric system number representation, including South Africa. A decimal

number consists of a "whole number part", followed by the decimal mark, and then followed by a "decimal fraction". I use the term "decimal number" when the entire number is being considered or where the form of the number is not specific to the context, and I use the term "decimal fraction" in cases where the fractional part alone is referenced. For this entire study I use the decimal point as the decimal mark, such as 23.456, rather than the South African educational standard which uses the decimal comma, for which the notational representation would be 23,456. This latter is potentially confusing for learners, and for almost everyone else in the country, since the decimal number 23,456 can also be read as "twenty-three thousand, four hundred and fifty-six" in different contexts.

My reason for adopting the decimal point as the standard for this study is that my primary diagnostic tests were conducted on the Web, which uses decimal numbers written using the decimal point, being the form of decimal numbers which learners are already familiar with from their use of calculators, computers, and spreadsheets. Within South Africa both of these decimal representations are used, with the decimal comma used in school-level education and in some government reporting, while business, science, and the media uses the more internationally-accepted decimal point.

2.2 Constructivism in Learning and Teaching

My first point of departure is to explore why learners make errors. I present the constructivist theory of learning which accounts for systematic learner errors, and specifically address how a learner moves from being a novice to being an expert in a specific domain of knowledge. Whereas this is a theoretical discussion, which would naturally fit within the outline of my theoretical model, the notion of misconceptions is so central to my argument that I have preferred to introduce the basis for this theory earlier in this thesis.

All followers of constructivism agree that the learners build their own knowledge, and they also accept a common assumption that meaningful learning is created by connecting new knowledge to be learned with pre-existing knowledge (Limón, 2001). Learning is seen as a "relatively permanent change in behavioural potentiality" and that "After learning, learners are capable of doing something that they could not do before

learning takes place" (Hergenhahn & Olson, 2005, p.3). Thus for learning to happen, some cognitive change is required.

Learning can be viewed as a sequence of incremental changes within the conceptual model of the learner, and these changes may be small or large. From one moment to the next, the learner's conceptual model is adapted to the needs of the situations that are presented and experienced. In the constructivist model, these adaptations are deemed to be driven by the learner's experiences of situations which do not match their current conceptual model. The process of learning occurs when the components of the learner's conceptual model—as their collection of schemas—fail and consequently trigger a cognitive conflict (Limón, 2001). These conflicts are the key drivers of the process of equilibration as outlined by Piaget (1985), which is the innate process of adaptation of the learners' world view to the external world they experience through their senses (Piaget, 1964/2003; Piaget, 1970). For example, such a cognitive conflict occurs when a learner is presented with a new form of mathematical object and its associated notations—such as the first introduction to common fractions—which does not fit into their prior knowledge of the whole numbers. For the learner, the only way to proceed successfully is to adapt his/her thinking. Such cognitive conflicts can be exploited in teaching to actively inform instruction (Limón, 2001). In such constructivist teaching the teacher modifies the instructional practices to expose cognitive conflicts and consequently trigger this natural learning process in the learners.

Piaget's model (Piaget, 1964/2003; Piaget 1970; Piaget, 1985), which is an historical basis for the theory of constructivism, postulates that every person is continually engaged in this process of equilibration, with the purpose of ensuring consistency between the external observed world and the internal world of the learner—his/her conceptual model. New observations are either "assimilated" into the learner's conceptual model, for situations in which the conceptual model contains schemas that are sufficient to consume this new input, or are "accommodated", requiring adjustments to the conceptual model to adapt to the internal-external inconsistencies.

A learner's external inputs or observations include two specific forms of communication. Firstly, mathematical examples and problems presented by teachers to the learner, with or without worked solutions, as part of instructional practices. Secondly, feedback from the teacher to the learner, where teachers respond to a learner's workings

and solutions. Such feedback may be provided at different levels of detail: as a simple mark indicating success or failure; as a worked-out explanation of the solution; or, as a detailed description of the specific errors made by a learner with guidance on how to improve. When more detailed feedback is provided by the teachers, then the learner has more inputs to help stimulate and trigger the constructivist learning processes.

However, there are two situations in which these external inputs will fail to increase learning and which may even cause learning to be negatively impacted. In the first situation, examples and problems presented by the teachers are incomplete, incorrect, or irrelevant in terms of the topic of instruction. This may create learning which is distorted and biased towards these examples, since these examples will constitute the entire external experience of the learner, thus are the only opportunity for learning. This situation was reported by Nesher (1987) in the case of comparing the magnitude of two decimal numbers, in which a random selection of pairs of numbers has a relatively small chance of helping to discriminate misconceptions. In such cases the teachers will, on the basis of tests conducted, deem the learners to be proficient even though known misconceptions have been overlooked. Thus the test results are biased by the selection of problems used in the test and a different conclusion may arise if different problems are presented. In this case, the test items used are incomplete when considering the range of learner schemas which are known to exist from prior research. This is firstly a problem of construct validity—that the test items do not match the construct being measured—and secondly is a problem of teacher knowledge; since if the teachers are not aware of the misconceptions then their concept of the construct is itself limited. In this latter case the test items may be measuring an incompletely defined construct, whereas in the former case the construct may be well-defined, but the test items are insufficient to measure this construct.

The second situation of learning failure occurs when the feedback provided by the teacher is incorrect, such as when a teacher marks a question as being incorrect even though it is correct. There is a common assumption that feedback will promote learning, but this is not necessarily always the case (Kluger & DeNisi, 1996). However, when the teachers themselves are struggling with their own pedagogical content knowledge of mathematics, then the feedback they provide may be incorrect or incomplete, and such feedback will certainly impact learning negatively, by causing the learners to construct

incorrect or incomplete schemas or models in adapting their thinking to meet the teacher's guidance. Thus the quality and accuracy of the feedback from teachers is an essential element of learning success and any incorrect or incomplete feedback is aggravated by the trust which learners place in their teachers' knowledge, so that feedback from teachers will be given priority as part of the adaptation process.

Learners begin a new topic as novices, bringing prior knowledge and pre-conceptions from other areas of knowledge in which they have previously developed some level of proficiency. Mack (1990, 1995) has examined the case of whole number knowledge which is brought into the learning of rational numbers and notes that learners may not be able to assimilate new symbolic representations with their own informal knowledge, and will attempt to generalize existing knowledge even when this knowledge is inapplicable. For example, a learner will see the fractions $\frac{4}{5}$ and $\frac{5}{6}$ as equivalent "because there's one piece missing from each" (Mack, 1995, p. 28). Thus, learners use their existing knowledge to address new situations and, since this is the only knowledge to which they have access, will adapt their thinking when told that this knowledge did not provide successful outcomes.

A learner who has reached a state of mastery in a domain will have developed a set of schemas which are sufficient to consistently succeed on problems presented within this domain. Thus the notion of domain mastery is dependent on the scope of the domains and the scope of the problems presented, and any set of schemas will suffice if they collectively are used to achieve demonstrable and consistent success.

A learner of mathematics does not move directly from being a novice to a master, but rather moves through intermediate stages of development which exist in the space between novice and mastery (Behr, Lesh, Post & Silver, 1983). When learners' schemas are applied to problems for which their schemas are suited then the learners will demonstrate proficiency. However, when these same schemas are applied to other problems they may fail to produce the expected results and are considered to be misconceptions. Thus, I consider that misconceptions are constructs which are not necessarily bad, or even incorrect, but which are not relevant or are not suited to a specific problem or domain to which they are addressed. A misconception is thus a conception which does not fit a new context.

The level of proficiency required by learners is identified by the curriculum statements (DOE, 2002; DBE, 2011a) for each type of problem within the scope of each topic. The curriculum statements do not necessarily define mastery within a topic but rather sufficiency for each level of education. However, whereas the curriculum indicates the required proficiency, it does not, and most likely cannot, state the nature of the schemas that the learners should construct to meet this proficiency. The conceptual model of an individual learner may be a single super-schema in which all mathematical knowledge is structured, or may be a set of special-purpose, semi-independent schemas which are triggered when needed. This conceptual model may also be something between these two extreme alternatives or may be something completely different. As long as a conceptual model is a useful tool to help the learner achieve success, then the structure of this model is of less concern, especially when we limit our measurement processes to determine only whether a learner has attained proficiency. However, when inquiring as to why a learner is not proficient, it is necessary to identify and examine the specific schemas that the learner has used to identify those schemas which are incomplete, are faulty, or are incorrectly applied, so that these can be attended to. This diagnostic practice is the responsibility of the teacher.

J.P. Smith (1995) has dispelled the notion that the development of mathematical knowledge moves towards a small set of general constructs, and has shown that expertise in the rational numbers occurs as a result of the development of more specialized methods. Thus, it appears that novice learners initially develop generalized schemas in their progress towards proficiency, while expertise is evidenced by the construction of more specialized schemas applicable for more particular problem situations. This is a shift from an initial focus on effectiveness in the novice learners—just getting it right—to an increasing focus on efficiency of application in proficient learners—getting it right with the minimal effort.

Given that learners construct their knowledge on the basis of external inputs, and that mathematics teachers are largely responsible for providing these inputs in the classroom, it becomes important for teachers to provide the right problems to each of their learners to maximize learning, which stimulates the cognitive conflict identified by Limón (2001). To achieve this, teachers need a fine-grained knowledge of their individual learners' understandings (Wiliam, 2011b). This in turn requires that teachers understand

the basis for learner errors and how this information can be used to provide personalized instruction. However, learners tend to make similar mistakes, many of which can be traced to using informal and prior knowledge in new situations (Mack, 1990; Mack, 1995), and thus many errors are both commonly experienced and hence predictable as natural steps in the advancement of learning.

The analysis of learner errors has a long history of inquiry in mathematics education. In an early work, some 90 years ago, Buswell and John (1926) report on a technology which captured learners' eye movements while they were adding a column of numbers. Buswell and John's approach was to unpack learner thinking by examining where learners are looking while engaged in arithmetic tasks and as they make mistakes. The eye movements were captured onto photographic plates which were analyzed to produce diagnostic conclusions. Buswell and John analyzed the methods used by 584 students in Grades 3-6 in the fundamental arithmetic operations. Whereas prior studies at the time had focused only on the identification of the errors, specifically in terms of which questions were answered correctly and which were incorrect, Buswell and John extended this prior work to examine the methods and mental processes of the pupils and how the learners' processes contributed to the incorrect answers. The introductory chapter to Buswell and John's study details a number of cases in which operations such as subtraction are analyzed for individual learners and where the approaches used by the learners are incorrect and inefficient, demonstrating significant misunderstanding of the number system. Their objective is summed up in the following quotation:

> It seems perfectly clear that the children described in the preceding paragraphs can never be efficient in arithmetic until they abandon their erratic and wasteful methods and adopt more direct and economical means of working examples. (Buswell & John, 1926, p. 4)

This highlights the mathematical focus of the time—that proficiency in mathematics was equated with the most economical procedures, and not seen as the development of a deeper understanding on the nature of numbers, and that research did

not include the identification and monitoring of schemas in the learners' conceptual models which may account for their responses.

It was from these initial studies that the focus of research shifted away from the analysis of errors, and towards the internal learner conceptions which give rise to these errors, and this led researchers to explore the mathematical conceptions and misconceptions of learners.

2.3 Misconceptions

Confrey (1990) has defined misconceptions as errors which are grounded in a theory and articulated by research—essentially as conceptions which are in conflict with accepted meanings—and also notes that the terms "errors" and "misconceptions" are often used interchangeably. J.P. Smith et al. (1993) note that misconceptions have been studied under many alternative names such as preconceptions, alternative conceptions, naïve beliefs, alternative beliefs, alternative frameworks, and naïve theories, but that the term "misconception" is the generic term commonly used and that all of these forms concern the relationship between novice and expert conceptions. J.P. Smith et al. (1993) also note that misconceptions originate in prior learning, by attempts to generalize knowledge. This is a natural part of constructivist learning, seen as experimentation and adaption, which gives rise to the construction of new schemas and conceptions with the goal of equilibrating individual experience with conceptual models, and consequently improving success.

Misconceptions are considered as the application of informal knowledge and prior learning to a new domain in which they do not fit properly (Nesher, 1987; Confrey, 1990; J.P. Smith et al., 1993). Mack (1990, 1995) suggests that learners overgeneralize informal knowledge, such as in the misunderstanding of the symbolic notations of common fractions by applying whole number knowledge. Resnick et al. (1989) conclude that learners commonly invent rules in the course of learning, in cases where the use of earlier concepts in a new domain results in a lack of success in new problems, and this causes a cognitive conflict.

Misconceptions are thus part of the totality of conceptions or schemas within a learner's conceptual model, and both misconceptions and other conceptions can exist side-by-side with each other (J.P. Smith et al., 1993). Thus it is possible to consider a

learner's conceptual model as a set of schemas of different levels of proven utility to specific problems, where the learners select the most appropriate schema when presented with a new problem. When problems of increasing complexity are provided by a teacher, schemas applicable to simpler examples will not be effective for more complex problems and will fail. Over time, and with a suitable set of problems coupled with appropriate feedback, the schemas which survive are those that are more successful and are given priority in future tasks. Given that we cannot observe learners' schemas directly, we must infer the existence and nature of these schemas from our observations of learner responses.

My exploration of prior work on misconceptions has shown that researchers are in general agreement on common attributes of misconceptions including that: they are theory-based, with cognitive explanations for their formation and usage; they arise from informal knowledge, from prior learning, or from overgeneralization; they co-exist with other conceptions; and, that learner development of misconceptions is an element of learning.

I have earlier indicated that for this study I consider a learner's mathematical understanding as a conceptual model which is composed of schemas which may themselves be connected or independent, and which may be small and special-purpose or large and general-purpose. These schemas are the only conceptual tools which a learner has available to address mathematical problems. In effect, I assume that all mathematical knowledge exists only in the form of schemas. All schemas are in a continuous process of adaptation towards the goal of consistent success on problems, and a schema which is stable and effective when applied to one type of a problem may prove ineffective, and thus be a misconception, when applied to more complex problems. Misconceptions are thus intermediate constructs created by the learner, in response to various inputs, and which are constructed on their path to mastery. Thus I argue that there is no essential difference between misconceptions and other conceptions, and all form part of the conceptual model of a learner. Therefore, in examining the prior work in this field, I consider that the concept of "misconception" is a categorization of a conception which is applicable when a schema is not fit for the purpose of a specific problem whereas it may have been used successfully for prior problems.

2.4 Issues concerning Rational Number Misconceptions

I now examine prior work on misconceptions in mathematics, and specifically within the rational numbers. In this section I firstly introduce common issues that span the rational numbers, and in the next section I discuss misconceptions in the different types of rational numbers.

Common misconceptions

One commonly-cited misconception is that "Multiplication Makes Bigger (and Division Makes Smaller)". Bell et al. (1981) examined the choices made by learners for whether a given verbal problem should be treated as a multiplication or as a division, and found that there were consistent mistakes which pointed to underlying learner assumptions of the nature of multiplication as compared to division. This misconception is also addressed by Feischbein, Deri, Nello and Marino (1985) and two of the problems they present are (PROBLEM 3) "From 1 quintal[2] of wheat you get 0.75 quintal of flour. How much flour do you get from 15 quintals of wheat?" and (PROBLEM 4) "The volume of 1 quintal of gypsum is 15 cm^3. What is the volume of 0.75 quintal?" (p.10). Both problems use the numeric values 15 and 0.75 and yet the results differ significantly. For PROBLEM 3 the success rate was 79% for Grade 5 and 74% for Grade 7, whereas for PROBLEM 4 the success rate was 57% for both grades. For both of these problems the learners were asked to select whether the right operation to apply is multiplication (the correct answer) or division (the incorrect answer). Thus the learners found PROBLEM 4 significantly more difficult than PROBLEM 3 when measured by these success rates. Feischbein et al. (1985) conclude that the two numbers used in each of the above problems have specific roles of "operator" (how many) and "operand" (of what) and that the tacit knowledge of "multiplication as repeated addition" does not work in cases where the operator is not a whole number.

Another commonly cited misconception occurs within the topic of common fraction addition. Behr et al. (1983) note that only one-third of 13-year-old learners could correctly perform the sum $\frac{1}{2} + \frac{1}{3}$, since the learners' prior knowledge of whole number

[2] A quintal is an historic unit of measurement which is 100 times a base unit, such as 100lb or 100kg.

addition causes them to see this as two separate additions for the numerator and denominator, resulting in the answer $\frac{2}{5}$ by many learners.

Issues of persistence

However, learners are not like computers which can be quickly reprogrammed when a bug is detected. The schemas constructed by the learners are personal conceptual tools which have been developed and refined, often over an extended period, to address some particular problem or to explain some external experience. When schemas are successful they become increasingly resistant to change (J.P. Smith, 1995; Nesher, 1987). When schemas are shown to be insufficient, remediation is required over an extended period since the schemas cannot adapt themselves on the basis of single, isolated examples. I contend that schemas are tools, and are constructed for a purpose, and that the process of schema construction involves the self-preservation of schemas with proven utility. Thus those schemas which are more successful are given priority, and they become persistent and thus naturally resistant to change once they have passed from short-term memory to the fixity of long-term memory. As a thought experiment, consider the alternative that schemas in the long-term memory could change rapidly, which would result in a lack of stability where schemas are changed 'willy-nilly' to accommodate every new situation. The result will be schemas which are only applicable to the latest experience, and for which older experiences and their schemas are discarded. Persistence is thus an important attribute of human learning, and it is essential that better schemas are persisted, with the corresponding need to ensure that incorrect or incomplete schemas continue to adapt before becoming persistent.

Whereas immediate learning occurs in the short-term memory of the learner, this is often insufficient to influence the long-term memory in which the persistent schemas are stored for future access. This argument was made around 80 years ago by Cooke (1931) who stated that there is no data available as to how long learner errors will remain remediated after the interventions, thus implicitly recognizing the persistence and strength of the misconceptions, even though the term "misconception" was not in common usage at that time. Cooke's follow up study (Cooke, 1932) identified that 80% of the errors he noted earlier did not remain following the remedial practices that he introduced, which then implies that 20% of these errors did remain.

The development of rational number concepts

The domain of rational number learning is a perennial challenge that is experienced by each new generation of learners (Behr et al., 1983). Rational numbers remain one of the most complex and difficult areas of elementary mathematics (J.P. Smith, 1995; Bart, Post, Behr & Lesh, 1994; Kilpatrick et al. 2001) and learning problems associated with the rational numbers have been researched for at least 100 years. However, consider that this is only the most recent 100 years of at least 4000 years in the development of rational number concepts. Ancient surviving records show that the Babylonians worked with base-60 fractional numbers, which remain with us today in our angular measures of degrees, minutes, and seconds, and in our time units of minutes and seconds (Kieren, 1976). The historical development of the rational numbers was wrought with difficulties and problems of acceptance, and the historian D.E. Smith (1958) states that the Romans avoided fractions to the extent that they developed a multitude of names for parts so that they would be relieved from having to represent and to undertake computations with fractions. D.E. Smith (1958) also states that the development of our current notational system of decimal fractions took hundreds of years to complete, with many competing methods and notations which were proposed during this long developmental history.

Multiple representations of rational numbers

The evolution of rational number notations and representation has led to our current eclectic mix of notations and representations, including common fractions, decimal numbers, number lines, percentages, ratio notation, and geometric diagrams, all of which are encountered both inside and outside of the school environment (Kieren, 1976). Whereas the whole numbers are used primarily for counting, the rational numbers have a variety of uses and interpretations. Kieren argues that understanding of the rational number requires “adequate experience with their many interpretations” (p. 102) and he identified seven specific interpretations: fractions; decimal numbers; equivalence classes; ratios; operators; quotients; and, measures on a number line. Kieren’s structure for the interpretation of rational numbers has been referenced and refined by many other authors over the intervening years, including Behr et al. (1983) who have used Kieren’s interpretations as a basis for the Rational Number Project (2014). Kilpatrick et al. (2001) note that Kieren’s classification has guided research in rational numbers for two decades.

Prior work shows that the domain of the rational numbers cannot easily be viewed as a single unified concept but is better viewed as a mix of notations and representations, each with a variety of interpretations and applications. Kieren (1976) has identified the need for the various fraction concepts to be understood both individually and as connected to each other in "isomorphic" interpretations. In his original seven interpretations, Kieren does not include verbal fraction representations, such as the term "three-quarters", and also excludes percentages as a specific form. These various forms of fractions and their notations have been questioned by Usiskin (1979), concerning whether common fractions still have a place in the curriculum given that calculators work suitably with decimal numbers alone. Usiskin concludes that different representations each have applications to which they are better suited and that calls for the removal of common fractions from the curriculum, in favour of a unified model of decimal numbers, are unlikely to succeed.

Given the multiple representations and interpretations of the rational numbers, I examine how the South African curriculum views the scope and development trajectory of this knowledge. The curriculum statements (DOE 2002; DBE, 2011a) show the development of fractional concepts from the earliest grades, leading to an initial understanding of common fractions and their notations, which are introduced prior to decimal numbers, percentage, and other forms. Thus, the specific forms of rational numbers are introduced in isolation from each other, but as the decimal numbers are introduced in Grade 6 (DBE, 2011a), the equivalence of decimal numbers with common fractions and percentages is a required proficiency. Learners commence their knowledge of rational numbers from the relative simplicity of the natural numbers, which is linked with concepts of parts and the notion of equal sharing in the early grades. The operations of division by 10, 100, and 1000 are then used to introduce fractions with these denominators, which leads to the notation of decimal numbers. From this point learners confront multiple representations, together with unique characteristics of rational numbers, such as density and the equivalence of different forms. Arithmetic operations such as addition are handled differently for common fractions, although for the decimal numbers much of the existing whole number knowledge is still applicable. Given this additional complexity in both representation and operations of the rational numbers, it is evident, purely on the basis of this analysis of the curriculum requirements, that learners will struggle to develop a consistent and successful conceptual understanding of these

numbers. This is aggravated by my earlier observation that the curriculum only allocates 40 hours to the common fractions and decimal fractions for the entire three-year Senior Phase of Grades 7-9.

The diagnosis of rational number difficulties

Early work on the diagnosis of arithmetic problems of learners, including problems in common fractions and decimal numbers, focused on identifying the types of problems that learners struggled with and with which frequency (Brueckner 1928a, 1928b), and one of the early approaches recommended a detailed analysis of the methods used by learners through observing their work (Uhl, 1917). Kieren (1976) cites Fish (1874) concerning the exclusive focus on mathematical operations and calculations required to solve these types of problems while lacking a focus on the conceptual understanding of fractions.

The study of learner errors in arithmetical operations, including decimal and common fractions, has been a continuously active area of educational inquiry for almost 200 years and highlights include the early diagnostic studies of De Morgan (1831/1898/2013), Buswell & John (1926), Brueckner (1928a, 1928b), Olander (1933), Sprague (1939), and Guiler (1945). In particular, the work of Olander (1933) is the earliest account I have found which considers that learner mistakes have a systematic basis. The work on rational number learning and misconceptions also includes: the extensive volume of studies published by the Rational Number Group (such as Bart et al., 1994, Behr et al., 1983); the work of Steinle (2004a) in categorizing learner errors in decimal ordering; and new theories of understanding learner errors by Siegler, Thompson and Schneider (2011). The focus on understanding and making sense of rational number errors has always been an issue for research in mathematics education.

The emphasis on procedural knowledge, especially in the rational numbers, has been highlighted by the Rational Number Project (2014) researchers (Behr et al., 1983; Bart et al., 1994), who identify the curricula focus on procedures and methods as being at fault, in that it fails to drive the teacher towards developing a deeper understanding of learner conceptions that give rise to the misuse of procedures. Behr et al. (1983) state that exposure to all interpretations and all forms of rational numbers is necessary to gain a full understanding and to establish competency.

Given my concern in this study for understanding the types of test items which are best suited for diagnosis, I examine how South African learners responded to two

specific test items in the TIMSS 1999 study (Mullis et al., 2000; NCES, 2015), both of which involve the ordering of a set of rational numbers. The first is item B10, which is to select the smallest of five decimal numbers being (1) 0.675 (2) 0.5 (3) 0.375 (4) 0.25 and (5) 0.125. For this item South African learners obtained an overall score of 6%, which is below the expected level of 20% if these learners had been randomly guessing and which, as for the examples presented in Chapter 1, reveals information about the conceptual development of the learners, inferring that guessing alone is highly unlikely to produce such a low result. The second item is D09, which asks the learners to select the smallest of four common fractions, $\frac{1}{6}, \frac{2}{3}, \frac{1}{3}$, and $\frac{1}{2}$, for which the South African learners, still close to the bottom of the list of countries, scored 28%, where this result may be explained on the basis of learner guessing which would result in an expected value of 25%. Thus these results show that item B10 provides far more evidence of the use of misconceptions than item D09, even though D09 has a higher relative success score. Whereas both of these items appear to have the potential to elicit similar misconceptions, B10 would be preferred for diagnostic purposes given its ability to trigger misconception schemas in the learners' responses.

The National Research Council (NRC, 2005) report on how students learn mathematics in the classroom, introducing three principles which can guide effective teaching and learning. Principle #1 "Engaging Prior Understandings" is described as "new understandings are constructed on a foundation of existing understandings and experiences" (p. 4), and this foregrounds the need to understand a learner's knowledge to ensure successful teaching, including what each learner knows on entry to each new topic. Chapter 7 of this NRC report explores new approaches to teaching rational numbers, and cites examples of research on various preconceptions and misconceptions which arise in the learning of rational numbers.

Stacey (2013) notes that the history of research into mathematics learning dates back to the early 1900s, with a "cognitive turn" from the middle of the 20th century onwards, and that this remains an active area for research, as suggested by 20% of the 2012 MERGA (Mathematics Education Research Group of Australia) conference papers being devoted to such research.

On the basis of the above general discussion of misconceptions in the rational numbers I now examine prior work on particular types of rational numbers as are relevant to my study.

2.5 Misconceptions by Type of Rational Number

Decimal Number Misconceptions

Misconceptions in the decimal numbers have been studied more than other topics in the rational numbers, and work can be traced back to the error analyses conducted by Brueckner, who notes that "the computations in the tests were kept as simple as possible so that errors due to faulty handling of decimal numbers would be revealed rather than errors due to difficult computation" (Brueckner, 1928b, p. 36). Brueckner thus asserts that more complex problems presented to the learners may cause errors resulting from the challenges in the computations which then reduces their effectiveness as diagnostic instruments. Brueckner's approach did not explore the cognitive causes of mistakes but rather limited his analysis to the identification of which steps in the standard procedures for computation were the cause of the observed problems. I concur with Brueckner that diagnostic test items should be as simple and as special-purpose as possible.

Hiebert and Wearne (1985) explored errors in decimal number addition and subtraction problems, and predicted the types of errors made by their students, such as the sum 5.1 + .46 resulting in a response of .97 or 9.7 rather than the correct response of 5.56. They conducted tests on learners in Grades 5, 6, 7, and 9 and found that the errors which they had predicted were the most frequently observed and accounted for 75% of the total set of errors. Hiebert and Wearne's hypothesis is that learners rely on learned methods—or "syntactic" methods—as distinct from conceptual understanding—or "semantics"—and thus their learners were likely to be using memorized rules rather than employing a deeper understanding of decimal numbers.

Wearne and Hiebert (1988) proposed a theory for how learners develop proficiency on the written symbols and notations of decimal numbers. Their theory is used to predict responses by learners to specific test items which cover place-value, decimal number addition, and decimal number ordering, and the theory considers a "connecting" process which predicts that meanings of the written decimal numbers will be developed gradually. Wearne and Hiebert illustrate this progression for the case of a

learner "Bonnie" who made various errors during the process of instruction in the decimal numbers. Bonnie initially selected .42 as larger than .5 by explaining that 42 is bigger than 5, and later explained that .8 is larger than .34 since "eight tenths is more than three tenths", and finally that 0.056 is larger than 0.05 since the "the five and the six is more than five hundredths…because it got five hundredths and six thousandths and that only got five hundredths." (Wearne & Hiebert, 1988, p. 378). These examples highlight a trajectory of misconceptions that occur through the development of proficiency at progressive stages of development.

Sackur-Grisvard and Leonard (1985) observed that 89% of the errors made by learners when comparing the magnitude of decimal numbers are accounted for by three rules used by learners. Rule 1 occurs when the decimal number is treated as though it is a whole number, which is in line with Wearne and Hiebert's (1988) example above of Bonnie when comparing .42 and .5. Rule 2 selects a number as smaller if it has more digits in its decimal fraction—thus both 12.94 and 12.24, with two decimal digits, will be considered as less than 12.7, which has only a single decimal digit. Rule 3 selects decimal fractions which start with a zero digit as being smaller. Sackur-Grisvard & Leonard (1985) conclude that Rule 3 appears later in conceptual development than Rules 1 and 2, and note that teachers and textbooks rarely use such comparison tasks in the classroom, since teachers tend to avoid introducing problems which are too difficult, so that they can produce higher success rates among their learners. For example, applying Rule 1 to decimal fractions with the same number of decimal fraction digits will always produce a correct response, so that Rule 1 only shows itself as a misconception when the number of digits in the two numbers are different. Sackur-Grisvard & Leonard also suggest that learners may achieve success when ordering two decimal numbers, but that incorrect rules may be exposed when using more than two decimal numbers. Rule 2 occurred with less frequency than Rule 1, and in all items either of these rules may be applied by the learners.

The basis for Sackur-Grisvard & Leonard's Rule 2 was also reported by Grossman (1983), on a large-scale study of more than 7000 applicants concerning a baseline assessment for placement at the City University of New York, observing that the highest percentage of students, when asked to find the smallest of five decimal numbers, selected the choice with the most digits to the right of the decimal point. Grossman also observed that these questions were the most difficult of all of the items in the test, as measured on

the basis of the proportion of the students who answered the question correctly, and that these items were more difficult for these applicants than all of the test items that involved operations of addition, subtraction and multiplication on decimal numbers. Thus the expectation that items which require more complex computations are naturally more difficult that those involving only conceptual understanding is misplaced. Grossman recommends that this type of question should be incorporated into schooling, and that this action could have the result of making these items on the comparison of magnitudes of numbers the easiest, rather than their being the most difficult.

Resnick et al. (1989) extended the study of Sackur-Grisvard and Leonard (1985) to explore the conceptual development of the learners, and suggest that "A child who has just been exposed to instruction on decimal numbers must build a representation of decimal numbers and related decimal numbers to other well or partially acquired number systems" (p.10). These "other number systems" include the domain of the whole numbers and the domain of measurement, both of which are introduced prior to the decimal numbers. Resnick et al (1989) refer to Sackur-Grisvard and Leonard's (1985) Rule 1 as the *whole number rule*, and they refer to Rule 2 as the *fraction rule* since it emulates common fraction reasoning. Consequently, Rule 3 is called the *zero rule*, and is a special case of the *whole number rule*. Resnick et al. (1989) also provide a comprehensive comparison of decimal number knowledge to both whole number knowledge and to common fraction knowledge, examining what is similar and what is different concerning the values, names, notations, and reading rules. This qualitative analysis provides the basis for predicting the areas which may cause confusion if prior knowledge of whole numbers or common fractions is applied to decimal numbers. The data gathered includes not only the results from tasks such as finding the smallest of 0.5 and 0.25, but also the verbal explanations of the learners concerning their decision processes. These qualitative analyses help to validate the predictive model of the rules.

Resnick et al. (1989) also found that some learners lacked a knowledge of place-value, with learners who are using the *whole number rule* and the *zero rules* mostly being unable to identify the place-value of the digit 5 in the numbers 1.54 and 2.45. An explanation for the 2.45 error is the misconception of identifying the digit 5 as being in the 'units' columns—caused by seeing the entire decimal number as though it is a whole number and effectively ignoring the decimal mark. Resnick et al. conclude that there is a

progressive usage of these rules in the learning trajectory of decimal number knowledge, with the whole number rule appearing early in learning. Their results show a high usage of these rules, accounting for 88% of the errors observed. They also conclude that such errorful rules are "intrinsic to learning" and "cannot be avoided in instruction" and are "best regarded as useful diagnostic tools for instructors" (Resnick et al., 1989, p.26).

Stacy, Helme & Steinle (2001) use the mirror as a conceptual metaphor to explain the common confusions experienced by both school students and by teacher education students between fractions, decimal numbers and negative numbers. They argue that the mirror metaphor is used in the psychological construction of these numbers and helps to explain why learners will see decimal fractions as negative numbers, such as 0.5 being understood as negative.

Steinle (2004a, 2004b) reports on a large-scale, longitudinal study on decimal misconceptions which was conducted with more than 3,000 students who jointly completed nearly 10,000 tests between 1995 and 1999 in Melbourne, Australia. The data was used as the basis to explore decimal misconceptions in more detail than is available with smaller studies, with a goal to better understand the nature of the misconceptions. Teachers can benefit from being aware of their learners' misconceptions to help to reduce the observed persistence of these misconceptions. Steinle (2004b) identified two primary behaviours, labelled as L (Longer-Is-Larger), and S (Short-Is-Larger), which are differentiated on the basis of how many digits are present after the decimal point in the two decimal numbers which are being compared on each test item. Steinle notes two other high-level behaviours being A (correct responses), and U (unknown). Steinle (2004a) explores the L and S behaviours in more depth with fine-grained analysis of the data sets which leads to additional classification within these L and S behaviours. Learners who respond with a majority of correct responses are placed into the A category, using a rule which allows for a single mistake in each of her Type 1 and Type 2 items, where these items are designed so that Type 1 would be answered correctly by learners with S behaviour and incorrectly by learners with L behaviour, and vice versa for Type 2. Learners whose responses do not exhibit sufficient consistency in the usage of L or S behaviours are categorized under the U behaviour. I make use of Steinle's L and S classification structure as part of my methodology in Chapter 4 and I also use these codes when analyzing my results from the decimal number ordering problems. Whereas

Steinle's identification of proficient learners, in her A behaviour, is based upon at most a single item incorrect for each type of item, I see the potential for a more formalized approach to measured proficiency, and I explore this within my study.

Steinle (2004b) reports that most learners who reach proficiency in the A behaviour in one cycle of assessment were not classified as either L or S in the previous cycle, concluding that they were probably using a combination of these two behaviours and were experimenting with their conceptual models. When considering these behaviours in conjunction with the range of intermediate conceptual models between novice and mastery that I discussed earlier, Steinle's conclusions on the changes in behaviour over time provide support for the nature of the intermediate conceptual models that lie at different points on this learning trajectory. Thus it is not only our understanding of these misconceptions which is important but also the potential for identifying the natural sequence of which misconceptions are likely to occur at the different stages of conceptual development.

These prior results provide support for my rationale for this study, and I question why misconceptions often persist for long periods and why they are not remediated more effectively through instruction. Is it because the wrong examples and problems are used in the instruction process, being posed at the class-level rather than at the individual learner-level and thus not meeting the requirements for constructivist teaching and learning? Or perhaps, as cited earlier, teachers prefer not to use "difficult" questions which may take up class time. It is expected that for instruction to be more effective, the instruction needs to be more customized and targeted to the needs of the individual learners. This argument is used by Griffin (2009) who states "…teachers need to have expertise in developmental assessment because it is integral to the formulation of personalized learning plans" (p. 181). I contend that such developmental assessment will consist largely of diagnostic assessments, to help to understand the current state of the learners' conceptions, since without such fine-grained knowledge of the learner, no such personalized learning plans are possible.

Common Fraction Misconceptions

Whereas, in the real world, decimal numbers are used to represent a magnitude or measured quantity, and generally have a single interpretation as a number, as one of Kieren's (1976) seven interpretations, the same cannot be said for the common fractions,

which have a number of different interpretations and conceptions, accounting for the other six of Kieren's structures.

The common fractions predate decimal fractions by at least 2000 years, with unit fractions used in ancient Egypt as the means of expressing parts (Gillings, 1982). The Egyptians did not develop fractional notations more specialized than unit fractions, which consist of fractions in which the numerator is 1, and which represent the aliquot parts of a whole. All other fractions were represented as the sum of a set of unit fractions, with the single exception of a specific notation provided for the common fraction $\frac{2}{3}$. I have previously cited the Babylonian's use of base-60 fractional concepts and how this has survived in our present time and angular measurement systems and the number 60 was chosen not by accident, but because of its nature as an abundant number—having a large number of factors—which provide for splitting into equal parts in many ways.

Unit fractions are used for the representation of equipartitioning—which is the primary concept used as the conceptual basis for the development of the fraction concept (Confrey, Rupp, Maloney, & Nguyen, n.d.). The task of equal sharing is first introduced in the Foundation Phase (DBE, 2011a) and leads to the notations of unit fractions such as $\frac{1}{3}$, which leads in Grade 3 to non-unit fractions such as $\frac{2}{3}$ and $\frac{3}{4}$. Thus common fraction notations are introduced around two years before decimal fractions in the South African curriculum.

Hiebert and Tonnessen (1978) explored the development of the fraction concept in a small group of young children between the ages of five and nine, using Piaget's model of seven interrelated constructs. They cite Kieren's (1976) interpretations of fractions as part-whole, ratio, quotient, or as a multiplicative operator, and conclude that whereas the seven Piagetian constructs are applicable to continuous quantities, such as cutting up a pie into parts, the part-whole construct is the only construct applicable to discrete quantities. They claim that Piaget's model can be used for diagnostic purposes of a child's conception of fractions.

Mack (1990, 1995) addressed the role of a learner's rich store of informal knowledge, such as partitioning and equivalence, in accounting for errors in fractions work. She draws on a wealth of prior research on fraction misconceptions, which indicate that rote learning of procedures often gives rise to misconceptions when these procedures are applied to different forms of problems. The informal knowledge associated with

whole numbers supports a learner's explanation for why $\frac{1}{8}$ is seen as larger in magnitude than $\frac{1}{6}$.

The understanding of the magnitude of fractions has been addressed by Stafylidou and Vosnaidou (2004) using the theory of conceptual change, concluding that students form synthetic models which reveal their misconceptions. These models are considered as intermediate conceptions that result from an attempt to make sense of the fractions, and they form a part of the continuously evolving conceptual model of schemas that I have referred to earlier. Stafylidou and Vosnaidou use an explanatory framework of a "Fraction as Two Independent Natural Numbers" in which both the numerator and denominator of a fraction are treated as distinct whole numbers. Using this framework, comparisons of the numeric values of common fractions are seen by the learners as being equivalent to the problem of comparing the numerators or denominators individually. This framework gives rise to a number of misconceptions such as "the value of a fraction increases when the numbers that comprise it increase", and "the value of a fraction decreases when the numbers that comprise it decrease" (Stafylidou & Vosnaidou, 2004, Table 3, p. 509), which applies for both numerators and denominators. As for Mack (1995), Stafylidou and Vosnaidou (2004) conclude that these misconceptions are derived from the conceptual development of the learners as they attempt to reconcile the fraction concepts with their prior knowledge and that this is a natural part of the learning of fractions.

Kilpatrick et al. (2001) outline the problems of multiple representations in the rational numbers, citing the problem of the equivalence of $\frac{3}{5}$, $\frac{12}{20}$, 0.6, 0.60, and 60%, and state that this presents a significant challenge for learners who are starting to learn the concepts of both common fractions and decimal numbers, given the large number of representations which learners are exposed to, and which they are required to understand and to compare. Kilpatrick et al. cite examples of common misconceptions from the rational numbers, where each example has diagnostic qualities which discriminate between learners who use a misconception and those who do not. One example they cite is that when asking learners to estimate the value of $\frac{12}{13} + \frac{7}{8}$ from the list of choices 1, 2, 19, 21, and 40, about one half of the learners selected either 19 or 21 as the answer, reflecting the use of whole number knowledge rather than common fraction knowledge.

I contend that learners who have reached mastery in the common fractions will have either constructed a single generic representation from which the various notations and interpretations are specific realizations; or they will have constructed specific representations for each notation, with a means of interchange between these forms. As cited earlier, this latter case is suggested by J.P. Smith (1995) in considering mastery in the rational numbers, by the creation of increasingly sophisticated representations which are suited to finer types of problem situation. The argument for a single general and integrated model for fraction understanding has been proposed by Vergnaud (1994) in the "multiplicative conceptual field" which shows that a relatively simple proportional representation can account for many interpretations of the common fractions based on the types of problems which are encountered in practice. Vergnaud (2009) later presented a theory of conceptual fields as a development theory which has as an aim "to describe and analyze the progressive complexity, on a long- and medium-term basis, of the mathematical competencies that students develop inside and outside school" (p. 83). Vergnaud's theory is generalized beyond the original focus on proportional representation and rational numbers and he employs Piaget's schemas as the basis for a model of mathematical understanding. Long, Dunne and Craig (2010) use the multiplicative conceptual field to model and to measure learner competence using the Rasch method and they recommend this approach as classroom practice.

More recently, Siegler et al. (2011) have proposed an integrated theory of whole number and fraction development, based upon the consideration of the magnitude of the numbers as the key unifying element of number knowledge. Their claim is that there is a strong relationship between learners' understanding of the magnitude of fractions and further proficiency in the rational numbers. In particular, Siegler et al. (2011) dispute the claim that systematic errors or misconceptions account for the majority of learner errors, and conclude that whereas systematic errors do account for some of the errors, that many other errors can be accounted for under their theory of the conceptions associated with the magnitudes of whole numbers and fractions.

Within the large body of work on errors and misconceptions in the common fractions, many of the examples cited by different studies are either identical or similar to one another, and I question whether these are the only examples of such behaviours, or

whether these represent classes of such behaviours and are rather viewed as specific examples of more generic templates.

Number Line Misconceptions

Novillis-Larson (1980) considers the number line as one type of "semi-concrete" representations that are used by educators to teach fractions, with other representations of this type being sets and geometric regions. She reports that number lines are more difficult to grasp than other representations and found that success rates and systematic errors differ between number lines which are labelled with the scale of 0-1 and those number lines labelled on the scale of 0-1-2. I have created this notational representation "0-1" and "0-1-2" to indicate the structure of a number line which has the standard horizontal line and with markings at specific points, so that 0-1 indicates that there are two marks on the line, for the 0 and the 1.

Bright, Behr, Post, and Wachsmuth (1988) claim that the number line is more complex than other representations since it has two aspects, being firstly the visual, geometric form of the line itself, and secondly the symbolic encoding of the numbers at specific points on a number line. Other representational forms for rational numbers are either purely symbolic, such as common fractions and decimal numbers, or purely visual such as sets and geometric regions. Bright et al. (1988) state that proficiency in the rational numbers requires the ability to translate between different representations, as was previously indicated by Kieren (1976). Bright et al. (1988) conducted an experiment in which learners were asked to position fractions on an empty number line, and they provide examples of a learner who marked a 0-1 number line incorrectly from left to right as $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4}$, and another as as $\frac{1}{3}$, $\frac{1}{2}$, and $\frac{1}{3}$. In the former case the learner is applying the whole number sequence 2, 3, 4 to the common fractions on the number line and in the latter the second $\frac{1}{3}$ fraction should have been $\frac{2}{3}$, but was written incorrectly by the learner. Bright et al.(1988) report that whereas learners develop an understanding with unit fractions they struggle to develop an understanding of non-unit fractions.

Pearn and Stephens (2007) developed a protocol to probe learners on their fractional concepts using a number line, exploring the connection between whole number knowledge and fractional knowledge. They used whole number knowledge as a preconception applied to fraction learning, as had been addressed earlier by Mack (1990).

One misconception which Pearn and Stephens identified was the misunderstanding between the number of ticks between the numeric labels on a number of lines and the number of parts into which the number line segment has been divided.

However, Ni (2000) has challenged the use of the number line to measure fractional knowledge, indicating that whereas proficiency in the number line implies proficiency in fractions, the reverse is not necessarily the case. Ni considers number lines as analogues to the fractions, but also notes that these analogues are not always more abstract than the symbols which they are analogous to, thus they do not provide a clear bridge between the concepts and the different representations.

Siegler et al. (2011) make use of number line estimation test items, using both a 0-1 number line and a 0-1-2-3-4-5 number line for larger fractions. They asked the learners, using a computer-based interactive test, to position common fractions onto the number lines based on the learner's conception of the magnitude of the fractions. These results were then assessed as predictors of the proficiency of the learners, where the proficiency had been earlier established on the state-wide tests. In this case the 0-1 number line results accounted for 41% of the variance in the state-wide tests, with the 0-1-2-3-4-5 number line results accounting for an additional 34% of the variance. This correlation can be paraphrased that if a learner has insufficient knowledge about the magnitude of fractions to be able to accurately position the fractions on the number line, then they will also be unable to achieve success on other rational number problems.

In this section I have outlined the nature of conceptions and misconception in decimal numbers and common fractions in the number line, and have identified some examples of test items which elicit evidence of these misconceptions. However, to be practically useful, this knowledge of misconceptions must find its way into the development of effective measurement processes, so that these can be embedded into assessment practices.

2.6 Educational Assessment and Measurement

Measurement vs. Assessment

Moss (1994) states that "Ultimately, the purpose of educational assessment is to improve teaching and learning" (p. 10). However, the terms "assessment" and "measurement" are both used in education, and often interchangeably, when discussing the context of the

discovery of learner abilities, and the nature of the information provided to various stakeholders.

Griffin (2009), citing Griffin & Fox (1990), clarifies the distinction between assessment and measurement by defining assessment "[…] as the process of observing, interpreting and making decisions about learning and intervention, whereas measurement was regarded as the process of assigning number to observations" (p. 184). For my purposes I needed both measurements, in terms of the extent to which learners were using misconceptions, and assessment, to ensure that such measures are fit for their purpose in making inferences which can support teachers in their daily teaching practice.

I thus treat assessment as more qualitative in nature and as purpose-driven, with each form of assessment having a distinct purpose, providing a benefit to specific stakeholders which would not be realized in the absence of this information. The curriculum statement for Mathematics (DBE, 2011a) indicates that teachers should apply four types of assessment: baseline, diagnostic, formative, and summative, and this curriculum statement also makes a distinction between "informal or daily assessment" and "formal assessment"—as I have cited earlier. I summarize the stated purpose of each of these types of assessment as outlined in the curriculum statement:

- baseline assessment—are used to ensure preconditions are met before entry into a new topic of study
- formative assessment—described as *assessment for learning,* is integrated into instructional practices in various ways
- diagnostic assessment—is concerned with discovery of specific challenges to learning, including both conceptual and non-conceptual challenges
- summative assessments—are used by teachers and schools at the end of a period of study to select learners for promotion

In contrast to assessment, I treat measurement as more quantitative in nature, implying a sense of exactness, such as the analogy to using a ruler to measure length and using a thermometer to measure temperature, which both yield numbers. This analogy between physical measurement and educational measurement has been used extensively in the Rasch literature to envision the ideal of educational measurement as a universal measurement unit which can be used invariantly throughout all times and places (Bond & Fox, 2012). For my purposes, educational measurement has the aim of delivering

numbers which represent the quantity of some construct which exists within an individual learner, so that these measures can be used for some assessment purpose. This reflects Matter's (2009) two paradigms of

> *measuring* how much of a certain quantity (single underlying dimension) is evidenced in the student responses; and *judging* what the evidence says about what the student has learnt and how well. (p.222, italics in original)

Both of these paradigms are of concern for classroom-based assessment practices, since it is required that not only are measurements taken, but that these are then assessed to determine what can be inferred from the measurement to guide action to improve learning. There is no benefit in measuring when the measured data remains unused, and I also contend that if these measurements are not used to the full, providing support to the needs of all potential stakeholders, then the resources and efforts in taking the measurements have been partly or totally wasted.

This argument hints that assessment tests should be developed as multi-purpose instruments, so that systemic assessments should then provide information to support individual learning, even those having different primary purposes. Dunne, Long, Craig and Venter (2012) have explored the assessment models which support classroom assessment practices, providing aggregated information for decision-making, as well as exploring how systemic assessments may inform teaching and learning. Dunne et al. argue for the use of Rasch measurement theory (RMT) to provide improved information on learner proficiency, coupled with an assessment practice based on mathematical development trajectories, and that the use of systemic assessments to provide more useful information for teaching and learning will require these systemic assessments to be more targeted, more focused, and more limited.

I continue this section of my literature review by exploring the role of formative and diagnostic assessments; by examining how Rasch measurement supports assessment practices; and by addressing particular applications of diagnostic assessment to cognitive constructs for mathematics. Whereas my concern is with diagnostic assessment, I position diagnostic assessment in the context of formative assessment, where diagnostic

assessment has the purpose of uncovering challenges to learning, and with formative assessment then using this diagnostic information as one of the inputs and practices which can improve teaching and learning.

Formative Assessment and Summative Assessment

Summative assessment, which is described as "assessment *of* learning", is distinct from formative assessment, as "assessment *for* learning". Formative assessment is seen as a process and as an approach to classroom practice rather than merely as a method for testing learners (Wiliam 2011a). Whereas my concern in this study is with formative and diagnostic assessments, I first compare these to summative assessments.

Summative assessment is the form of assessment traditionally understood by the term "assessment" when this term is used alone. Summative assessments provide a measurement of a learner's proficiency suitable for grading purposes, such as in symbols (A, B, C, etc…), or levels (Level 5, Level 6, etc…), and these results are often used for promotional purposes or to provide a certificate of proficiency. Summative assessments do not have the purpose of providing useful information to support learning, since they address the entire subject curriculum, often covering work conducted over many years of study, and are not focused on the particular topics as they are being taught in the classroom from day to day. Rather, the outcomes of summative assessments will often have a major implication for the learners, either positive or negative outcomes, and are referred to as "high-stakes" assessments. Summative assessments lack depth in the individual topics, sacrificing this depth for increased breadth, and are measuring a high-level construct, such as the learners' general proficiency with the subject matter of mathematics as a whole. Whereas the purpose of the summative assessment is to obtain a broad-based statement of ability, there may be some fine-level information which can be derived from these assessments. The high-stakes summative assessments, such as the Grade 12 Senior Certificate examinations conducted in South Africa, are managed and controlled from a single external source, being the national Department of Basic Education, and are thus also systemic in nature as well as being summative. Dunne et al. (2012) propose that systemic assessments should use better-targeted instruments, which have the power to provide more formative and diagnostic information, and these recommendations can also apply to the annual summative assessments.

Throughout the past 25 years there has been an important shift from assessment *of* learning to assessment *for* learning (Gardner, 2012). The terms "formative assessment" and "summative assessment", which are in common usage today, can be traced back to 1967, a mere 45 years ago, and the term "assessment *for* learning" was first used as recently as 1997 (Leahy & Wiliam, 2012). Teacher guides often indicate that classroom practices should involve time for in-class assessment but the teachers for whom these guides are written do not have the time to carry out these formative assessments (Pellegrino, Chudowsky & Glaser, 2001). There have been calls for teachers' classroom practices to be changed to incorporate formative assessment in 300,000 UK classrooms and 2 million USA classrooms (Leahy & Wiliam, 2012). Rather than there being an overarching theory of formative assessment, there are guiding principles which have been derived from practice and which should be incorporated into such a theory (Black & Wiliam, 2009).

For my purposes, diagnostic information is that which helps to identify the cause of learner conceptual problems, and this is an integral element of any formative assessment structure, such as the conceptual model proposed by Black and Wiliam (2009, p.8), and in particular their Strategy 2 "Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding". I argue that evidence of student understanding as referred to by Black and Wiliam is required, for diagnostic purposes, to also elicit evidence of learner "misunderstandings" or misconceptions.

Pellegrino et al. (2001, p.1) state that "educational assessment seeks to determine how well students are learning and is an integral part of the quest for improved education", and they provide an analytical model for assessment which they refer to as the "assessment triangle", which consists of three elements:

- a *conceptual model* of the students, identifying the target of the assessment, being the knowledge and proficiency of the learners
- *observations* as assessment tests which yield measurements of the students, and
- *inferences* drawn from the measurements, in terms of the nature of the conceptual model of a student, which is, in effect, our inferred model of a student's hidden conceptual model

This assessment triangle requires that classroom practice must include the measurement ("observations") of learners' knowledge, it being only possible to infer this model through observing the learner and by interpreting these observations. Pellegrino et al. (2001) applied this model to both classroom and high-stakes assessments, and have noted that most current assessments of achievement are not in line with current theories of human cognition. Some of their conclusions are highly relevant to my study:

- measurement should provide evidence of multiple aspects of proficiency rather than a single factor such as ability
- better assessment practices are needed to diagnose at-risk learners
- there is enormous potential for creating valid classroom assessments, but more study is needed to determine the effectiveness of the approaches
- most work is directed towards large-scale and high-stakes assessments and not to classroom assessments

Each of these conclusions resonate with the goals of my study, in the need for assessments to provide more diagnostic information, and the need to focus on effectiveness of such assessments.

Wiliam (2011a) states that "assessment is a, perhaps the, central process in effective instruction" (p. 3) and argues that assessment information alone is insufficient, unless it is embedded into a feedback system, since "the use of assessment information to improve learning cannot be separated from the instructional system within which is it provided" (p. 4). Wiliam reports that the implementation of formative assessment practices has had a profound impact, increasing the rate of student learning of around 70%, which he estimates as being equivalent to an additional eight months per year of learning. Wiliam also provides definitions for the terms "formative assessment" and "assessment for learning" as they have been applied in prior work, with the key difference being the emphasis in formative assessment on the usage of the assessment information to improve learning. For my purposes both of these terms embody a need to extract useful information about the conceptual model of the learner and I do not find it necessary to distinguish these terms.

Wiliam (2011a) asserts that the goal of "assessment for learning" is not to merely indicate gaps, nor to indicate specific difficulties, but rather to provide evidence about

insights into learner thinking which can provide guidance to the teacher as to instructional activities which should be undertaken. An example used by Wiliam concerns the case of a learner using only the denominator when comparing fractions, which is cited as an example of formative assessment, but which is more diagnostic in nature. Thus, to dive to the levels of learner thinking which are needed to guide instruction within a formative assessment environment, the assessment tasks need to meet diagnostic requirements.

Stacey (2013) reports on the use of a computer-based assessment tool to conduct "smart" tests ("specific mathematics assessments that reveal thinking") which covers 60 specific mathematics topics for learners in Years 5-9. These smart tests are introduced into the classroom as a formative assessment task, with the goal of impacting directly on learning.

Diagnostic Assessment and Development Stages

Whereas I have cited Stacey (2013) as an example of the use of formative assessment, this work on "smart" tests is also relevant to my work when it is considered as a diagnostic tool. Whereas the formative purpose of these smart tests is to inform teaching, it achieves this by identifying specific systematic errors, which is essentially a diagnostic task. These smart tests help to position learners into development stages, which are intended as a way to report learner status to teachers in a simple manner and to inform teaching practices. The learning stages used by Stacey for the topic of "Reflections" are structured on increasing complexity in the problems on which learners achieve success, with each stage building on the proficiencies of the previous stage. For example, Stage 3 is defined as Stage 2 plus the ability to "reflect a simple shape (such as circle) in any line, including oblique lines" (Stacey, 2013, p. 18). Whereas the individual test items have a diagnostic quality and encapsulate specific misconceptions or systematic errors, the development stages of the smart tests are rather defined in terms of capabilities, or increasing levels of what the learners are expected to be able to do.

There is a natural link between the structuring of such development stages and the systematic errors, or misconceptions, which are more likely to occur at particular stages, so that evidence of particular misconceptions can help to position the learners into these stages. These development stages are similar in nature to learning trajectories, a concept introduced by Simon (1995) as a teacher's prediction about the path through which learning will proceed. Simon considered that teachers will build such models naturally as

a part of their teaching practices based on their observations of learners' difficulties. Wilson, Mojica, and Confrey (2013) explored how such learning trajectories can be used as core progressions to help teachers make sense of mathematical thinking. Their approach uses the learning trajectory of equipartitioning as a long-term development of knowledge, spread over many years of learner development, rather than as short-term and finer-grained development stages. Whereas such development stages are topic-specific, I see the possibility for a topic-agnostic set of development stages, as a generic progress of development common to all topics and I propose such a model in Chapter 3.

Whereas I have presented diagnostic assessment as being coupled with learners' development stages, this coupling is a modern discourse as exhibited in the above-cited works of Stacey (2013) and Wilson et al. (2013). The majority of the prior work in educational diagnostic assessment views diagnostic assessment as a distinct research focus area in which a teacher gathers information on learner conceptual problems, requiring a set of tools to gather this diagnostic data from the learner. Among the tools which are required to aid teachers are improved test instruments, and Bejar (1984) has argued for the development of test instruments to provide more informative and diagnostic results than are provided by more traditional testing practices, firstly by analyzing students' weaknesses and secondly by analyzing the patterns of errors. Bejar concludes that that there is little guidance on how to conduct such diagnostic tests and also highlighted the lack of instrumentation for diagnosis.

Wylie and Wiliam (2006), as well as Ciofalo and Wylie (2006), explore the possibility of teachers using diagnostic questions, one at a time, in a classroom setting as an efficient means of gathering diagnostic information. However, their method sacrifices reliability in return for a gain in efficiency, and thus their approach is unable to provide learner-specific inferences, and consequently is unable to identify the development stage of learners. This focus on efficiency, at the expense of validity and reliability, is contrasted against the detailed, 30-item diagnostic test set as used by Steinle (2004a) for detecting fine-grained evidence of learning thinking on a valid and reliable basis. This opens up a line of inquiry, as contained in my research question RQ2, on whether there exists a minimum number of items which are sufficient to make an inference, and also to what extent this minimum number of items differs between various topics, and is perhaps dependent on the specific items used. Whereas a single diagnostic item is the ideal

situation—minimizing diagnostic effort so long as validity is assured—it appears intuitive that a single item may be insufficient, given the complexity of the types of learning thinking. However, it is also an intuitive expectation that 30 test items may be too many. Thus in the absence of further evidence on how many items are needed we are merely guessing our way towards an effective and efficient diagnostic practice.

To answer this question on how many test items are needed, there is also a requirement to inquire into the nature of good diagnostic test items, on the assumption that not all test items are equally suited for diagnostic purposes. This problem has been addressed by Bart et al. (1994) in formally specifying the properties of good diagnostic questions, which are applied to problems of proportional reasoning. Bart et al. consider the properties of a perfect diagnostic test item, which they refer to as a "semi-dense item", and which they define thus: "an item is semi-dense if one can exactly interpret the errors students make when they respond to cognitive rules from the responses to the item" (p. 2). They describe a "cognitive rule" as the cognitive operations that learners will use to provide their response.

The "semi-dense" item of Bart et al. is a theoretical ideal for how to analyze the effectiveness of a test item for diagnostic purpose, exploring the structure of the item as a semantic, qualitative object, as distinct from empirical results and consequent measurement analysis arising from the use of an item among a group of learners. The analysis of Bart et al. is conducted using MCQ items which have a specific set of choices, where the items are required to meet five properties, paraphrased from Bart et al. (1994):

- *response interpretability*: every item choice can be interpreted as the application of at least one cognitive rule
- *response discrimination*: every individual choice can be interpreted by only one specific cognitive rule
- *rule discrimination I*: the item has response discrimination, and each cognitive rule which interprets a choice will interpret only one choice within the set of choices. The suffix I is used to provide for differentiation between different rule discrimination rules, the others being II and U which are not covered here. For simplicity this is referred to as simply "rule discrimination".

- *exhaustive rule set usage*: every cognitive rule can be applied to obtain at least one choice
- *semi-density*: a test item has rule discrimination and exhaustive rule set usage.

These properties are related to one another and ordered into a hierarchy in which Semi-Density has preconditions of Rule Discrimination I and Exhaustive Rule Set Usage, and both of these have the precondition of Response Discrimination, which itself has the precondition of Response Interpretability. This qualitative analysis proposed by Bart et al. (1994) is an important stepping stone for my study, since this is the only formal model I have found which identifies the properties of good diagnostic items. My approach uses some of these ideas but I rather explore a statistical approach to measure the goodness-of-fit of such diagnostic items.

Cromley and Mislevy (2004) argue that test items which are suited for diagnostic purposes, and which are consequently used to identify learner misconceptions, are different from other types of assessment items, and need to be purposely created to meet the goals of diagnostic assessment. Ciofalo and Wylie (2006) propose that when diagnostic test items are used one at a time, and thus not within a test of many items, there is little need for consistency in whether there are two, three or more choices, but they do suggest that the main concern for developers of diagnostic test items should be that the incorrect test items are linked to different types of understandings, and thus the selection of every choice will provide insight into the learner thinking.

Cognitive Diagnostic Assessment

To measure the ability of a learner, various tests are conducted and then, based on the results of these tests, a score is determined for the learner. However, this notion of "ability" is an artificial, aggregate construct, since ability is a combination of a set of schemas, in which each schema may, or may not, come into play to address specific situations. When a learner makes a mistake on a test item, this may be traceable to the lack of a suitable schema, or to a deficiency in one or more of the schemas that have been used. Thus by examining the mistakes made by the learners, in conjunction with the cognitive requirements for each of the test items, it is possible to infer which of the sub-constructs are the most likely to be deficient.

As stated earlier, the learners' schemas are hidden and unavailable for direct observation, whereas the "construct" is the theoretical component being measured. For each learner, a given construct is likely based on a combination of schemas, each of which is linked to a specific sub-construct. The structuring of the construct of "ability" into its sub-constructs is the essence of the work in Cognitive Diagnostic Assessment (CDA; Leighton & Gierl, 2007a; Leighton & Gierl, 2007b) and Cognitive Diagnostic Modelling (CDM; Rupp, Templin, & Henson, 2010).

Both CDA and CDM comprise a range of models which differ in how they infer the extent of the sub-constructs used by the learners and how they account for slips and guessing. These various models have some common attributes:

- there are a set of sub-constructs, or cognitive attributes, which are constituents of the construct being measured, and which may be combined in various ways
- every learner either has, or has not, proficiency in each of the sub-constructs, which may be considered as discrete (on/off), or as continuous (range of proficiency)
- every test item either requires or does not require proficiency in each of the sub-constructs, which is structured as a "Q matrix" (Tatsuoka & Tatsuoka, 1997) showing which constructs are required for each choice on each test item
- the test results indicate which learners selected which choices for each test item

All CDA models have at their basis a mathematical model of the probability that a person with a particular set of attributes will obtain a specific response on a given test item (DiBello, Roussos, & Stout, 2007). By dividing a construct into sub-constructs we have a window into finer-gained thinking of the learners. However, there are some key assumptions made in these CDA models, including whether the sub-constructs used by a learner can be validly determined as being either on or off, or whether these sub-constructs exist as a continuum of proficiency, or are perhaps rule-based in nature and are used under a variety of conditions and not consistently applied. Another key assumption concerns the approach of considering the sub-constructs of ability as being a target for diagnosis. Whereas deficiency in a sub-construct is one approach to undertake diagnosis, the assumption that these sub-constructs are either on or off does not readily address the nature of misconceptions, which could be called "mis-constructs" and which I consider

as different, and as distinct, from the absence of ability. The CDA approach is ideally used when the questions have a complex level of proficiencies required, and represent an approach to diagnosis that would be suited to the analysis of the results from summative assessments. When the sub-constructs are extended to include erroneous cognitive rules, this leads to a potentially very large and open-ended set of constructs, which then cannot easily be mapped onto a limited set of test items.

Early work in CDA includes the Rule-Space method of analysis, as developed by Tatsuoka (1983), which was initially applied to the problems of integer subtraction problems and associated learner misconceptions. The Rule-Space approach analyses the responses of the learners to infer which rules were likely used to achieve each response, and from this to determine the most likely learner proficiencies and the corresponding conceptual deficiencies.

An alternative, and earlier, approach to the identification of faulty components in integer subtraction problems, Brown and Burton (1977) developed a program called "BUGGY" which used the artificial intelligence (AI) technique of procedural representation to simulate student behaviours in response to problems, and then used the student responses to explain why the students were making mistakes, rather than merely identifying these errors. Brown and Burton's approach considers misconceptions as distinct elements of behaviour rather than as part of a learning trajectory. Their work examined diagnosis in terms of a structure of knowledge in which certain operations are undeveloped, and which may then be the cause of incorrect results. This has formed the basis for further work in which mathematical knowledge is seen as primarily operational in nature and largely ignores the conceptual nature of knowledge.

CDA considers the totality of the attributes, or schemas, which are required to be developed within a learner's conceptual model in order to achieve success on problems within the construct of interest. CDA uses a set of test items, linked to the set of attributes, to infer the most likely deficiencies of the learner. I position my work as distinct from CDA since rather than attempting to address the complexity of the attributes, I used a targeted set of simple tests, each of which was designed for fine-grained elicitation of specific misconceptions, and which were likely to require smaller sets of items, and which I argue are better suited for classroom application. My approach was to conduct analyses in parallel for each of the misconceptions, resulting in a pattern of responses, and which

made use of simpler test items which were designed for the specific purpose of fine-grained diagnostic measurement.

The Rasch Method

To conduct diagnostic assessment using empirical data from learners, a method is needed to analyze the data received so that deductions can be drawn, and from these deductions useful information can be derived to inform decisions and actions.

The most common approach to the scoring of test results for a learner is to simply count their correct answers and then to represent the learner's score as a percentage correct. This remains the predominant method for tests conducted in the school environment to measure learner ability and is referred to as Classical Test Theory (CTT). By providing a range of test items, from easy to difficult, learners are expected to succeed on those items which are within their ability and to fail on those items beyond their ability. Thus most learners should get the easy items correct, but only proficient learners will achieve success on the difficult items. However, CTT fails to indicate by how much learners will differ in their ability when they obtain different scores, since it is dependent on the particular mix of items chosen for a test. CTT is dependent on the mix of easy and difficult test items, and a different mix of test items will produce a different score. What is needed is a method which can provide true invariance, so that the scores are consistent when tests are applied to different groups of learners, using different sets of items. The Rasch method was developed to provide the basis for calculating invariant measures of constructs so that educational measurement practices can reach the ideal of "fundamental measurement" as has been developed in the physical sciences (Wright, 1997), as described earlier.

This notion of invariance is a fundamental requirement since, when comparing two learners with different scores, we are able to say that a learner who achieves a score of 80% is thus 10% better than another learner who scores 70%, and that this same 10% is the difference between a learner scoring 70% and another scoring 60%. This is not guaranteed with CTT and, to repeat the earlier analogy, what is required is the educational measurement analogue of the ruler to measure lengths. A 30cm ruler has marks indicated for each 1cm, and this is constructed so that each 1cm distance on the ruler is the same as each other 1cm distance on each ruler ever made, within a small provision for minor

errors in its construction. As a result, the measurement of length of an object will always be the correct and consistent, no matter the instrument used.

The Rasch method was developed by Georg Rasch in the 1950s, based upon the requirement to provide such a fundamental measure, and the historical developments which led to Rasch's discovery have been described by Wright (1997). The Rasch method provides a model which determines the probability that a person of ability B will have of succeeding on a test item of difficulty D, and is expressed as a logistic function on the difference (B-D). An outline of how the Rasch method has been used for mathematics can be found in Dunne et al. (2012), and the application of the Rasch method for measurement in the human sciences is provided by Bond and Fox (2012).

Whereas the Rasch method can be used for measuring any construct, it has been primarily applied in education for measuring learner ability, where learner responses are typically scored as correct or incorrect. This is known as the dichotomous Rasch model, and is distinct from the polytomous model in which each item may have a range of possible scores. A learner response to a test item is based on a potentially large set of schemas, and it is the presence, quality, and maturity of these schemas which collectively define the notion of learner "ability" as the construct being measured.

Whereas constructs in mathematics are measured by items for which there may be a correct or incorrect answer, thus being dichotomous, there is commonly a need to attribute a more complex score, such as an item which carries a result of between 0 and 3, depending on how much evidence the learner has shown in his/her response. For the each of the Grade 12 published examinations, there is an associated Memorandum which describes the scoring approach and how marks are to be allocated. For this a simple correct/incorrect dichotomous model is insufficient, and measurement is better served by the polytomous Rasch model. This is an extension of the basic Rasch method to account for a range of scores from a single item. There are two primary applications of this, firstly in rating scales, such as the Likert[3] scale as used in attitudinal surveys. The second is the partial credit model, as explained above, in which there are differing scores for partial success on a problem. Whereas these offer some potential for use in a diagnostic

[3] A Likert scale is a rating scale for subjective information, mostly structured into five categories such as from "completely disagree" to "completely agree".

environment, I have chosen to examine a single fine-grained construct at a time, and thus have selected the dichotomous model for my study.

Individual test items may measure more than a single construct, and this may confuse the learners and bias the scores, and there is an important assumption of unidimensionality in the construct which is required by the Rasch method. This implies that the test items are all directed to measuring the construct of interest and not others. The example I cited earlier of "What is 1+1" is not appropriate to the measurement of rational number ability since the construct is different. The Rasch model requires that there is a single construct underlying the behavior of the responses. This does not ignore the potential for multi-dimensional constructs, but these are extensions to the basic Rasch method. For usage in diagnosis, the Rasch method needs to work with test items that are targeted to the measurement of fine-grained constructs, rather than the aggregate construct of "ability". Unidimensionality is threatened with vague and poorly written test items, as well as items which test extended constructs which are not core to the test requirements (Bond & Fox, 2012).

Within any test of a complex construct, such as general ability in the rational numbers, it is possible to theorize the steps of ability which a learner will proceed through, and to provide a range of items to test different attributes of the construct. When measuring ability in arithmetic, such items may test simple addition up to complex long division, and these can be mapped onto a diagram which reflects the expected development of proficiency of this construct. The individual test items will each fit this theoretical model of construct development to some degree. Those items with a good fit are better estimators, and they match the expected path of development from novice to mastery. Those with a poor fit are not good as estimators and should not be considered for use in practical settings.

The practical application of the Rasch model proceeds by understanding the nature of the individual test items in terms of the construct, and then using the learner scores to calculate the measure of the learner. This learner measure is expressed in terms of the difficulty of items at which they have a 50-50 chance of succeeding. Items that are easier they have a greater chance of succeeding on, and those with a higher measure of difficulty will not be answered successfully.

The Rasch method thus not only helps to determine the learner measures, on a fair scale of measurement, but it also provides a range of associated statistics on the goodness-of-fit of the items to the construct being measured, as well as conventional reliability statistics. Moss (1994) has proposed that educational measurement, and in particular psychometrics, places too much emphasis on reliability, and that to obtain validity it is necessary to examine a range of other elements. As a result, there can be validity without necessarily obtaining high levels of reliability. This is a direct effect of the nature of psychological data, as compared to measures of a physical nature where reliability of measurement is highly significant for accepting results.

When using the Rasch method for diagnostic purposes, I argue that the constructs being measured are specific misconceptions, which can be measured in isolation from each other. All of the remaining elements of the Rasch method are applicable, but require a shift in the interpretation, and this approach is covered in Chapter 4.

2.7 Construct Validity

Sadler (1989) has proposed that validity should be prioritized over reliability in all diagnostic contexts in which learner improvement is expected and indicates that, whereas reliability is normally considered as a precondition for validity, with formative assessment this is reversed so that validity is seen as a sufficient, but not necessary, condition for reliability.

In all assessment situations it is essential that the test results are a valid indicator of what they are intended to measure. Messick (1989) explored different forms of validity and concludes that construct validity is an essential element for all educational assessment and that validity must be applied to the process of drawing inferences rather than to the data collection tasks. Thus data alone cannot be considered as being valid or invalid, and it is when this data is used to obtain measures that construct validity becomes of concern. The construct being measured is a theoretical aggregate of an individual's schemas as constructed by his/her personal learning experiences, and to be a true and valid measure it is essential that the assessment process accommodates this construct during the development of test items, and in the processing of the results.

Validity is of particular importance for diagnostic assessment in order to derive maximum value from the assessment, and in turn to guide remediation and further

teaching. There is a spectrum of validity, ranging from completely invalid to completely valid so that it is not only the inference to be drawn from the results of the items which should be identified as being valid, as Messick (1989) proposes, but also the nature of the test items themselves, since it is clearly not possible to draw valid inferences from invalid test items.

To ensure validity of diagnostic inference for individuals there is a requirement for not only the afore-mentioned validity in the individual test items, but there will also be a minimum number of such questions that are needed to ensure that the misconception being diagnosed is the true cause and that no other extraneous factors account for the test results obtained from the learner. My study is thus concerned with both of these influences, being the quality of individual test items as valid measures of specific misconceptions, as well as the economics of assessment in terms of how many questions or test items are required to reach valid inferences on the existence of these misconceptions on the evidence of the learner responses. The nature of the construct is thus important, since validity is expressed in terms of this construct. This is addressed for this study in the Development Stage model of learning that I introduce in Chapter 3, together with the design of the diagnostic assessments I have used as part of my data gathering, as outlined in Chapter 4.

2.8 Domains and Learning Trajectories

Mathematical knowledge is traditionally structured into topics or domains, such as the whole numbers, the rational numbers, geometry, and algebra. The concept of a mathematical "domain" does not have a strict definition, and I have used this term to name each broad topic, as are found in the topic headings within curriculum statements. I introduce the term "micro-domain" to describe particular elements, problems, and proficiencies located within a domain, and which, in my opinion, are better suited for the diagnostic work. Thus, whereas a domain is a broad topic which is typically covered over 2-4 weeks of schooling, a micro-domain may be the subject of 1-2 school periods of instruction. My study was concerned with the domain of the rational numbers, and I used a number of micro-domains which provide support for the discovery of misconceptions.

One purpose of dividing up mathematical knowledge into domains is to describe the learning trajectories which a learner is expected to move through in developing

proficiency in the domain. Graf (2008, p.7) has suggested that "understanding the trajectory of development in the target domain is necessary to help students reach learning expectations". Thus, when designing test items which are suited for diagnostic purposes, these items should be considered within the context of a learning trajectory and not as isolated or independent from a context.

2.9 Web-Based Assessments

It became evident at the start of my inquiry, and has become increasingly pronounced during the intervening period, that the World-Wide Web (the "Web") is the preferred medium for new computer applications in all fields, which are now being delivered less over traditional computers and more over laptops, mobile phones and tablets. Another modern trend, which I have not explored in this study, is the potential for mobile "Apps" which are hosted on the various mobile devices. These Apps mostly utilize the Web for their data access and to support interaction between the users and shared data and processes, and for my purpose do not offer anything radically different from the use of the Web alone. I position mobile Apps as an alternative medium for the delivery of content and data processing. It is clear that any solution which is designed to support improved information about learners is required to use digital data stores, as well as a means of processing and communicating this data, and thus computer-based and Web-based solutions present a significant opportunity to take this forward into the classroom.

The DIAGNOSER system (DIAGNOSER, 2012) has been developed by the Hunt Lab in the Department of Psychology at the University of Washington. The purpose of DIAGNOSER is to provide on-line, Web-based formative assessments in science and mathematics. It claims to provide formative assessments of student thinking to inform instruction and has been implemented in various topics in physics, chemistry and biology, but I could find no mathematics assessments on the teacher home page. The particular questions designed to explore student thinking are called "elicitation questions" with the intention to stimulate learner thinking as a prelude to further instruction and discussion. All of the questions are in MCQ format, and the various answers are linked to "facets" which indicate the particular nature of the student thinking which would have caused the selection of each of these. Essentially, the alternative answers are distractors which may be related to particular misconceptions. These facets are based on an attempt to organize

all of scientific knowledge into a single structure including identification of topics which represent the more problematic types of thinking. The DIAGNOSER system is available for teachers and their classes using a simple logon process without restrictions, and it provides a reporting environment to help inform teachers about their students' thinking. However, it does not appear to be based upon a strong theory of misconceptions in the fields in which it operates.

Recent work conducted at the University of Melbourne, by Stacey and her associates (Stacey, 2013; Price, Stacey, Steinle, Chick & Gvozdenko, 2011; Steinle & Stacey, 2012) is bridging the gap between research and practice, ensuring that the wealth of research into learner thinking can be coupled with the practices of assessment for learning, to bring diagnostic assessment into the classroom, and identifying the role of the teacher in such classrooms. Their project is called SMART ("Specific Mathematics Assessments that Reveal Thinking") and is a Web-based diagnostic assessment system to "provide teachers with a quick and easy way to conduct assessment for learning" (Price et al., 2011, p 3). This SMART approach has a number of purposes and approaches which are in line with my own work, although I am exploring only one element of classroom-based assessment practice, being the understanding of learner thinking. My work concerns how such classroom-based diagnostic assessments can be improved with more reliable and valid evidence and with specific attention to the nature of good diagnostic test items, using a universal model of development stages.

Stacey describes the current status of the work as "...experimental and incomplete, yet demonstrates what is possible" (Stacey, 2013, p.14). By combining academic research with government educational requirements they have created an open resource for diagnostic assessment on 60 topics of interest in the middle school. Among these topics are the rational numbers, and among the tests are the pair-wise comparison of the magnitude of decimal numbers, which is used to expose a range of misconceptions, and which is based upon extensive prior research (Steinle, 2004a; Stacey, 2005; Stacey, Price, & Steinle, 2012; Stacey & Steinle, 2006).

At the University of North Carolina, Confrey and her associates (Confrey, Maloney, Nguyen & Corley, 2012; Confrey & Maloney, 2012) have developed a pilot diagnostic assessment system, called LPPSync, with a focus on the learning trajectory of equipartitioning. This system uses wireless mobile devices, and this work is a part of the

GISMO programme (Generating Increased Science and Math Opportunities). This is not a Web-based application, but it shares information over wireless devices.

It is evident from these recent cases that Web-based diagnostic assessment is emerging as a discipline to take misconceptions theory into practice.

2.10 Learner Self-Knowledge

I noted earlier that empirical evidence arising from educational measurements can be complemented with evidence provided directly by the learner, in terms of their self-knowledge of what they know and what they don't know.

This self-knowledge has been referred to as a "Confidence Index" in prior work (Webb, Stock, & McCarthy, 1994; Huntley, 2008) which focused on the extent to which the individual being assessed is guessing or whether their responses are drawn from their prior knowledge and expertise. Webb et al. (1994) used a five-point Likert-type scale (1-5) to collect information on the testee's confidence in their own knowledge with points 1, 3 and 5 labeled respectively as "NOT CERTAIN', "SOMEWHAT CERTAIN", and "ABSOLUTELY CERTAIN", and with points 2 and 4 unlabeled. Their study was conducted with undergraduate students enrolled in tertiary studies. Huntley (2008) has included a confidence index as one component of a model she used to identify and measure "good" mathematics questions. Huntley's study does not address diagnostic assessment but rather deals with general mathematics questions and her model explores the role of feedback within mathematics assessment practice.

In my study I worked with younger learners, and my approach was to ask how difficult the learner found the individual test items rather than to ask them how confident they were in their answer, which I expected may have been misunderstood by younger learners. I refer to my measure as a Difficulty Index, to distinguish it from the Confidence Index of the cited former studies. However, this is not merely a change of name of the index, since I have shifted the focus away from a subjective assessment by the learner of their own self-knowledge and towards the objective statement on the learner's view of the question in terms of its difficulty.

2.11 Conclusions from the Literature Survey

Diagnostic assessment is not a singular discipline which can be disconnected from the world and studied alone. It is strongly related to the constructs which it is required to diagnose, and I have explored prior work in the domain of the rational numbers.

I have examined how diagnostic assessment is positioned within formative assessment practices, and how these practices can aid learning. The formative practices which can benefit from diagnostic assessment include the need to know a learner's state of learning in particular topics of study. I have also reviewed educational measurement and the Rasch method, which I expand on in Chapter 4 as part of my methodology.

Finally, the application of diagnostic assessments using computers and web-based assessment environments is examined for its potential in future classroom-based assessment practices.

In this review of prior work, I have positioned my inquiry and my research questions to see how my research problem and questions fit into the evolving discourse and knowledge base and with the emerging practices of web-based diagnostics.

CHAPTER 3 : THEORETICAL MODEL

3.1 Introduction

My inquiry has a model-building focus and I use the term "model", as distinct from "theory" to accentuate my purpose to describe the phenomena I am observing rather than to make explanatory or causal claims (Mouton, 2001, p. 177). I do refer to my "theoretical model" in the course of this chapter to indicate that whereas this is a model, it is also grounded in theory.

I use the term "assessment" to include situations in which a learner is presented with mathematical problems to be solved and for which the results are used to infer something about the learner, for a pre-defined purpose. As discussed earlier, there are many forms of assessment, each having a specific purpose, and my concern is with assessments which provide information to support learning, which encompasses both formative and diagnostic assessments. I consider assessment to be formative when the information it produces concerns the state of learning, and that assessment is diagnostic when the results are an indication of the nature of content-related challenges to learning. For both formative and diagnostic assessments, the information produced can be used by teachers to guide their teaching activities, and can help learners to guide their learning. All diagnostic assessments may have a formative value, but diagnostic assessments have a specific purpose and thus not all formative assessments will provide diagnostic information.

To explain the role of diagnostic assessments further, I draw an analogy to diagnostic practices in other disciplines, such as medicine and vehicle maintenance.

We visit a doctor when we suspect that something is wrong with our health. The doctor performs some initial tests, such as looking at our tongue, checking our pulse, listening to our heartbeat, and making other direct observations. This may be followed by laboratory blood tests which produce quantitative results. The doctor will not pronounce us healthy on the basis of a diagnostic test, but will be able to indicate probable causes of our symptoms by reference to some normal range of values and may be able to remove certain causes from consideration.

A similar analogy is drawn from problems with our vehicles. We firstly observe a symptom: we hear a funny sound when the car turns a corner; there is a sudden drop in performance; black smoke is coming out of the exhaust; or some substance, perhaps oil or water, is dripping under the car. A single symptom alone gives little indication of which of hundreds of possible causes is the true cause of the symptom. With many modern vehicles, a mechanic will simply connect a diagnostic computer to the vehicle that will examine a large number of vehicle parameters and from this will identify likely problem causes. The purpose of these diagnostics is not to declare our vehicle roadworthy, but rather to help to isolate the problem as quickly as possible so that it can be fixed.

In both of these analogical situations—for personal health and vehicle operational health—it is not the individual measurements which provide the diagnostic value, but the inferences drawn from these measurements which help to identify possible causes of the problems.

Similarly, for educational diagnostic assessment, we have some basic symptoms, such as a learner who is not succeeding on a particular class of mathematical problems, and we want to isolate the causes, which will include the cognitive obstacles that are preventing the learner from achieving success. To carry out a diagnosis we need access to targeted tests that are suitable to isolate specific causes, which for educational requirements will include schemas which are incomplete or used incorrectly and which are thus treated as misconceptions. Once we know the cause of the learner errors then, analogous to the diagnosis of health or vehicle issues, we can plan to address these problems. Educational diagnostic assessment is thus the process of identifying the causes for observed problems so that these can be used to inform instruction.

3.2 Modeling Diagnostic Assessment

My theoretical model, as introduced later in this chapter, is used to support diagnostic assessment practices in the context of a fine-grained progression of learning within micro-domains. The development of this model is explained in the context of my research questions, and this model has been designed to account for the following:

- The constructivist theory of learning, applied at a fine level in terms of conceptual development in micro-domains of the rational numbers.

- The progression of a learner through development stages, as learners develop and refine their schemas, to increase their success in solving mathematical problems in a micro-domain.
- The role that self-knowledge plays in the identification of misconceptions.

This model then forms the basis the basis for my research approach as discussed in Chapter 4.

Prior to a detailed explanation of the model, I reflect that a key element of my research problem concerns the nature of misconceptions that arise in the learning of the rational numbers and how the detection and identification of these misconceptions can be improved to support teachers engaged in an assessment *for* learning context.

My model identifies five stages that occur in the conceptual development of learners as they gain increased levels of proficiency within a micro-domain. I name these stages ABSENT, EMERGENT, ACTIVE, IMMINENT, and STABLE, and I use these stage names in capitals throughout this thesis. These stages are progressive and they represent an abstract learning trajectory which can be used to model conceptual development in a variety of micro-domains. In effect, these stages identify points of how learners progress in their conceptual development from being a novice to a master, using a progression which is independent of the particulars of specific micro-domains. These stages are distinguished from one another by the extent to which the individual is able to address problems in a limited micro-domain context and how the individuals employ their intermediate conceptions, or misconceptions, in their attack on these problems. One of the significant features of this model is how it addresses low-performing learners, for whom traditional measurement of ability does not yield adequate measures to support remedial interventions, as discussed earlier in terms of the interpretations of the scores from the TIMSS surveys.

I position proficient learners, who are in the IMMINENT and STABLE stages, into a Zone of Competence. I position learners who are in the process of developing proficiency, and who are in the developing stages of EMERGENT and ACTIVE, in a Zone of Learning. I propose that these two Zones should be addressed differently to ensure that diagnostic assessment practices are effective. In essence, the manner in which proficient learners are assessed differs from how non-proficient learners are assessed. One critical feature of this approach is that proficient learners are separated out before the non-

proficient learners are assessed, and thus the learner population is divided into two groups as part of the detailed diagnostic analysis, with proficient learners, who require no diagnostic assessment, removed prior to the diagnostic assessments.

I also include an ABSENT stage for learners who show no evidence of any conceptual development and who have not yet reached the novice stage, and for whom diagnostic assessments are not possible since they have too little knowledge to benefit from any form of direct remediation. I place these learners into a Zone of Incompetence. These learners will require development of their prior knowledge to move up to the level at which new learning is possible. These may be learners who are more than one year behind other learners.

I define these development stages at a high-level, and I expand on these later in this chapter:

- STABLE = the learner knows the micro-domain
- IMMINENT = the learner almost knows the micro-domain
- ACTIVE = the learner is getting to know the micro-domain
- EMERGENT = the learner is just starting to know the micro-domain
- ABSENT = the learner does not know the micro-domain at all

I explore each of these stages in terms of the conceptual models which learners use during the development of their proficiency, such as preconceptions brought into a new micro-domain from prior learnings. Other conceptions are developed as intermediate conceptions and are constructed during learning. Thus, at every stage of development, a learner will have used and constructed a set of conceptions, and these are developed, modified, and refined until the learner reaches the STABLE stage, at which point the learner exhibits proficiency in the range of tasks deemed sufficient by the curriculum statement or by other standards of proficiency. Given the time constraints in the classroom, many learners do not reach the STABLE stage before the class moves on to other topics. Learners may thus remain at a state in which they continue to hold incomplete intermediate representations, and these will continue to persist as misconceptions as they progress into further topics and micro-domains and as this knowledge itself becomes a building block, and thus as prior knowledge, for the study of new areas of mathematics.

Every micro-domain consists of a range of problems which are commonly used by teachers and which are found within textbooks, and these problems can be placed on a measurement scale from easy to difficult. A measurement process, such as the Rasch method, can score each test item with a numeric difficulty factor based upon the probability of persons with a particular level of proficiency succeeding on that item. The higher the difficulty score for an item, the fewer learners will succeed on this item. The Rasch method places both learner proficiency and item difficulty onto the same scale of measurement, and this helps to differentiate responses that are merely slips, by an otherwise proficient learner, from responses for which there is evidence of the usage of specific misconceptions.

I use the term "conceptual model" to represent the set of schemas which are constructed and refined by the learner during the learning process. Various schemas are used by learners when solving rational number problems and they complement other resources, such as calculators or external help from a peer, a teacher, or a textbook. However, these schemas are the only conceptual tools available to a learner during problem solving, even with access to external resources.

I use the term "test item", or "problem" where appropriate, for the questions and problems used in my research, which are designed specifically to elicit learners' conceptual understandings and misunderstandings in the context of my research questions.

My unit of analysis consists of the individual learners who have taken these tests, coupled with diagnostic test items which are specific to the learning of the rational numbers, and linked with the usage of a Web-based diagnostic testing environment. I investigate the cognitive obstacles which learners face in developing proficiency and also how diagnostic assessment practices provide evidence of particular misconceptions used by learners in responding to the test items.

My study is played out at two levels. Firstly, at the level of the learners who are trying to make sense of the rational numbers through their answering of my test items, and secondly at the level of me, as the researcher, who gathers data to validate the elements of my model. Thus I make sense of the data I obtain from learners making sense of the problems they are presented with. There are thus two minds at work here, being firstly my own as researcher, and secondly that of the learner. I cannot merely be a passive

observer to this process, since the choices I make in what data I gather, how I collect this data, and how I interpret and present the results, reflect my own world view of what this data is, what it means to the learner and to the research, and why this should mean something to you, the reader. These levels have been addressed by Steffe (2000) in terms of his definition of first-order models, being the conceptual model of the learners we are studying, and a second-order model, which is our own model of the learners' model.

3.3 A Constructivist Theory of Teaching and Learning

I explore the development of my five-stage model through its constructivist roots, and then proceed to examine the model in further detail. These roots can be traced to the seminal work of Vygotsky (1978) who provides two important clues as to the nature of effective teaching and learning:

> … learning which is oriented toward development levels that have already been reached is ineffective from the viewpoint of a child's overall development
> (p. 89)

and

> … if a child's mental functions (intellectual development) have not matured to the extent that he is capable of learning a particular subject, then no instruction will prove useful
> (p. 80).

These statements identify an upper and lower limit for effective learning, and it is only between these limits where learning can occur. Vygotsky's (1978) model of learning is conceptualized in the Zone of Proximal Development (ZPD) as

> … the distance between the actual development level as determined by independent problem solving and the level of

> potential development as determined through problem solving under adult guidance or in collaboration with more capable peers. (p.86)

My model builds onto the ZPD by conceptualizing a "Zone of Learning", as described earlier, which I apply specifically to micro-domains, and also specifically for diagnostic purposes. The difference between my Zone of Learning and Vygotsky's ZPD is my formalized approach embodied in a fine-grained theory and model of the development stages of learning within a micro-domain, which I argue is required for diagnostic modeling purposes. I position the Zone of Learning as a model of learning which can help to identify the role that specific misconceptions, as intermediate representations, and diagnostic assessment, play in learning. The Zone of Learning enables learners to be positioned in terms of their conceptual proficiencies and misconceptions on the subject matter of a given micro-domain. My model can thus be seen as a type of learning trajectory, as described by Empson (2011), which is applied to micro-domains in the rational numbers, and which may also be applicable to other micro-domains beyond the scope of my study.

Whereas my Zone of Learning provides a model for the positioning of learners in terms of their stage of cognitive development within a micro-domain, it does not explain the progress, or trajectory, through these stages, and why learners who are not in the STABLE stage have developed and have used a range of partial and incomplete conceptions in the various stages in their learning process. My concern here is not with models of expertise such as described by J.P. Smith (1995) in which STABLE learners will continue to refine their schemas to be more efficient, effective and generalized, but rather I am concerned with the progress only to the point where a learner has developed sufficient conceptual maturity to consistently demonstrate proficiency in a micro-domain. For the majority of classroom mathematics this maturity is achieved when a learner meets the curriculum criteria as indicated by consistent success on representative problems.

In the Zone of Learning, learners actively build their own knowledge in response to external experiences. This active response to problems both confirms and challenges the learners' existing schemas and consequently requires that the schemas change to accommodate these new situations. The constructivist model accounts for why different

learners are in different stages of knowledge development, given that each will have a different history of external experiences. There is no direct way to determine how schema construction takes place in the mind of an individual learner and our only glimpse into a learner's conceptual model is to observe how he or she responds when presented with specific and targeted mathematical problems.

Piaget identifies the processes of organization and adaptation as being innate functions of the human being, and which manifest at a certain stage of physiological development (Simatwa, 2010; Piaget, 1985). The organizational element of human knowledge and mental capability is conceptualized as the collection of schemas which comprise their conceptual model. The adaptation element is enabled by three internal processes of assimilation, accommodation, and equilibration. Assimilation occurs when the existing schemas are able to process new inputs, and accommodation occurs when new inputs cannot be addressed, requiring changes in the schemas. Equilibration is the process which takes the inputs and attempts to equalize the internal schemas with the external experiences, using either assimilation or accommodation. Since adaptation is innate in terms of Piaget's model, this also implies the innateness of equilibration, and consequently equilibration is applied continuously as we adapt to each new external observation and attempt to make sense of these by modifying our schemas. This modification is the process of learning, no more and no less, and thus all of our life experience can be considered a continuous process of learning through adaptation.

In addition to adaptation driven by external experience, we also organize and re-organize our internal schemas, and this can take place through self-reflection in the absence of external observation. This occurs, for example, when we reflect on errors we have previously encountered in attempting to solve a mathematical problem, when we review our current approach and discover a new approach which works better—and all of this takes place in our mind without external observation, by a process of inner reflection.

3.4 An Approach to Assessment

As outlined above, the ZPD of Vygotsky (1978) models the learning process as it occurs from a social perspective when a learner is in contact with an adult or peer. Only interactions within the ZPD are effective for learning and it is thus necessary for teachers

to know the cognitive level of the learner so as to ensure that the interactions with the learner are within their ZPD. Assessments can be used to determine this cognitive level, and consequently to inform instruction. Teachers will engage in such practices as a natural part of their teaching practices and will conduct assessments to help drive the learners towards the curriculum goals.

The curriculum is a tangible social construction which defines what learners need to know to be considered as proficient in mathematics, and curricula are presented as structured, sequenced, performance-based subsets of the intangible social construction of the totality of mathematics. The curriculum is introduced piece-by-piece during the year and from one year to the next, and this is the moving target to which equilibration is addressed in education, and learners are expected to adapt by continually modifying their base of schemas—their conceptual models. Assessment is ideally conducted as each new topic of mathematics is introduced in terms of both the curriculum goals and the learners' conceptual models which they bring to bear on each topic. With regular assessment, and with suitable feedback, as promised in assessment for learning practices (Wiliam, 2011b), learning has the potential to be accelerated.

Piaget (1985) submits that a mental conflict occurs when a schema is not fully developed or when an inappropriate schema is used for a given situation. This situation may arise in learners through misreading a problem or failing to select the most appropriate schemas. It may also arise through learners not being exposed to a sufficient variety of mathematical situations and contexts within each domain of learning, as cited earlier for the micro-domain of the decimal numbers (Nesher, 1987). In the domain of mathematics education, a cognitive conflict occurs when a learner provides an incorrect answer to a question and when this error is shown to the learner by the teacher or by another external agent engaged in the ZPD interaction, such as an on-line assessment system. Whereas such errors may result from cognitive or from non-cognitive obstacles, my concern is only with cognitive obstacles, and in particular those identified as misconceptions. In the mathematics classroom every problem presented to the learners may potentially help with the identification of misconceptions, but to be effective such problems must be at the right level to match the learners' knowledge and must also be suited to elicit evidence of misconceptions. However, given that different learners are at

different stages of development, and also have different conceptual models, it is not possible to target individual learner's needs while conducting class-level assessments.

In terms of the ZPD I consider the points beyond which learning is not effective, being beyond the limits of my model's Zone of Learning, as introduced earlier. These two points are a high point and a low point in terms of the problems presented to learners and the learners' capability to address these using their own conceptual models. Beyond the high point, problems are too difficult for the learner, who will not have developed schemas sufficient to address the problem and in this situation the learner can only make a random guess or may choose not to answer the problem. At the low point, the problems are too easy so that whereas such problems may reinforce the conceptions already learned, they cannot add to learning in a meaningful way by the construction and refinement of schemas.

3.5 From Constructivism to Theory

Arising from the previous discussion on the constructivist basis for my model and the assessment considerations, I now describe my model in depth. This model is a fine-grained theory of learning which refines the ZPD into development stages to reflect the high and low points identified in the previous section, as being those points between which learning is possible, and outside of which no learning takes place.

My model treats misconceptions as first-class elements of a learner's conceptual model and as an integral part of the entire conceptual model of the learner. The conceptual model consists of schemas which are conceptions at various levels of maturity, including preconceptions arising from prior learning, intermediate conceptions which are actively being developed during learning, and stable conceptions, which are suited for usage on the mathematical problems in the micro-domain of interest. In this context, a misconception is thus any conception which is not stable, and which is either incorrect, inapplicable, or incomplete. I continue my argument that all conceptions can be considered as misconceptions in some context, since all conceptions may be subject to change and improvement, although I do account for stable schemas which reach a level of maturity at which they appear to be universally applicable to the micro-domain of interest. I use the term "first-class" to express my view that misconceptions should be treated in the same way and with the same level of importance as stable conceptions, since

both stable conceptions and misconceptions will contribute to learner responses. From a measurement perspective, I give misconceptions the same level of recognition as "ability" as an object of study, and treat misconceptions as distinct objects of study for my analysis.

My model distinguishes between the perceived "negative" attribute of misconceptions and the "positive" attribute of ability. A schema or conception is either negative or positive in the sense of measured proficiency and the role that the schemas plays in learner success. I focus on the negative because we can determine with confidence when a learner's mistakes are attributable to a specific misconception, but on the other hand, we may not be able to determine with a similar level of confidence the nature of the schemas a learner is using when he/she is measured as proficient by his/her successful responses. There exist a myriad of different schemas which may all lead to measured success in a micro-domain of interest, whereas systematic patterns of errors are more likely to reveal a single way of thinking, and this type of thinking may be highly refined for a particular case and may be discoverable through a suitable set of appropriate test items. My argument here is analogous to Popper's argument that we can prove falsity with certainty but may never be able to prove truth, as "no matter how many instances of white swans we may have observed, this does not justify the conclusion that *all* swans are white." [italic in original] (Popper, 2002). That is, a single black swan will disprove the conjecture that all swans are white. When I gather data from learners who provide responses to the test items that I present, I can dive into a range of possible reasons for the responses on the basis of the misconceptions which I conjecture that they hold. If there is sufficient evidence of responses that point to a particular misconception, then I have learned something about the conceptual model of the learner. If all learners achieve success on all of the items in a test, this may tell me little about the learner's true ability or about their conceptual model.

This approach of proof by falsification can also be found in the "null hypothesis" in inferential statistics, in which it may not be possible to prove a statement to be true, while it remains possible to prove a null hypothesis that states there is no impact or effect (Howell, 1999). I consider that the ability of learners to successfully *not* fall into the traps which are included as distractors in good diagnostic test items is an indication of the advancement of learning. This is thus an application of negative inference, or falsification, in the teaching of mathematics by the identification of these negative traits

within each learner—where negative is regarded as being inappropriate to the problem at hand. In other words, we can learn much from the mistakes made by learners, but we may not learn much from their successes.

3.6 A Fine-Grained Theory of Learning

My argument above has led me from a general constructivist theory of the ZPD to a finer-grained model which incorporates the role of misconceptions and which formalizes the stages of learning and development. I visualize this fine-grained model of learning as in Figure 1, which presents the previously introduced three zones and five stages that represent a learner's maturity within a micro-domain.

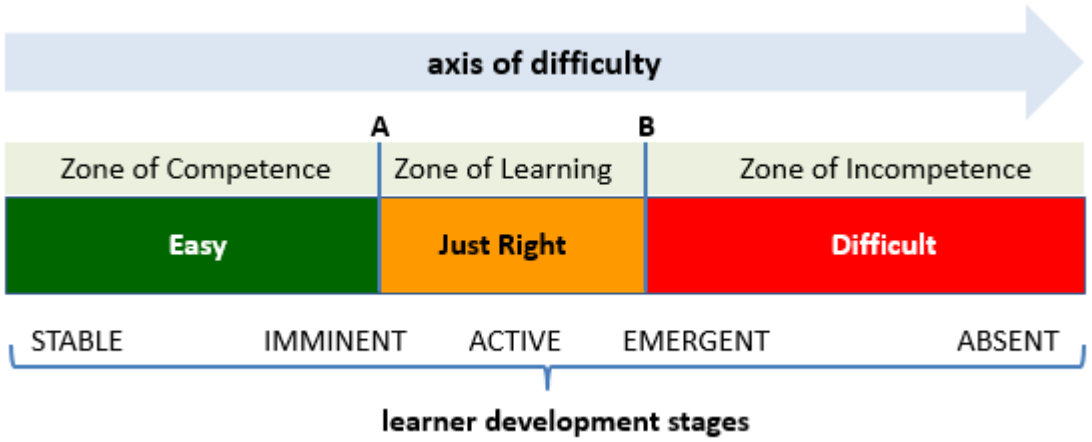


Figure 1. Fine-Grained Model of Proficiency and Learning

This model represents a single learner in his/her progress towards proficiency in a micro-domain of mathematics. The axis of difficulty represents the measured difficulty of problems within the micro-domain, and this increases from left to right. As a learner progresses, then the problems which are found to be Easy, Just Right, and Difficult will change. An expert learner will likely find all of the problems Easy, but for a novice all problems may be Difficult. The range of problems which are within the Zone of Learning of the learner are indicated by the end points A and B, representing the low point and the high point as described earlier. In this context, the terms Easy, Just Right, and Difficult, are features of the learners, rather than the problems, and are expressed in terms of how a learner will classify a problem based upon self-knowledge of his/her own proficiency. This same classification can also be obtained through empirical evidence of learners'

attempts to solve problems. Using Rasch Measurement Theory, the items can be measured on a scale of difficulty which is invariant to changes in the set of problems and learners. All such measurement is in terms of a particular construct being measured which for most assessment purposes is "ability"—which translates to item "difficulty". Items which fit the Rasch model do so because all learners tend to get the easy ones correct, and high-performing learners will get both easy and difficult items correct. Items which fit this Rasch model are thus good indicators and predictors of a person's ability since they can be used to infer more accurate ability measures. Thus we can determine, from the tests, whether a learner finds the items easy or difficult.

The conceptions used by the learners will adapt through learning, and it is possible to position all learners onto a scale of "ability" based upon their success rates when attempting a range of test items of known difficulty. However, this does not point directly to the specific schemas, or intermediate conceptions, which cause the learners to make mistakes, and thus more can be gained from knowing not just that a learner is positioned at a particular point on the scale, but to explain why he/she is at that point in terms of the schemas which are held and used, —with the schemas of interest to my study being the common misconceptions which occur in the various micro-domains of the rational numbers.

I now unpack this model into its constituent parts, and explain how each part contributes to answering my research questions.

3.7 Individual Learning Trajectories and the Points A and B

In Figure 1, I identify two specific points, labelled as A and B, which represent the low point and the high point at the extremities of the Zone of Learning. As outlined earlier, it is in this ZPD-based Zone of Learning that learners are constructing and refining schemas as they progress from novice to mastery in a micro-domain. Thus, this Zone moves during learning, so that at the start of learning there may be only the Zone of Incompetence, encompassing the full extent of the micro-domain, and by the end of learning there is ideally only a Zone of Competence, since all learning has been completed, and the learners are thus proficient in all types of problems, so that they find all such problems easy and within their proficiency. Thus for a given micro-domain the model represents

not only a snapshot of a learner's proficiency in the micro-domain, but also represents how this proficiency changes over time.

As discussed earlier, we cannot directly observe the process of schema construction and reconstruction and thus evidence of this process is obtained though indirect observation of the responses of learners to test items—and this is accentuated with items which are diagnostic in nature, which are designed to identify specific conceptions. It is only such diagnostic items which can help to elicit evidence of specific misconceptions, rather than more general items which provide evidence of learner ability. Various schemas are constructed by a learner on their personal development path, and there is no reason to expect that every learner will follow the same path or even a similar path to others. However, many studies of systematic errors in the rational numbers, conducted over the past 30-40 years, have shown a considerable consistency in the ways of thinking that lead to errors. These ways of thinking are intermediate conceptions which are progressively developed over time. As a result, these ways of thinking can help to represent the specific stages of development for each micro-domain. My model is thus essentially a generic trajectory of learning through which learners move in their passage from novice to mastery within each micro-domain. This generic trajectory is different from the learning trajectories identified by Empson (2011) who has defined a learning trajectory as a standard progression of conceptual development in a particular topic, as exemplified by the learning trajectory identified by Confrey at al. (n.d.) for equipartitioning.

Individual test items can be positioned onto this axis of difficulty from novice to mastery, in terms of the suitability of these items for learners who are at particular stages. These test items not only help to determine ability, but also, when used for diagnostic assessment, help to identify the particular misconceptions that are being used by learners who are at a particular point on this spectrum of capability. For example, some test items may elicit late-stage misconceptions, which are used by otherwise proficient learners who are in the IMMINENT stage in their final thrust before reaching mastery in the STABLE stage.

3.8 Micro-Domains

I introduced the notion of "micro-domains" in section 2.8 "Domains and Learning Trajectories" on page 69, and I now define this notion in further detail, considering a micro-domain as a small, bite-size unit of teaching and learning, which is mostly derived from stated curriculum competencies, and which may consist of a specific type of problem, such as decimal number ordering. Learner proficiencies at the level of micro-domains are combined to create higher-level proficiencies for a domain as a whole, as part of instruction in mathematics.

My model of learning, using micro-domains as the key unit of assessment for diagnostic analysis, is a bottom-up approach in which knowledge is gained in fine-grained areas and then combined to create more generic knowledge. This is in contrast to the model in which learners will first learn generics, and will then apply these generic structures to finer levels of problem situation and micro-domains.

The rational numbers are an important domain of middle-school mathematics, and this domain can be divided into micro-domains such as "decimal place-value", "rounding of decimal numbers", "decimal number ordering", and "adding common fractions". Each of these micro-domains is sufficiently small to be packaged as a self-contained unit of knowledge suited for classroom teaching and also for fine-grained diagnostic measurement. However, each micro-domain has pre-requisite proficiencies, and the lack of success in a specific micro-domain may be traced to limitations in this prior knowledge.

I consider micro-domains as atomic learning units, in the sense that they cannot easily be divided any further for teaching or assessment purposes without losing their holistic identity and structure. I position diagnostic assessment, in support of learning, as being best addressed at the level of micro-domains rather than being applied to an entire domain such as the rational numbers. Every response from a learner to a test item will cause the learner to either use their schemas or to guess if they lack appropriate schemas to address the problem. These schemas are continually evolving as learning develops and thus for my purposes it is less important to know whether a learner is capable or not in a specific micro-domain, but more important to understand the set of schemas which they bring to each problem. These schemas will include both those which are stable and effective for the micro-domain as well as those which are under development.

For my study, these micro-domains are not necessarily aligned exactly with the curriculum, since such curricula change over time and are often country-dependent. I have rather aligned these micro-domains to misconceptions' research.

3.9 Item Difficulty Reconceptualized

Of all of the elements of this study, it is my diagnostic-based reconceptualization of the term "item difficulty" which may cause the most confusion and I now unpack this modified definition.

One of my concerns is with the nature of test items that have good "diagnostic value", such as is contemplated in a qualitative sense from the semi-dense items of Bart et al. (1994) as I have referenced in Chapter 2. To address the notion of diagnostic value, I am asking a research question concerning why some test items appear to have better diagnostic value than others. This is an important question because if we know the diagnostic value for individual items within a given set, then we can rank these items in terms of their relative suitability and effectiveness for diagnostic purposes. Whereas Bart et al. (1994) propose a qualitative approach, I am rather exploring a quantitative approach to the discovery of diagnostic value and suitability.

This notion of diagnostic value demands a specific definition, since it is integral to my study and also to my approach to the evaluation of specific items. I consider that one test item A has a greater diagnostic value than another item B, for the purpose of diagnosing a particular misconception, if item A is more likely to elicit evidence of the misconception—when applied to learners who actually are using this misconception—than item B.

This notion of diagnostic value gives rise to the two-dimensional model in Table 1, which is represented by the vertical axis (item potential to elicit evidence), and the horizontal axis (learner possessing a specific misconception as one of their schemas).

The goal is to find the items which are positioned in the cell marked with "(X)" which are suited to detect learners who hold the misconception, and to distinguish these items from those which appear in the other three quadrants. In addition, the goal is to determine the best fitting items to meet this requirement, as measures of this construct, being the misconception, and which are then better for this purpose than others.

Table 1: Items with Good Diagnostic Value

ITEMS / LEARNERS	Learner does not have misconception	Learner has misconception
Item can elicit evidence of this misconception	Error – item elicited evidence incorrectly. **Poor diagnostic value.**	(X) Correct – learner misconception is detected. **Good diagnostic value.**
Item cannot elicit evidence of this misconception	**No diagnostic value.**	Error – item is not suited to this purpose. **Poor diagnostic value.**

This notion of diagnostic value can be considered in terms of the traditional notion of item difficulty, when the construct is a misconception which is being measured.

The determination of item difficulty is a standard process of both Rasch analysis (Bond & Fox, 2012; Linacre, 2013) and IRT (Hambleton & Swaminathan, 1985), and both Rasch and IRT determine the probability that a learner with a specific ability will be successful on an item which is at a particular level of difficulty. In both Rasch analysis and in IRT the items and the learners are positioned on a single scale of measurement. For my purposes here, Rasch analysis is preferred over IRT due to its suitability for small samples, whereas IRT is recommended for a minimum sample size of 200 (Wright, [2005]).

This measurement scale, and consequently the notion of "item difficulty", is traditionally aligned with the construct of learner "ability". The probability of a learner being successful on a more difficult item will be smaller than the probability of the same learner being successful on a less difficult item. However, both Rasch and IRT can be applied to the measurement of any cognitive construct, such as personal attitudes, and these methods are not restricted to measuring "ability" as the sole trait of interest. Whereas this is the primary interest for summative, systemic, baseline, and much of formative assessments, with diagnostic assessment the construct of interest changes radically.

Ability is measured as a single construct but, as discussed earlier, learner ability is better viewed as a set of schemas, each with a specific purpose and function, which are used collectively to enable a learner to answer test items, and thus to demonstrate proficiency. When measuring a learner's ability, we are measuring whether a learner's set of schemas is suited to a test item as presented to the learner. We are consequently not measuring the existence or nature of individual schemas, but rather measuring the extent

to which the combination of the learner's schemas is sufficient for the learner to answer the item. We can measure ability without ever knowing what these schemas are in detail, since the schemas themselves are hidden inside the 'black-box' of the mind of the learner.

The same argument does not hold as I shift my focus to misconceptions as the construct of interest, in which each misconception is a single construct for which I seek evidence, rather than the composite trait of "ability". I thus explore the extent to which a specific misconception can account for errors in the learner responses to one or more items. My construct of interest can be defined as "learner propensity to use a specific misconception in a particular situation" and the items are also measured as the extent to which they are suited to elicit this misconception, as per the matrix presented in Table 1 on page 90. For these special-purpose diagnostic items, the notion of "item difficulty" is better relabeled as "item as elicitor of misconception". The terms "easy" and "difficult", as are used in traditional item measurement, also require redefinition since they now measure the extent to which a learner who has a particular misconception will use this misconception to select the multi-choice response which is based on this misconception rather than selecting other choices. In this context, the term "easy" applies to items in which most learners who use this misconception will select the response which is a "rich distractor" for this misconception; whereas "difficult" means that only a few learners will select this rich distractor. I define a "rich distractor" as a choice in a multiple-choice question which is designed to elicit evidence of one or more misconceptions. Thus, as long as the test items fit the construct being measured, which in this case is a specific misconception, then "easy" items are likely to catch more learners in a diagnostic test situation. Expressed in probabilistic terms, I determine the probability that a learner who is using a misconception will then "succeed" in selecting the choice which indicates this misconception. Once I have established the validity of a set of test items which provide evidence of a misconception, then I argue that the best items to use for diagnostic purposes are those with the highest probability of selection of the rich distractor for the misconception. This is how the diagnostic value of test items are calculated, and this provides the basis for determining which ones are more suited than others for this specific purpose.

I thus conclude that there is a significant difference in how learners should be measured between the proficient learners, who are in the STABLE and IMMINENT

stages, and the active learners, who are in the ACTIVE and EMERGENT stages. Measuring learner ability using standard Rasch scores is effective to achieve the goal of ranking high-performing learners. However, low-performing learners may obtain their low scores for a number of reasons, and diagnostic assessment has the purpose of identifying which of many possible cognitive causes is the most likely explanation for each learner. Exploring the patterns of the learner responses against known misconceptions will help to identify patterns of learner thinking to account for the observed errors. Thus, whereas there is a single "ability" score that is suited for the high-performing learners, the low-performing learners should be given specific scores based upon each individual misconception that they may have used to answer the test items. For the case of low-ability learners, the measure of item difficulty is thus less significant than a measure of the usage of a specific misconception. I see this as a reconceptualization of the traditional measurement attribute of item "difficulty".

Whereas traditional assessments on ability are conducted using the entire set of learners, I have chosen to apply a different approach to measurement, so that those with less than high proficiency are analyzed using methods which are specific to the discovery of misconceptions. As a result, the high proficiency learners are not included when conducting diagnostic assessments. This is a claim arising from this study, that lower-performing learners should not be measured in the same way as high-performing learners.

3.10 Zones and Development Stages

Whereas I have reconceptualized item difficulty from the viewpoint of low-ability learners, there remains the need to position learners in terms of their raw ability scores, as the first step in an analysis of cognitive causes.

Each individual learner can be mapped onto the range of abilities, as represented by the various test items and problems with known "difficulty" measures. I structure the spectrum of proficiency from novice to mastery into three zones which I have previously introduced as the Zone of Competence, the Zone of Learning, and the Zone of Incompetence. Coupled with these zones are the development stages of the learners. I use the Zones to reflect the general positioning of the learner into a micro-domain's spectrum of proficiency, and I use the development stages to reflect the cognitive constructs which a learner in a zone will have in terms of their ability to solve a problem.

A learner's Zone of Competence indicates the problems which they are able to solve successfully within the micro-domain. If the learner's Zone of Competence encompasses the entire micro-domain, then this learner will have developed and refined a set of schemas which address the common misconceptions within this micro-domain.

At the other end of the spectrum, the Zone of Incompetence identifies the range of problems that the learner cannot address since they have no schemas to use to understand or internally represent the problem as presented and thus can only guess at an answer. Whereas the learner may bring in prior knowledge from other domains, the problems in this Zone of Incompetence are beyond the learner's ability to even commence an attack on the problem—they cannot read the question and cannot understand what is required. Presenting such a problem to this learner is a waste of effort for both the teacher and the learner. However, detecting which problems fall within this zone for a particular learner is not trivial, and I explore this within this study. Prior to commencing a new micro-domain, many learners will have the entire micro-domain within their Zone of Incompetence.

I agree with Vygotsky (1978), as cited earlier, that no learning takes place within either the Zone of Incompetence or the Zone of Competence, given that the problems presented are either too difficult for the learner to make sense or, or they are too easy and can be answered with ease. The problems within these Zones do not challenge the learner, and thus learning only occurs within the Zone of Learning which exists between the Zones of Incompetence and Competence. However, there are no clear divisions between these zones, but rather fuzzy transitions, and it is for this purpose that I have identified the transition stages of EMERGENT – in terms of the initial commencement of learning from the Zone of Incompetence to the Zone of Learning, and also IMMINENT – from the Zone of Learning into the Zone of Competence.

My purpose in introducing development stages into the model is to identify, at a fine-level, where learning is and is not taking place for each individual learner, and to determine this by asking the right diagnostic questions to isolate the cognitive causes of mistakes, rather than basing the position of the learner on ability alone. To make this point, I repeat my argument from Chapter 1 to explain the very low success rates of South Africa on some items in the TIMSS surveys, and the inferences which can be drawn from these, with this same argument also applying at the level of standard school tests. My

argument is that the ability measure alone is insufficient to determine a learner's position within the development stages, and that different methods are needed for the evaluation of low-performing learners.

Knowing the development stage in which a learner is located helps to introduce an economic approach to learning, in which the minimum interaction with the learner achieves the maximum growth in sustainable proficiency within the shortest possible time. I argue that nothing is lost in the learning process by being more effective and more efficient, even though this is a clinical and somewhat theoretical approach to the true realities of the mathematics classroom. A critical success factor to achieve this effectiveness and efficiency is to empower teachers with a fine-grained knowledge of learner misconceptions, and it is consequently important to embed such diagnostic and formative assessment practices into the classroom as recommended by Wiliam (2011b). An extension of this critical success factor is to provide the teacher with the right tools to enable them to conduct diagnostic assessments quickly and easily. I originally introduced these development stages at the start of this chapter, and I expand these in Table 2.

My model of Development Stages is used to support my research question RQ1 (EFFECTIVENESS) where my concern is to discover which test items are more effective than others in isolating misconceptions which occur at a specific development stage. By assessing the results of learners, such good diagnostic questions allow an improved approach to position learners at a particular stage of development, being a more accurate representation of learner conceptual development. These good diagnostic questions are thus those which have a higher diagnostic value, as I have outlined earlier.

RQ2 (EFFICIENCY) follows from RQ1 by asking how much diagnostic work is needed to produce a sufficiently valid score, on the basis that the less work required to meet the goal of effectiveness, the more efficient is the assessment process.

Table 2. Learner Development Stages in micro-domains

STAGE	DESCRIPTION
ABSENT	*The learner does not know the micro-domain.* The learner has no schemas sufficient to address the stated problems. He/she cannot understand the questions posed or the choices, whether expressed in words, diagrams, or in symbols, and thus cannot recognize what has been given and what is expected to be produced. The responses to test items presented follow no systematic approach and guessing is the only alternative for the learner.
EMERGENT	*The learner is just starting to know the micro-domain.* This is the novice state, where the learner has an initial understanding sufficient to read the questions and to recognize the inputs given and the outputs expected, seeing symbols, words and other expressions with an initial recognition of them and their place. He/she has not developed schemas to solve the problems but is recognizing the nature of the problem. There are initial schemas which are being used, perhaps from pre-conceptions.
ACTIVE	*The learner is getting to know the micro-domain.* The learner has developed some schemas for how to use the symbols to understand and to solve problems and is getting some problems right and some wrong as the schemas are adapted to deal with new situations. Some solutions are based on a process of educated guessing in which the success of such guessing improves over time and with the right quality of feedback. There is dynamic and regular change in the schemas, and this stage has the fastest rate of conceptual change within the development stages.
IMMINENT	*The learner almost knows the micro-domain.* The learner is achieving a high degree of success on most problems presented and final refinements are being made to the schemas to deal with increasingly difficult problems. Some of the schemas reflect late-stage misconceptions, which are the final changes to the schemas before reaching the STABLE stage.
STABLE	*The learner knows the micro-domain.* This is the state of mastery in which the learner is able to deal with all problems as presented and to provide a successful response. This mastery evolves by the learner developing increasingly more efficient schemas to deal with the problems. It is possible that this process of mastery development will continue forever.

3.11 Self-Assessment

A final element of my model is the inclusion of the learner's own self-knowledge on the difficulty of items. This self-knowledge is an alternate source of evidence which may help to identify misconceptions based upon the prior work identified in Chapter 2, such as that of Huntley (2008), in which misconceptions are likely to exist whenever a learner identifies an item as easy and yet selects an incorrect response.

So, whereas items can be positioned onto a measurement scale of difficulty using the quantitative, objective Rasch process, all learners will also have their own individual qualitative, subjective perception of how difficult an item is, and should be able to reflect on their own abilities to state whether the item is Easy, Just Right, or Difficult. In answering this question, the learners must have sufficient understanding of their personal capabilities in terms of each of the questions presented. It was my assumption that if learners identify a question as Easy or Just Right, then this implies that they are likely to have a schema which they have identified as suited for their attack on the item. If this schema results in an incorrect answer, then one inference is that the schema is wrong and thus they may have selected the wrong schema. Another inference is that they selected a schema which is incomplete. Whatever the outcome, something can be learned by gaining access to the learner's understanding of the difficulty of the item.

3.12 Commentary on the Theoretical Framework

My model of the Development Stages of learning explicitly positions misconceptions as central to the entire learning process, specifically within the domain of the rational numbers in the mathematics curriculum. My model elevates misconceptions to first-class constructions in the learning process which are not merely there to be remediated but are treated as sufficiently important that we cannot position a learner onto a scale of development stages without understanding the learner's intermediate representations developed during the process of learning.

I have argued previously that conceptions and misconceptions are essentially similar and both are considered as schemas in the conceptual model of the learner. This includes stable conceptions, preconceptions, and intermediate conceptions, all of which are in a continuous state of adaptation and improvement. Thus there is no perfect schema,

but merely schemas that become increasingly adapted to the mathematical experiences that learners are given in their pedagogical environment. Those schemas which do survive are best adapted to the needs, and will become increasingly stable over time. However, every schema can be placed into a context in which it may fail, and thus misconceptions are not an inherent quality of a schema, and it is rather the case that a schema becomes a misconception at the time that it is applied in a context in which it does not fit. As part of the learning process, the learner may adapt their schemas to a new situation, or may create a new schema for the special situation as observed, while retaining the existing schemas for its previously proven purposes. Since we cannot actually inspect these schemas, and also cannot observe their creation and adaptation, it is necessary to theorize their nature and the process of modification.

The detection of misconceptions is not a guaranteed outcome of traditional formative testing, and I argue that specific diagnostic tests are required to elicit evidence of each misconception. Knowing which misconceptions are used by a learner will support the positioning of learners into development stages and can thus aid personalized instruction. Teachers can only apply the practices of assessment for learning in the situation in which such information on specific misconceptions is available for each of their learners.

CHAPTER 4 :
RESEARCH DESIGN AND METHODOLOGY

"Never send a human to do a machine's job" Agent Smith, The Matrix (1999)

4.1 Introduction

In this chapter I outline the research design and methodology to address my research questions, in the light of the Development Stage model for learning in micro-domains introduced in Chapter 3. I describe the sampling method, the approach to assessment for each of the individual micro-domains, and the plan for how the assessments were conducted for both pretests and online lessons. This includes the details of the data collection tools which were designed for the assessments. I explain the usage of the Rasch method for measurement of the data, and the details of how the data was analyzed.

The primary data for this study was gathered during "online lessons" which consisted of learners accessing a web-site which was specifically developed for this study. The online lessons were composed of "lesson elements" which were presented in sequence. These lesson elements included "tests", "information elements", and "results". The information elements consisted of text and graphic explanations, which were provided before and/or after the tests. The tests used for this study, also referred to as "online assessments", were collections of test items which were presented to each of the learners in a pre-defined sequence. The results were provided at the end of each test as feedback to the learner and showed the results of the learner's performance on the test. These results were provided to assist the learners to identify which questions they answered correctly and which were incorrect, but these results did not identify or explain the possible causes of incorrect responses.

In order to build up the online test items a pretest was conducted, for which the results are presented in Chapter 5.

4.2 Population

The population for this study is, at the extreme, all learners who are in the Senior Phase (Grades 7-9) of South African schools, and who are required to take Mathematics as a compulsory subject. The rational numbers are introduced in the Senior Phase as a distinct mathematical topic and the curriculum includes varieties and combinations of representations and notations. The knowledge of the rational numbers is an important foundation for the work covered in Grades 10-12 Mathematics and Mathematical Literacy. By the end of the Senior Phase, learners are expected to be proficient in each of the types of rational number, including decimal numbers, common fractions and percentages, and to know how to convert between these types.

In 2010 there were 2,991,254 learners in the Senior Phase (Grades 7-9) covering all of the schools in South Africa (DBE, 2012). This is a particularly large population for a study but is informed by the results from large-scale studies conducted throughout the world, which provide evidence that many of the rational number misconceptions pervade time and place. This is reflected in the international TIMSS 2003 study (Mullis, Martin, Gonzales & Chrostowski, 2004), as well as in multi-national studies such as Resnick et al. (1989), and in longitudinal work conducted in Australia by Steinle (2004a, 2004b). In particular, the TIMSS 2003 study shows a very low success percentage of South African learners on rational number items, and this justifies my selection of the entire country as the population for this study. In essence, the problems I am exploring are likely to pervade the entire country.

However, given that my approach was to use web-based assessments I needed schools with working computer laboratories, where there were both teachers and learners who were fully computer-literate, and thus my study population was limited to schools with the computer facilities and capacity to enable a study such as this. Data on how many of the national schools were sufficiently resourced was not available but is expected to be relatively small, perhaps no more than 5% of the schools in the country at the time of this study.

4.3 Sample

My investigation was conducted at two schools in the Northern Suburbs of Johannesburg[4], which I refer to as School A and School B throughout this thesis. These two schools were selected on the basis of meeting a range of criteria concerning sufficient computer facilities, and sufficient computer and web proficiency for both the learners and the teachers. The schools were chosen on the basis of convenience, given that they were both available as study sites, and also were conveniently located in the geographical region of Johannesburg.

The specific criteria I established for participating schools were:

- a functioning computer room, with one computer per learner and with every computer having access to the Internet
- suitable furniture to allow the learners to use the computers without being too close to one another, including one chair per learner
- computer teachers who are proficient with computers, and who generally know more than their learners
- learners who are computer-literate and who have been using computers at school for several years

Whereas these requirements may create the perception that this study was exclusive and only for the wealthy schools[5], I have assumed that there will be a future widespread implementation of computing and communication facilities within most schools, in which all teachers and learners are proficient with computers.

I conducted the study for School A in 2009 with two Grade 7 classes with a total of 56 learners. School A adopts a policy of allocating learners to classes randomly at the start of the school year, and this is relevant considering that at the time of this study I had initially chosen to divide up the two classes into a control group and an experimental group. Thus for School A there would be no bias to this class selection given the schools policy on learner allocation. School A is a public primary school, falling under the Gauteng Department of Education (GDE).

[4] The northern suburbs of Johannesburg have traditionally been viewed as the more affluent areas within the broader metropolitan area of the city.
[5] As at the time of this study, there was a relatively low percentage of schools in South Africa with working computer laboratories satisfying these strict requirements.

Following the initial assessment of the results from School A, changes were made to the structure of the study, and a second school (School B), a private secondary school (School B), was selected. For School B, no control group was used, and the online lessons were conducted with all three Grade 8 classes of size 25, 25, and 24—a total of 74 learners. This second study was conducted in 2011 with a two-year delay from School A which was partially caused by the impact of the Soccer World Cup in 2010, which disrupted normal school schedules and which consequently made it difficult to accommodate the additional disruption arising from research studies.

Information sheets and consent forms were prepared for each of the schools and provided to the learners for their own information, as well as for their parents or guardians, since the learners were too young to give their own consent to be involved in this research. The consent forms were returned to the schools to ensure approval had been given to participate.

4.4 Method

Two forms of tests were used for this study. A pretest was given to all five participating classes from the two schools. This was followed by four of the five classes being given the online tests as part of an online assessment and instructional system.

These online tests were conducted over four school periods, one lesson per week, during the Computer Studies periods so as to minimize any interference with mathematics lessons. The online tests were supplemented with introductory materials, instructional content, and feedback. The online lessons were conducted without direct involvement of the teachers, who were not pre-informed about the nature of the specific tests and who were not shown the test items until the full set of tests had been completed. My decision to not inform the teachers about the tests was to minimize any bias in how the teachers may have conducted their teaching from week to week if they had had prior knowledge of the types of questions being used within the diagnostic assessments.

The pretest was a paper-and-pencil test conducted in the classroom, and these were conducted and completed in a single school period in the week prior to the first online test. The pretest included basic instructions on the front page of the test paper, which was read aloud to the classes at the start of the pretest. If any learner could not complete the full test in the time allotted then they handed back incomplete papers, and

no extensions of time were provided. The pretest was used in this study to inform the test items to be used for the online assessments; however, the pretests were also a part of the original experimental research design.

For School A, the research methodology was originally planned as an experiment, using pretest and post-tests, with learners divided into control and experimental groups. Prior to the commencement of the research the mathematics teacher selected one class randomly as the control group and the other as the experimental group. Both classes were given the pretest during the same school period. The experimental group was then given access to the online lessons, and one school period per week was allocated for these assessment sessions, which was planned for four consecutive weeks. The first three of these assessment periods were conducted over three consecutive weeks, and the last of the four online lessons was conducted after a short school holiday break. However, the original approach to use an experimental design was changed after the end of the online lessons at School A, with the scope of the study reduced to focus solely on the research questions which were presented in Chapter 1. The other research questions that had been originally part of this study, concerning the impact of diagnostic assessment in a web-based classroom, are left to further studies due to the requirement to limit size and scope of the project, which had become too large.

The four online lessons in School A were structured along the lines of a "design experiment" (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) in which the results of each lesson informed the content and structure of the following lesson. This methodology supports flexibility in the approach and enabled me to adapt the lessons. Following each weekly online lesson, I performed an initial evaluation of the results and from this I planned the next lesson. This design experiment approach was performed only for School A, with the lesson structures then used, being essentially the same for School B.

My original research design was to conduct this study at a single school. However, based upon my analysis of the data from School A I decided to repeat the study at School B, consisting of three Grade 8 classes. The inclusion of School B allowed me to increase the quantity of the learners involved in the study, providing a larger sample of results for analysis, and this also provided the experience of working in a different school environment at a different grade level. The data that was collected from the two schools was merged into a single database for the data analysis.

For School B I made some small changes to the structure of the assessments to accommodate the findings of the first study, using the same tests and items as I had used in School A. This facilitated the merging of the data between the two studies even though the tests were not exactly the same. This provided a larger base of data from the online lessons than I had obtained from the single class at School A.

I planned this study to run in isolation of any knowledge of the individual learner's rational number proficiencies, and I chose to have no engagement with the mathematics teachers either before or during the assessments, with the sole exception being communiques which I provided to the head of school and the head of the mathematics departments concerning the progress of the study.

MCQ assessments were used exclusively in the online assessments although both MCQ and provided-response formats were used for the pencil-and-paper pretest.

4.5 Tools for Data Collection

Prior to describing the instrumentation of the tests used in this study, I outline the data that I planned to collect in this study, specifically concerning the micro-domains and the corresponding item types, and the misconceptions that the test items may, or may not, elicit. Whereas this is more detailed than would normally be included into a chapter on methodology, this information is relevant as the bridge between the research questions and the theoretical model on the one hand, and the data analyses of the results obtained on the other. Thus in this section I explain the specific items used to gather the data from the learners.

To illustrate these micro-domains and the types of test items I have used I present examples from the online test bank, showing how these test items appeared to the learners during the online assessments. The test items used within the pretests are provided in Appendix B.

Good Diagnostic Items

Each of the test items was developed for diagnosis, and none of these test items were designed to support grading of the learners in terms of rational number proficiency. Thus each of the test items can be viewed as a "trick" question—designed to trip up the learners. Whereas the use of such trick questions may be seen as unethical, I consider that it is

exactly these kinds of questions which are ideal diagnostic instruments that maximize the opportunity to elicit incomplete or incorrect mathematical thinking.

However, the diagnostic value of the items I used was unknown in advance, although some items were drawn from prior studies. Other items were introduced without any prior knowledge of their diagnostic value. An outcome expected from these tests was to establish the suitability of these items by ranking them in terms of a measure of diagnostic value, using the definition of diagnostic value as I introduced in Chapter 3.

The MCQ test items used in the online assessments each had between 2 and 11 choices. Most of the choices were rich distractors which were included to elicit evidence of particular misconceptions. These rich distractors were distinct from choices that have no linkage to known misconceptions, which I refer to as "random distractors".

Introducing the Micro-Domains

Each of the test items used in the study was codified to indicate the micro-domain which they are part of and the specific form of test item—such as PV1, which is the code for the first type of place-value test item. If a choice in an item could have resulted from more than one misconception, or from both a misconception as well as it being the correct choice, then the choice was codified with each of these alternative conceptions. The misconceptions used were selected both from prior research as well as from my personal experience in tutoring learners in the rational numbers. The nature of the coding was such that new codification could be introduced in the future for new misconceptions that may be discovered, allowing for further processing of the data sets.

I now explain each of the micro-domains that I used for this study, showing examples of the types of test items and the corresponding misconceptions that can be elicited in each micro-domain.

Micro-Domain PV – Decimal Place-Value

Place-value knowledge is fundamental to the understanding of decimal numbers and thus also to decimal fractions. Place-value misconceptions in decimal numbers occur when learners do not know how to determine the value of a particular digit in a decimal number and the learners resort to prior knowledge, such as using whole number place-value knowledge. For example, in the decimal number 12.345 the learner may see the digit 4 as tens or tenths, instead of hundredths, which likely results from the learner ignoring the

decimal point and seeing this as the whole number 12345. Place-value misconceptions may also be intermediate conceptions which are constructed as learners develop their knowledge of place-value in decimal fractions.

As discussed earlier, throughout the pretests and the online testing I exclusively used the decimal point, rather than the decimal comma, for representing decimal numbers. This is for the reason that learners are more familiar with the decimal point in their work with computers.

I have not encountered prior studies during my literature review that were designed to specifically explore misconceptions of the place-value notational system as a separate topic of study. However, place-value knowledge is implicitly required for understanding how to compare the magnitude of decimal numbers, and these types of questions have a long history within misconceptions research, and thus the wealth of these prior studies has indirectly explored learners' understanding of place-value. My intention in introducing this place-value micro-domain was to explore the nature of the misconceptions using decimal notation alone and without placing decimal numbers into a context of operations, such as ordering decimal numbers by their magnitudes.

I introduced two types of diagnostic test items for place-value knowledge, each of which links the numeric representation of a decimal number with the names used for the place-values of the individual digits in a decimal number. For both types of test items, the decimal numbers were constructed so that the same digit was not used more than once in the decimal number, to reduce ambiguity in the learners' responses.

The test types were encoded as PV1 (find a digit at a stated place in a decimal number) and PV2 (find the place of a given digit in given number).

Item Type PV1 : Find the digit at a named place-value in a decimal number

Figure 2 was taken from the Item Bank for this study, which is presented in full in Appendix D. Each of the examples shows an Item Number, such as 10003 in Figure 2, which is used as the reference to the item in the item bank.

The selection of which digit in a decimal number corresponds to a particular position can help to expose a lack of knowledge concerning the relationship between the whole number place name and the fractional place name, such as between "thousands" and "thousandths". The response of the learner may also reflect the uncertainty of where the counting should start to determine the position of digits within a decimal number,

such as units, tenths, hundredths, and also in which direction the learner is required to count to find the different place-values—such as whether they count from the units digit as for whole numbers, or from the decimal point.

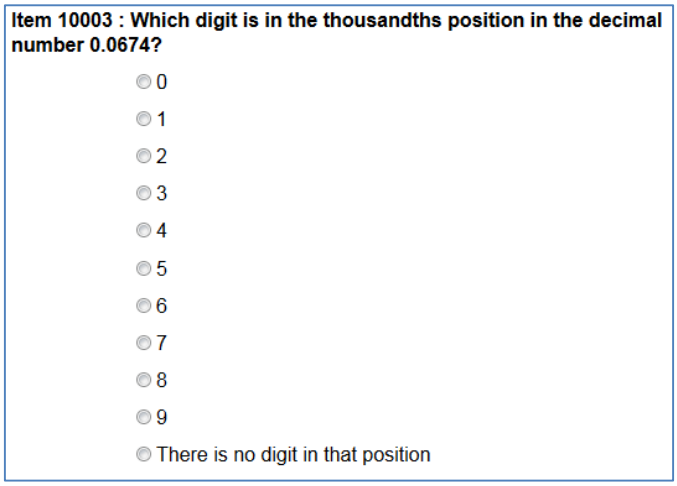


Figure 2. PV1 – Place-value sample test item - select digit at named position

To clarify the definitions which I introduced earlier, a decimal number, such as 13.45, has a whole number part (13), a decimal mark (.), and a decimal fraction part (.45). The place-value system uses the convention that the position of the digit relative to the decimal mark determines its magnitude, so that in the whole number 123, the digit 2 represents "tens", and has the value 20, and not 2. Place-value knowledge is an important element of mathematics instruction in the early grades, as learners come to understand the decimal notational system of place-value within the whole numbers. When the place-value notational system is extended to include decimal fractions, the process of learning introduces a number of misconceptions.

One source of these misconceptions concerns the naming system we use for the place-values. For the whole number part of a decimal number the place-values have the largest place-value on the left and the smallest on the right (thousands, hundreds, tens, units), increasing in magnitude from right to left. However, when examining the words used for the digits in the fractional digits of the decimal number these appear to increase from left to right—tenths, hundredths, thousandths, etc.—if the "ths" is dropped from the end of the place-values, which then become tens, hundreds, and thousands.

Another source of misconception is the decimal mark itself, its nature and its purpose, given that the learners' prior exposure to whole numbers stops at the units columns. Some learners will see the decimal number as a single whole number by

ignoring the decimal mark altogether, and thus making sense of this new form of number using whole number knowledge. There are also well-known misconceptions in seeing decimal fractions as the same as the negative numbers (Stacy, Helme & Steinle, 2001).

For this PV micro-domain, I use the misconception of seeing the entire decimal number as a whole number by ignoring the decimal mark. Other place-value misconceptions can be analyzed in a similar manner in future studies.

For Item 10003 in Figure 2 the correct response is the digit 7. However, if the learner used the whole-number misconception, then he/she would probably have selected the digit 0 as being in the thousands place, which is misusing the term "thousands" for "thousandths". This place-value misconception predicts that learners using this misconception will then have selected 0 for this test item.

Item Type PV2: Find the named place-value of a given digit in a decimal number

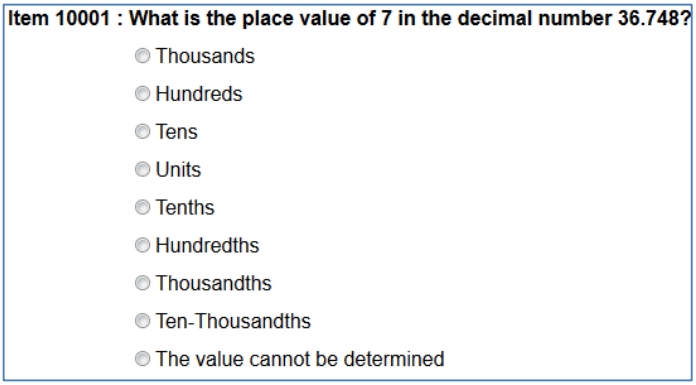


Figure 3. PV2 - Place value sample test item - select position from digit

For the PV2 test items, the goal is to recognize the place-value name for the digit selected. These test items are the logical complement of the PV1 item type above.

Item 10001, in Figure 3, asked the learner to select the place value name for the digit 7 in the decimal number 36.748. Whereas the correct answer is tenths, the learner who sees this decimal number as a whole number would select hundreds, or may have selected hundredths, confusing the meaning of the "ths" suffix. These are diagnostic test items designed to expose learners' misunderstandings by requiring the learners to unpack the decimal number into its parts, and to understand the role of each digit in a decimal number.

The item bank includes a range of PV1 and PV2 test items and there are differing numbers of decimal places in the fractional part of the decimal numbers, based upon the

assumption that the learners may already know how to unpack 2-digit decimal fractions, as is found in the local monetary representation (such as R 1.65 for one Rand and sixty-five cents) but may struggle with decimal fractions containing 1, 3, or more digits which numbers are not as familiar from real-world experience.

Misconceptions in this micro-domain

Whereas I initially considered a number of misconceptions that may arise in the development of place value knowledge, I focused on the conception that the learner sees the entire decimal number as a whole number by ignoring the decimal point in making his/her judgment about the value of a digit or its place name. This has been identified in studies which have addressed misconceptions occurring in the comparison of decimal numbers, such as Resnick et al. (1989).

Table 3. Misconceptions in the PV micro-domain

Code	Misconception Name	Description
WHOLE	whole-number knowledge	The learner treats the decimal number as a whole number, ignoring the decimal point.

Micro-Domain DO - Decimal Number Ordering

The ordering and comparison of decimal and other numbers is included within the assessment criteria in the National Curriculum Statement (DBE, 2011a). This micro-domain consists of test items which are used to identify misconceptions in decimal numbers, and consists of problems which ask learners to select the smallest or largest from a set of two or more decimal numbers. This micro-domain has been studied extensively in rational number misconceptions research (Sackur-Grisvard & Leonard, 1985; Resnick et al., 1989; Steinle & Stacey, 2005; Steinle, 2004a) and has proven value in eliciting a range of misconceptions held by learners as they attempt to make sense of the decimal numbers. For my study I have restricted the test items to problems in which the learner is required to select the smallest or largest from a set of two or five decimal or whole numbers. I have not included the more challenging problems which require the sequencing of a set of numbers into ascending or descending sequence, as was used by Sackur-Grisvard and Leonard (1985).

Item 10021 in Figure 4 illustrates the use of staggered decimal structures, in which the learners were presented with decimal numbers with differing numbers of digits in the decimal fraction, such as where 2.39 has two decimal fraction places and 2.4 has only one

decimal fraction place. This type of comparison helps to identify misconceptions in learners who have not yet reached the STABLE stage in this micro-domain. A single test item of this type will not provide enough information to determine a learner's misconceptions or development stage, since it is not possible to determine from a single item whether the learner was guessing or was applying some conceptual basis to his/her selection. Thus a number of such test items were needed to provide sufficient evidence—which is my research question RQ2.

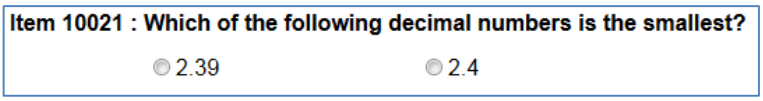


Figure 4. Decimal ordering sample test item: two choices

In addressing problems in this micro-domain, learners will make use of many conceptions and misconceptions, and when an incorrect response is given by a learner it is not evident which misconception has been used or whether the learner has simply guessed the answer. Steinle (2004a) has documented a range of behaviours used by learners to address these observed problems and has analyzed how these behaviours change with student age and grade. My purpose in this study is not to contribute further to this extensive body of prior work into these misconceptions, but rather to explore which particular test items can detect misconceptions more effectively (RQ1) and efficiently (RQ2). Consequently, I am limiting my study to a selection of the ways of thinking documented by Steinle (2004a), which is the most comprehensive analysis of this type of misconception conducted to date. As an example, in Steinle's L1 misconception, "whole number thinking & decimal point ignored thinking" (Steinle, 2004a, Table 3.1, p.47), learners will ignore the decimal point and will consider 1.53 as either two whole numbers of 1 and 53 (whole number thinking) or will consider 1.53 as 153 (decimal point ignored).

Whereas Steinle (2004a) has used pairwise comparisons, other researchers have used alternative formats with more than two choices such as item B10 from the published item set from the TIMSS 1999 study (Mullis et al., 2000; NCES, 2015), which was noted by Kilpatrick et al. (2001), and which was described previously in Chapter 2. This B10 item asks the learner to select the smallest from the set of choices (1) 0.675 (2) 0.5 (3) 0.375 (4) 0.25 and (5) 0.125, where these choices include rich distractors which can expose misconceptions in the decimal numbers. However, even this widely-cited test item may be faulty as a diagnostic item, and Steinle (2004a) has pointed out that the correct

response of 0.125 may be selected by "denominator focused thinking" which is indistinguishable from true proficiency when using this item alone. Thus a single test item used alone is often insufficient for accurate diagnostic assessment when considering the range of ways of thinking and it is thus necessary to determine how many such questions may be required. When there is evidence of many distinct ways of thinking then this challenges the requirements of "semi-dense" items (Bart et al., 1994) which are deemed to be qualitatively sufficient criteria for diagnostic items. For situations with a large number of identified fine-grained misconceptions it is not possible to embed all such misconceptions into a single test item.

I have used a general template for both two-choice and five-choice test items, using a range of misconceptions to construct the rich distractors, as well as providing random distractors. For instance, the example presented in Figure 5 includes choices that include both leading (0.075) and trailing (0.090) zeroes in the decimal fraction, which are not present in the B10 item from the TIMSS 1999 study as cited above.

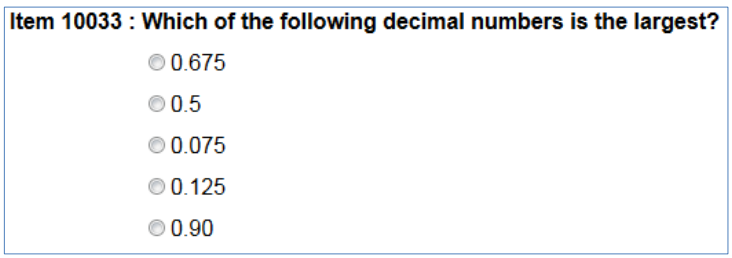


Figure 5. Decimal ordering sample test item with five choices

Misconceptions in this micro-domain

I have used Steinle's (2004a) detailed coding structure for misconceptions in decimal numbers. This is described, using my own examples, in Table 4 below. However, all of these codes are attributed to Steinle, and I have not contributed further to this coding structure in this study, but have rather explored how these can be used to answer my research questions.

These codes were applied to each of my 28 test items for this micro-domain, and in cases where a test item did not include provision for one of the misconceptions then that item was not used for the analysis of that misconception. Thus every choice for each of the 28 test items was linked to zero or more of the misconception codes, as well as being identified as the correct choice where this applied. In many cases the correct choice could be accounted for by one or more of the misconceptions and this required further

analysis to determine which misconception the learners were likely to be using on the balance of the evidence of their response to the total set of items. The assumption is that a learner's results are likely to be consistent if he/she is using a well-developed misconception. Due the fine-grained nature of this classification scheme, it is common that a single choice is accountable for by more than one of the codes.

Table 4. Steinle's (2004a) Classification of decimal ways of thinking

Code	Description	Example
A	**expert mode**	
A1	task expert	Score high on most examples
A2	money thinking	Score high, and truncates or rounds to two decimal places as for money
L	**longer is larger**	
L1	Whole number thinking and decimal point ignored	Whole number thinking: see 1.53 as two numbers 1 and 53 in which only one part is used for answering, such as choosing longer decimals as larger. Decimal point ignored: 1.53 > 1.8 since 153 > 18. This also includes the value of trailing zeroes, so that 1.40 ≠ 1.4 These two ways of thinking are not treated separately in Steinle's study.
L2	column overflow	0.70 > 0.7 since 70 is 70 tenths, in which the unit value is determined by the first non-zero place in the decimal fraction. Thus 2.39 > 2.4 since 39 tenths is more than 4 tenths.
L3	reverse thinking	0.37 seen as 7 tens and 3 units or 73. Removing the "ths" in the place-value. Trailing zeroes are ignored since they contribute no value to the number produced.
S	**short-is-larger**	
S1	denominator focused thinking	Since one tenth is larger than one hundredth, then any number of tenths is larger than any number of hundredths. 1.999 < 1.50 < 1.3 since 999 thousandths is smaller than 50 hundredths which is smaller than 3 tenths, thus in increasing sizes : 1, 1.101, 1.998, 1.11, 1.49, 1.2, 1.8, 2.
S3	reciprocal thinking	In which 3.86 is seen as 3/86 and is thus larger than 3.87 since for denominators smaller is larger. For this I ignore leading zeroes, but include trailing zeroes, so that 1.040 is treated as 1/40.

Micro-Domain CR – Common Fraction Representation

This micro-domain concerns the relationship between the word description of a fraction and its standard notation. This is analogous to my PV micro-domain for the decimal numbers, exploring the relationship between mathematical notations and the corresponding words used for these concepts.

Whereas prior studies have explored misconceptions and errors in the addition of common fractions, I have not encountered studies that address the relationship between

the textual and notational representations of common fractions, and my inclusion of this type of test item is based upon my own experience with learners in tutoring sessions, encountering their inability to see these two representations as equivalent in all but the simplest cases. This relationship between the fraction words and notations is not specifically identified as a proficiency required in the National Curriculum Statement (DBE, 2011a).

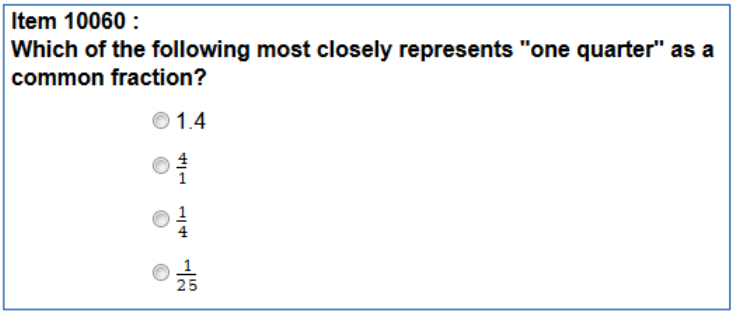


Figure 6. Common fraction word representation sample test item

As an example, Item 10060 in Figure 6 requires knowledge of both language and mathematical notations. The words used in this item, being "one quarter", should have been introduced, used, and reinforced throughout grades R-6. This test item has distractors which are designed to trip up learners who are using incorrect schemas which will cause them to make an incorrect educated guess at the choice.

- The first choice (1.4) examines the possibility that the learner is confusing the representation between common fractions and decimal numbers. Whereas there is explicit reference to "common fraction" in the item stem, this may not be understood or read properly by the learner.
- The second choice ($\frac{4}{1}$) explores whether the learners may confuse numerators and denominators.
- The final choice ($\frac{1}{25}$) is a more complex distractor which is intended to provide an alternative to learners who may know that the decimal number 0.25 equals one quarter.

I position this CR micro-domain as addressing EMERGENT knowledge in the larger context of the domain of the rational numbers, and the items within this micro-domain are designed to distinguish between learners who have no usable prior knowledge, being in the ABSENT stage, from those learners who have started to develop some conceptions and schemas, being the EMERGENT stage in my model. This

explanation illustrates that development stages exist at various sizes and levels of mathematical domains, such as being applicable to the entire domain of the rational numbers as well as to micro-domains that are small component elements of the rational numbers. Each of the micro-domains, such as this CR micro-domain concerning common fraction representation, will have its own development stage structure which leads to proficiency.

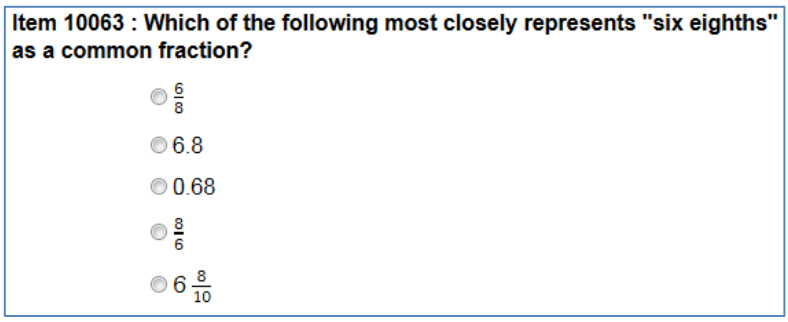


Figure 7. Fraction representation sample test item

Item 10063, as presented in Figure 7, was more challenging than the Item 10060 presented in Figure 6 since it required an understanding of the difference between the numerator and the denominator, rather than simply understanding the concept of "one quarter" as a unit fraction as in the previous example.

Whereas the first choice was the correct option, each of the other choices may have elicited a response from a learner based upon his/her level of understanding. For example, choice 3 (0.68) was a likely choice if the learner confused common fractions with decimal fractions and had not yet developed the conceptual models to see these as distinct from each other.

This proficiency should have been established in the Intermediate Phase in Grades 4-6, for which the curriculum includes the understanding of common fractions with 1- and 2-digit denominators. The test items in this group helped to determine whether the learners had incomplete schemas concerning the core knowledge of common fraction notation. Such incompleteness in their knowledge would have a serious impact on other fractional work, such as with the addition of two common fractions, and the conversion from common fractions to other rational number representations.

Misconceptions in this micro-domain

I have identified two misconceptions for analysis of this data, presented in Table 5 below, and I consider the possibility of discovering other misconceptions from the data

arising from the online tests. There is naturally some overlap between the misconceptions in this micro-domain and in other micro-domains, such as the usage of whole number thinking which applies to various types of problems. Whereas the detailed coding structure of Steinle (2004a), as outlined in Table 4, was developed primarily on the basis of decimal number paired comparisons, these ways of thinking may also be applied to common fractions. For this study I separate these micro-domains and explore specific misconceptions as applicable to each, rather than exploring common ways of thinking across micro-domains.

Table 5. Misconceptions in the CR micro-domain

Code	Misconception Name	Description
RECIPROCAL	fraction reversal	Selecting $\frac{4}{1}$ instead of $\frac{1}{4}$.
DECIMAL	decimal leakage	Selecting the decimal number rather than the common fraction

Micro-Domain NL - Number Line and Common Fractions

The number line is used throughout the early school grades, from Grade 1 up to Grade 7, as a visual image and artifact to aid learners' understanding of number systems. Commencing in Grades 1-3 in which small whole numbers are depicted and manipulated on number lines, this moves to the representation of common fractions on number lines Grade 4-6, and the representation of decimal fractions in Grade 7. Whereas there is no explicit statement within the curriculum that the number line can be seen as a common basis on which to compare different types of numbers, I expect that this is an implicit outcome of using the number line for different purposes.

Thus the number line is used to aid the representation, ordering, comparison and operations of various types of numbers. It is a dual representation, including both visual and symbolic elements which complicates the development of number line knowledge over the purely symbolic, such as common fractions, or purely visual, such as geometric or set representations of fractions as I cited earlier from Bright et al. (1988).

The number line has the potential to act as a universal representation on which all different types of numbers can be seen as parts of a single system of numbers, on the basis of their relative magnitude.

Ten test items are included for this micro-domain, of which two are on the scale 0-1, three are on the scale 0-1-2, and the remainder cover scales of more than 2 units. All

of the test items have the same question stem: "What is the value of the red arrow on the number line as a common fraction?", and each test item has four possible choices. Most of the ten items provide tick marks between the units, which are not labelled with values. Two of the items have the red arrow pointing to a place between the ticks, and for Item 10073 there are no ticks provided between the units. The test items differ in terms of the scale or range of the number line, the number of ticks between the items, and the specific choices for the learner to select from.

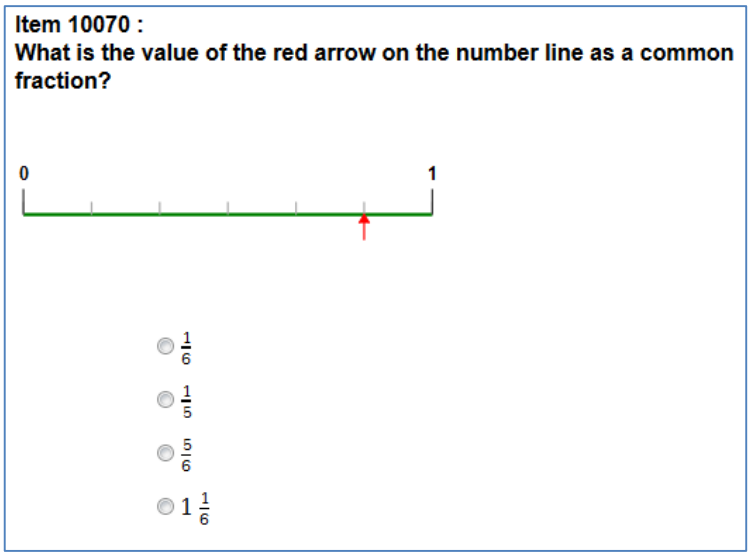


Figure 8. Number line sample test item

Item 10070, depicted in Figure 8, is presented on the scale 0-1 and has 5 tick marks between the 0 and the 1 with the arrow pointing at the 5th tick mark. The choices for this item include rich distractors which may elicit evidence of particular misconceptions, such as $\frac{1}{5}$ which may be chosen by a learner who notes that the red arrow is positioned at the 5th tick.

As outlined in Chapter 2, there has been prior research into number line conceptions and errors and there has been some identification of systematic errors that occur during learners' attempts to position a number on the number line. These misconceptions include treating the entire number line as a single unit, no matter how large the scale as marked by the numbers.

My intention in using the number line was to explore how new misconceptions may be discovered on the basis of the patterns in the learner responses. When there is a non-random consistency in the selection of incorrect choices then an opportunity is opened to discover and document the ways of thinking which give rise to these choices.

Misconceptions in this micro-domain

Two misconceptions have been included into this study arising from prior work. The first is where the entire number line is seen as a unit, and common fractions are seen in terms of the entire line, disregarding the ticks and number marks. I have called this the WHOLELINE misconception and this was reported by Novillis-Larson (1980), and further studied by Bright et al. (1988).

The second misconception is where the number of tick marks between two numbers on the number line is seen incorrectly as the same as the number of parts into which the number line is divided, so that 4 ticks on the scale 0-1 would be seen as a division into 4 parts rather than 5 parts. This has been reported by Pearn & Stephens (2007) and I refer to this as the TICKPARTS misconception.

In addition, I have included a third misconception, which I call DECIMAL being the selection of a decimal number rather than the common fraction which was requested, which was also used in the CR micro-domain under the label of "decimal leakage".

Table 6. Misconceptions in the NL micro-domain

Code	Misconception Name	Description
WHOLELINE	whole number line.	The learner sees the entire number line as a single unit, no matter how many whole numbers this represents (Novillis-Larson, 1980; Bright et al., 1988).
TICKSPARTS	ticks vs parts.	The number of ticks is seen as the number of parts (Pearn & Stephens, 2007).
DECIMAL	decimal leakage	Selecting the decimal number rather than the requested common fraction when both are available as options.

Micro-Domain CG – Common Fraction Graphics

The learning of fractional knowledge commences in the early grades with the identification of physical, tangible objects, which are used as manipulatives to illustrate the notions of equal sharing and equipartitioning. These concepts are introduced in Grade R and learning progresses through Grades 1-3 with increased sizes of the sets and groups that are to be shared and divided into parts. Geometric shapes and sets of visual objects serve as surrogates for the physical objects and fractional knowledge can be applied to these visual object in the same way as to the physical objects. Even as early as Grade R activities include the dividing up of groups of items by colour, by shape, by size, or by

other observable attributes. By the time that the learners reach the Senior Phase in Grades 7-9 they should have experienced at least six years of exposure to this type of thinking.

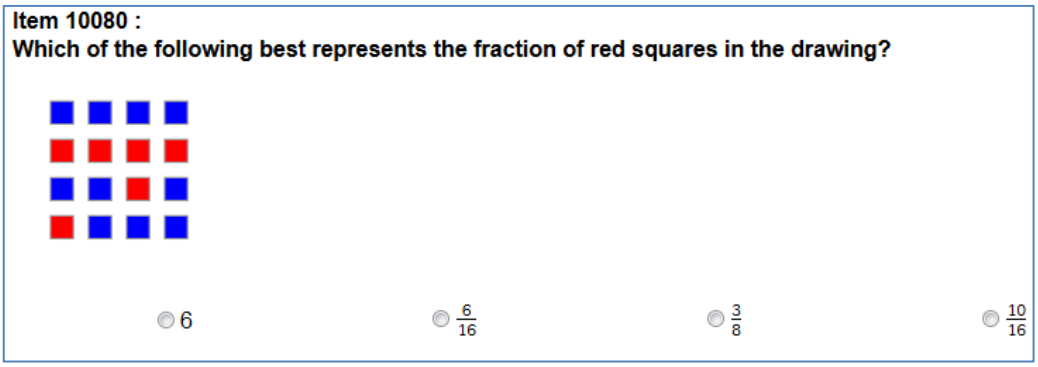


Figure 9. Graphical fraction sample test item

In Item 10080, shown in Figure 9, the learner is presented with 16 squares arranged in a 4x4 grid structure, with each square being coloured either red or blue. The item stem says: "Which of the following best represents the fraction of red squares in the drawing?", and the word "best" is included in this question to point out to the learner that there may be more than one correct choice and also to suggest that one choice may be a better correct choice than others. For Item 10080 the "best" representation could be interpreted as $\frac{6}{16}$ since this represents the diagram in which there are 6 red squares out of a total of 16 squares. The choice $\frac{3}{8}$ is the reduced value of the fraction $\frac{6}{16}$, and is not wrong, but the learners were not asked to provide the simplest or reduced form of the fraction but rather the "best". However, the term "best" is possibly not a term with which the learners are familiar in this context, and they may consider such best to imply what they are asked to do in class, which in most cases will involve reduction to the simplest form. Both of these choices are correct as mathematical expressions, but there are situations in which non-reduced forms could be better representations than the reduced form, since the $\frac{6}{16}$ form provides the added information that there are 16 units out of which 6 are identified, rather than seeing this only as a proportion.

The other choices in this example have been included as distractors to elicit different types of thinking. The choice of 6 is included for when the learner does not know common fractions, and this may provide evidence of the ABSENT stage of development. The choice $\frac{10}{16}$ has been included for learners who misread the question and think that the

fraction represents what is not selected, analogous to the confusion between fractions and negative numbers.

It was my plan to explore new potential misconceptions resulting from observed systematic errors, rather than to use any prior research within this micro-domain.

Misconceptions for this micro-domain

This is a new type of item that I have experimented with to determine if these items elicit any evidence of misconceptions and will be discussed during the analysis process and are thus not being conceptualized in advance. The analysis explores whether there were any patterns of errors in the learners' responses that may have pointed to systematic behaviours.

Micro-Domain CO - Common Fraction Ordering

Misconceptions arise in the understanding of the magnitude of common fractions, as exemplified in the following test item.

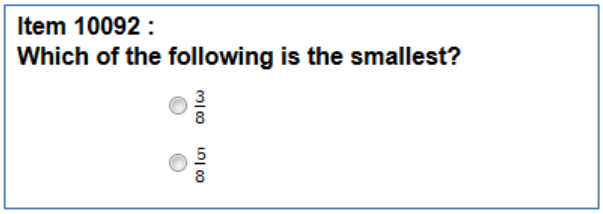


Figure 10. Common fraction ordering sample test item

Item 10092 in Figure 10 provides two choices in response to the question stem: "Which of the following is the smallest?" in which both are common fractions with equal denominators, in this case 8.

This micro-domain consists of 20 test items, each of which has exactly two choices which are both common fractions, and which have a range of complexity in the relationship between the denominators with the simplest being equal denominators as in Item 10092 above. Other types of item in this micro-domain are: denominators which are multiples of one another, such as Item 10094 which asks for the smallest of $\frac{5}{6}$ and $\frac{9}{12}$; denominators with common factors, such as Item 10096 asking for the smallest of $\frac{2}{6}$ and $\frac{4}{9}$; and finally, where the denominators are mutually prime, such as Item 10100 asking for the smallest of $\frac{7}{10}$ and $\frac{5}{7}$. These items represent increasing complexity in the arithmetic effort required to solve the problem. The first ten items ask for the smallest, and the

second ten items ask for the largest. Given that there are only two choices given, the correct wording should have been “smaller” or “larger” but this was not noted during the piloting.

Whereas the items with the same denominators can be solved by inspection of the numerator, all other forms of relationship between the denominators require some effort in computation to be performed to obtain the correct response. However, learners may attempt to use a short-cut method, perhaps derived from whole number knowledge or from other misconceptions.

Misconceptions for this micro-domain

When presented with two common fractions with a task to find the smallest (or largest) a learner may use the short-cut method of selecting either the smallest (or largest) numerator or denominator alone, and without first determining the true magnitude of the common fraction.

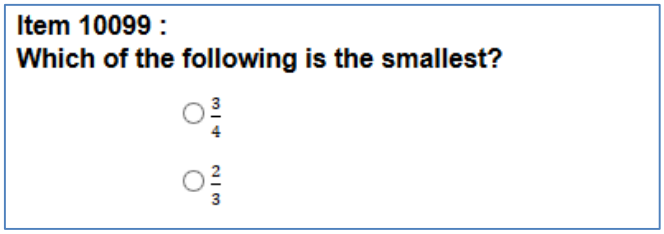


Figure 11. Sample CO test item

For example, Item 10099, as shown in Figure 11, asks for the smallest of two common fractions with different numerators and denominators. Considering that the learners would be expected to honour the requirement to find the smallest, they would select the second choice. However, in this case this is also the correct option and so there is no way to distinguish between the misconception and proficiency.

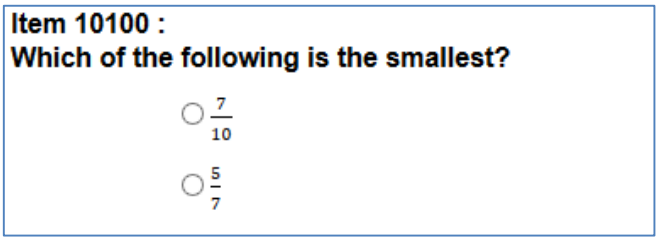


Figure 12. Sample CO test item

Another example, Item 10100, in Figure 12, would result in the second choice being selected if the learner was using a rule of selecting the denominator or numerator. This choice is the largest, but by only a very small amount.

Based upon these examples I explore one misconception, which I refer to as WHOLE which is when the learner selects the smallest (or largest) using either the numerator or the denominator of the fraction, rather than determining and using the true magnitude of the common fraction.

Table 7. Misconceptions in the DO micro-domain

Code	Conception Name	Description
WHOLE	whole number thinking	Selecting the smallest or largest depending on either the numerator or the denominator as being smallest or largest.

Micro-Domain CE - Common Fraction Estimation

This micro-domain concerns problems where the learner is required to find the closest match between common fractions and decimal numbers. I designed two types of items, with the first (coded as CE1) asking the learner to select the closest whole number or decimal number to a given fraction, where the choices were not common fractions. The second item type (coded as CE2) asks the learner to select the closest fraction to a given whole number, common fraction, or decimal number.

Given that whole numbers are addressed in the South African mathematics curriculum prior to the introduction of common fractions, I have made the assumption that a knowledge of the whole number system is more developed than the knowledge of the common fraction system for all learners in these grades. However, the same cannot be said for the decimal number system, which is introduced in the curriculum after the common fractions. The curriculum also identifies the requirement that learners be proficient in converting numbers between common fractions and decimal numbers.

The test items I introduced for this micro-domain were loosely based on items cited in prior studies, such as Kilpatrick et al. (2001) who makes reference to the misconceptions that are exposed when a learner is asked to estimate the whole number closest to the common fraction sum $\frac{7}{8} + \frac{12}{13}$ given the choices 1, 2, 19, 21, or 40. This is a complex common fraction addition operation when performed by hand, requiring the discovery of the common denominator of mutually-prime 8 and 13. This problem appears to be suited for diagnostic work, and Kilpatrick et al. (2001) report that the majority of the learners selected 19 or 21 as their answer, which is explained on the basis of the learners disregarding the common fractions and employing whole number thinking.

I simplified the form of the test items I used for estimation with 10 test items for each of the types CE1 (Items 10112-10116, and 10122-10126), and CE2 (Items 10117-10121 and 10127-10131).

These test items were structured to detect two specific whole-number-thinking misconceptions which I call NUMERATOR and DENOMINATOR. The NUMERATOR misconception is when learners use the numerator of a common fraction for estimation purposes, and the DENOMINATOR misconception is when they use the denominator of the common fraction as the basis for estimation.

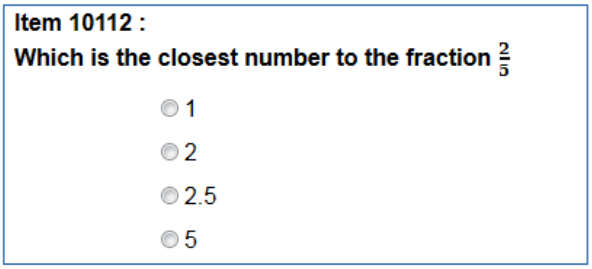


Figure 13. Fraction estimation sample: CE1

As an example of the CE1 test item type, I show Item 10112 in Figure 13, where the learner was asked to select the number which is closest to the common fraction $\frac{2}{5}$. This problem is challenging for those learners who lack a conceptual understanding of decimal numbers and common fractions, given that there is no correct choice of 0.4 as a possible answer and all of the available choices appear to be too large as an accurate estimate. The correct choice is 1, being the closest to 0.4, and the other choices are included as distractors which highlight particular cognitive behaviours, using the numerator and denominator digits of the common fraction in the item stem.

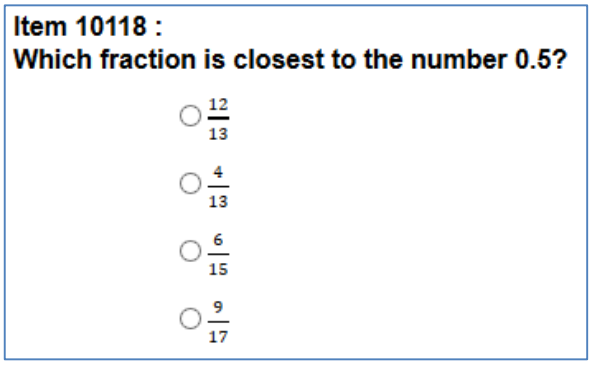


Figure 14. Fraction estimation sample: CE2

As an example of the CE2 test item type, I illustrate Item 10118 in Figure 14, and these CE2 items are more complex to solve, requiring more work by the learners.

However, given that these are problems of estimation, the intention was for the learner to identify which is the closest, without having to perform these complex calculations. The choices given in Item 10118 are relatively close to the decimal numbers 1.0, 0.25, 0.25, and 0.5, so that the fourth choice is correct.

This micro-domain is potentially a rich resource for discovery of misconceptions concerning the common fractions and I was exploring how learners attempted to answer these questions by informed guessing.

Each of the 20 items for this micro-domain had four choices, which were mostly a single type of number, such as a whole number or a decimal number, but there were two items which provided a set of choices which included a combination of whole numbers and decimal numbers.

Misconceptions in this micro-domain

Whereas this micro-domain has been identified as a useful source for the discovery of misconceptions, I have not found a prior study in which these misconceptions have been named and formalized, as is the case with decimal number ordering.

For the CE1 type of items, learners may select the choice based upon their true understanding, or may be influenced by the presence of specific digits in the various choices. I assumed that in the absence of any other information, the learners would have looked for similar digits and numbers in the numerator or the denominator. For the CE2 types of items, the data arising from the learners was used to explore the potential for identifying misconceptions.

Table 8. Misconceptions in the CE micro-domain

Code	Conception Name	Description
DENOMINATOR	use denominator	Select the choice which is equivalent or similar to the denominator.
NUMERATOR	use numerator	Select the choice which is equivalent or similar to the numerator.

Micro-Domain CA – Common Fraction Addition

This micro-domain consists of problems concerning the addition of common fractions. This type of test item has been studied as a basis for understanding learner errors and diagnostic measures for a period of at least one hundred years, and has been shown to expose misconceptions in learner thinking (for example Brueckner, 1928a).

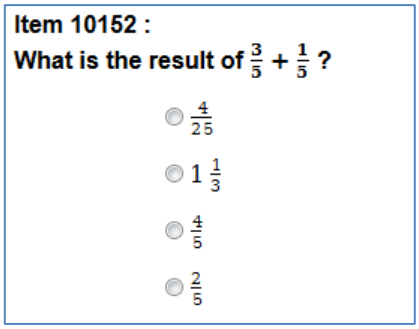


Figure 15. Fraction addition sample test item

The choices have been included on the basis of my assumptions of incorrect rules that the learner might apply, resulting from my own experience with learners, and also from some prior studies in this field. Some of these choices are rich distractors, such as $1\frac{1}{3}$ in Item 10152 in Figure 15, which bears a similarity to the item stem in the usage of the digits 1 and 3. Item 10152 is an example of the simplest case of fraction addition, in which the denominators are the same. Other items within the item bank include those with denominators in which one is a multiple of the other (such as 2 and 4); with two denominators having common factors (such as 4 and 6); and denominators which are relatively prime (such as 3 and 4, with no common factors).

Misconceptions in this micro-domain

Only one misconception is identified for analysis, which is the case where the sum of two fractions is seen as the common fraction which is produced by adding the numerators and the denominators respectively.

Table 9. Misconceptions in the CA micro-domain

Code	Conception Name	Description
ADDITION	addition	Select the choice in which the numerators and denominators are added individually rather than performing the full fraction addition.

Summary of the Micro-Domains Used in this Study

I have outlined above the range of data capture instruments for each of the micro-domains that I used for this study, exemplifying a few selected cases and identifying the specific misconceptions which these instruments were designed to elicit. These are however only eight of a large potential base of such micro-domains in the rational numbers and these specific micro-domains were used to explore the potential for common principles and practices which may then apply to other micro-domains.

4.6 Self-Knowledge of Item Difficulty

The second element of my data capture instrumentation concerns qualitative information which is available from asking the learners how difficult they find each of the items.

When learners respond to a test item in the pretest or the online assessments, they respond with some personal level of confidence in their ability to tackle the item as presented. This is self-knowledge which all learners will have to a greater or lesser extent and which informs their approach to every mathematical problem which they answer throughout their entire school career.

A knowledge of the confidence that learners have in their ability can aid the identification of the development stage in which a learner is located. To address this possibility, Huntley (2008) proposed a 2x2 decision matrix structure in which the two columns represent confidence measures of low and high, and the two rows represent whether the choice selected by the learner was correct or not. Each cell in Huntley's matrix model is used to infer whether the learners are proficient, are guessing, or whether they are using a misconception to account for their response.

I proposed a 3x3 decision matrix which extends Huntley's model of confidence by considering three levels of difficulty and three types of response. My difficulty scale consisted of Easy, Just Right and Difficult, which are words familiar to young people, and this question on perceived difficulty was included into every item posed in the pretest and in the online tests. I used these three terms (Easy, Just Right, Difficult) with initial capitals throughout this thesis to distinguish them from other uses of these words. The response scale provides three types of learner response, from the choices available in the MCQs: firstly, when the response is correct; secondly, when it is an incorrect response which has been designed as a rich distractor; and finally, when it is a random distractor.

This 3x3 decision matrix is depicted in Table 10 below and this identifies, for each cell of the matrix, which learning zone this corresponds to, as well as the specific Development Stage which is inferred.

Table 10. Decision matrix on learner confidence vs. response type

RESPONSE TYPE	LEARNER PERCEPTION OF DIFFICULTY		
	Easy	Just Right	Difficult
Correct response	*Zone of Competence* STABLE	*Zone of Competence* IMMINENT / ACTIVE	**guessing** *Zone of Incompetence* ABSENT
Incorrect - Rich distractor	**misconception** *Zone of Learning* ACTIVE / EMERGENT	**misconception** *Zone of Learning* ACTIVE / EMERGENT	*Zone of Incompetence* ABSENT
Incorrect - Random distractor	**potential misconception** *Zone of Incompetence* ABSENT	*Zone of Incompetence* ABSENT	*Zone of Incompetence* ABSENT

Table 10 is a decision matrix, identifying the decisions made given a particular learner response to an item with known types of response for each choice and distractor.

As an example of how this matrix was used: when answering a specific test item, if learners selected a rich distractor and then marked this as Easy or Just Right, then I inferred that they believed they knew what they were doing when they answered, but they selected the wrong answer. This points to a misconception, and the matrix positions them in the ACTIVE stage in which this misconception is being used actively, or alternatively in the EMERGENT stage, in which they have recently started to use this as a new schema as they experiment with ways to address the problem. However, whereas this may apply to a single item answered by a specific learner, it is unlikely that a single test item is sufficient, so these results are signs of misconceptions rather than accurate measurements.

Alternatively, if learners selected a random distractor then we cannot establish a conception or misconception that could account for their response, within the limits of our current knowledge of the micro-domain, and the conclusion is that the learners have insufficient knowledge to address the problem, with the inference that they have guessed and are in the ABSENT stage.

This approach applies to both the pretest and to the online assessments. For the pretest the following instruction is included on the front-page of the test sheet:

> Indicate whether you found the question Easy, Difficult, or Just Right for your own level of knowledge of rational numbers by placing an X into the block next to this statement… For example **[X] Easy.**

Every test item presented in the online assessment included an additional question of whether the learner found the test item Easy, Just Right or Difficult. The learners were prevented from continuing with the next test item until they had answered this question on the level of difficulty.

One important element of my study was to determine whether this information on learner self-knowledge is useful to improve diagnosis. This constitutes my research question RQ3.

I address the potential for this Difficulty Index model in Chapter 7 following the initial analysis of the data from the online assessments for the individual micro-domains in Chapter 6.

4.7 Instrumentation

Whereas this is a study on diagnostic assessment, I structured the online lessons to include not only assessment, but also to include introductory explanations, remedial content and feedback. This structure was designed to be close to the classroom situation in which formative and diagnostic assessments are integral elements of classroom activity.

In this lesson structure, the results of the diagnostic assessment, at the level of measuring learner misconceptions, were likely to be impacted by the instructional pages and the feedback. However, my primary goal, as expressed in the research questions, was to determine the value of the diagnostic items, rather than to determine the development stage of the learners as the major outcome. The development stages are introduced to determine learner stages for the end goal of understanding the nature of good diagnostic test items, thus they are a means to an end, rather than being an end in themselves for this study.

This lesson structure may also have impacted my use of the Rasch method, which I used for the analysis of the results, and which required that the items were independent of one another since it is likely that the provision of feedback and instruction will have

had some impact. However, as discussed in Chapter 2, the Rasch method provides a method of measurement which is invariant, and it is very difficult to achieve total independence of the items when all items are similar in nature and are from a single micro-domain, since even the act of answering one item may subtly inform the learner how to answer further items. This impact occurs even when feedback is not provided immediately after each test item. For this study, I made an assumption that the introduction of instructional content and feedback into the lesson structure would not impact the assessment results.

The question arises as to whether a set of tests, such as these diagnostic tests, constitutes a research method. I note that Cohen, Manion and Morrison (2007) provide guidance on the use of tests as a research method, in which they note the importance of being clear about the nature of what is being tested. They specifically cite that "diagnosis of difficulties" (p. 414) is one example of a research project suited to be conducted by means of tests. I concur with Cohen et al. (2007) that diagnostic testing, coupled with my model of development stages, is the most suitable methodology for gathering data for data analysis, given the particular focus of this study.

I previously outlined the micro-domains I have chosen to use in this study, and I have used some of my test items as exemplars of how they were used to extract diagnostic information from the learners. I now outline each of the testing instruments I have used in terms of what is being tested and measured, and how this has been developed and piloted. This section is supplemented by the outline of the lesson structure in Appendix E, the pretest structure in Appendix B, and the item bank in Appendix D. I continue with a description of how the web-based assessments were conducted.

Implementing the Web-Based Assessments

I have chosen to use MCQ items exclusively within the online assessments, since this both facilitates gathering data and is suited for detailed analysis with the Rasch method.

In my implementation of MCQ, each test item was structured with four distinct elements: (1) a question which is posed; (2) a set of choices, including a correct choice and a set of distractor choices, in which each may be linked to one or more misconceptions (rich distractors) or which are not linked (random distractors); (3) a method for selecting a choice; and finally (4) a question on learner confidence. This is an extension of the

standard MCQ structure as outlined by Kehoe (1995), who uses the term "stem" for the question part of the MCQ, and "options" for the choices.

Using the Web for mathematics assessments presents a number of challenges which result directly from the nature of mathematics in general, and from the rational numbers in particular. These challenges include the handling of mathematical notation in web pages and the presentation of mathematical graphical elements. These two special challenges must be incorporated into the preparation of both the question text in the item step as well as in the choices.

I conducted pilots of the online tests using a variety of technologies for representing mathematics and graphics and I selected the MathML (Mathematics Markup Language) standard for the presentation of mathematical notation, which at the time of this study was only available on the Firefox Web browser. This posed the restriction for the study that FireFox had to be installed on every computer used during the assessments.

I also explored alternatives for the display of fractions in graphical form. I decided to use the Scalable Vector Graphics (SVG) standard which provides support for all of the types of graphical elements I needed but which again was not widely available in the common Web browsers.

Both MathML and SVG have helped me to present the mathematics to the learners in a form familiar to them from their text books and classwork. My usage of MathML and SVG is described in Appendix C.

Lesson Structure

Four online lessons were conducted for each of the two schools, and each lesson was structured as a sequence of activities, as outlined in the introduction to this section on Instrumentation. Three types of activity were used and each activity was presented to the learner as one or more Web pages using the FireFox Web browser. These three types of online activity are:

- Information pages: these include introductory information, instructional material, and commentary before and after the tests. When the learners are finished they move to the next activity automatically.
- Test pages: each test consists of between 4-20 MCQ test items, and is presented to the learner in a fixed sequence. The learners cannot continue until they have answered the item, and they cannot move back to change an answer.

The system captures various data elements which are then stored on the assessment database:

- the choice selected by the learners for each test item,
- the learners' self-assessment of the item difficulty,
- the learners' indication of whether they don't know how to answer the test item, and
- the start and end time of the learners' responses for each item.

- Results pages: on the completion of each test the results are presented, indicating the original question together with the learner's response and the correct response, and a score for the test itself. It was not my original plan to present this, due to the impact that this may have for diagnostic purposes, as opposed to formative purposes, but the piloting indicated that the users felt short-changed by the lack of feedback after they had spent time answering the questions.

The lesson structure used for this study, including how each of the lessons is structured into these three types of activities, is detailed in Appendix E.

4.8 Piloting of the Instruments

Piloting is recommended for questionnaires (Cohen et al., 2007; Mouton, 2001) and I used a pilot implementation to assess the usability of both the individual test items and the test environment as a whole, as well as to check for any 'bugs' in the encoding of the test items. For my purposes there is no essential difference between a questionnaire, as outlined by Cohen et al. (2007), and an assessment test conducted on paper or online, so this recommendation applies directly to my work.

My sample for the pilots consisted of a group that included graduates, professionals, as well as some learners in Grades 10-12. I selected this sample on the basis that they were expected to be sufficiently proficient in mathematics to be able to comment on the content, and sufficiently Web-literate to be able to comment on the user interface for the online assessment. Some of the individuals on whom this pilot test was administered were Computer Studies graduates for whom mathematical proficiency is an entry requirement for their course of study.

Piloting the Pretest

The initial version of the paper-based rational number test was administered to the pilot group. My intention was to explore how mathematically-literate people would experience the test prior to using it on younger learners.

An informal discussion was held with each of the pilot individuals after completing the pilot test, and this helped to inform me about test items which were too difficult; items which were ambiguous; and items which did not serve a purpose. On the basis of this feedback I modified the pretest.

Piloting the Online Assessments

I used a pilot version of the online assessment for two workshops I conducted on Computer-Based Assessment at the AMESA conference in 2008 at the Nelson Mandela Metropolitan University in Port Elizabeth. The lessons learned from this pilot helped me to improve the way I presented the questions to the learners and the manner in which I reported results and provided feedback during the online sessions. I also learned more about what data can be captured online and how this data may be useful. From this, I designed the online assessment to collect more data than I need for this study, such as the response time, which may prove useful in further studies.

I conducted a final pilot of the online tests with four of the individuals who had previously helped with the pretest pilot and these were conducted on the day prior to each of the actual lessons with the classes. This helped me to make final adjustments to improve the performance of the assessments and to correct any outstanding issues with the test items themselves. This piloting resulted in the identification of some spelling mistakes and improved wording of some test items to render them more meaningful and less ambiguous. The timing of the test was checked to ensure that most of the learners would be able to complete the test in the allotted time. The pilot group also reported back on the results pages, as well as the information pages. This helped considerably in minimizing the occurrence of errors in the test items and in the result pages prior to the actual lessons.

4.9 Planning Meetings with the Schools

Planning meetings were held in advance with each of the schools, attended by the head teacher and the head of the mathematics department. This was followed up with more detailed meetings with the mathematics teachers responsible for the classes in my study, and with the technical personnel who were responsible for the computer facilities. The teachers were informed that the study involved the rational numbers, but no further details were provided on the specified tests to be conducted. At these meetings I reviewed the computing facilities, ensuring that the computers and their software were suitable. Both School A and School B have computer rooms which were used for Computer Studies lessons and which were made available for my study.

4.10 Administering the Tests

Detailed data was gathered from the tests presented to the learners in the five classes across the two study schools. A pretest was conducted for each of the classes as a paper-and-pencil test administered to all five classes. The online lessons were given to four of the classes, being one class from School A and three classes from School B, at the rate of one lesson per week.

Pretests

The pretests were conducted at the schools during a mathematics lesson. The pretest was designed to be completed within 40 minutes and, with the school periods being 50 minutes in duration, this allowed sufficient time to hand out the papers and to instruct the class on the tests. At this time, I also explained how to access the online assessments and I handed out the user codes. I made a specific point of explaining that the decimal point was being used throughout these tests rather than the decimal comma which they would have been familiar with at school. The learners indicated that they were familiar with both forms of writing decimal numbers from their school work, as well as from using spreadsheets on the computers. The learners confirmed that the knowledge of the decimal point was not an issue.

All of the learners wrote their user codes, but not their names, on the question sheet for the pretest and this user code then provided the link between the pretest and the Web-based assessments.

The results from the pretest were captured into a Microsoft Access database. These results were analyzed at a high-level during the days after the pretest, and the findings were used to inform the test items to be used within the online assessments.

Online Lessons

Each of the online lessons was conducted in the computer rooms of the study schools. For each of the sessions I arrived early and ensured that the computers were set up properly in advance.

School A - Lesson 1

In School A, the computers were running the Ubuntu variation of Linux, and FireFox was installed by default. Each of the learners was competent with using the Web for research and they had no trouble accessing the Web-based assessment engine.

I explained how to log on and to run through the tests and any learners who had forgotten their codes were reminded of these by the computer teacher, who was available throughout the assessment lessons. Most of the learners had opened the Web site and had entered their user code and password before I had even completed my explanation. I emphasized again that my tests presented decimal numbers using decimal points rather than decimal commas, to refresh what I had explained during the paper-based tests in the previous week.

Some learners finished the first online lesson in 10 minutes whereas others took the entire 50 minutes of the lesson. All of the learners in the class were quiet for the entire lesson and I saw no evidence of the learners speaking to one another or collaborating silently.

Based upon the results of this first lesson, I planned for the second lesson, with the following changes:

- Additional explanation was provided in the instructional content for the use of the decimal point in the tests as opposed to the equivalent decimal commas. This was included to minimize the possibility that some of the whole number misconceptions (such as reading the digit 5 in the number 2.45 as units rather than hundredths) were due to misreading the decimal number.
- Some of the test items used in the place-value tests in Lesson 1 were repeated to see if this explanation on the usage of the decimal point would have any

effect. Special consideration was given to these duplicate test items in the analysis, to avoid double counting.

- The number of test items was increased from 20 in the first lesson to around 40 in the second lesson.
- The size and number of instructional pages was reduced, with a focus on the explanations used between the tests. These pages are a short set of instructional pages and the manner in which these are used was outlined earlier.

A communique was prepared and emailed to the key staff involved in this research, explaining some of the tests performed during the pretest, while omitting any advance information about the upcoming tests.

School A - Lesson 2

Lesson 2 was primarily concerned with the problems of ordering decimal numbers, by selecting the smallest or largest from a set of numbers.

After Lesson 2 was completed I prepared another communique for the staff involved in this study from the school. I particularly noted an initial result showing that there was a significant improvement in the results from the previous test on the recognition of places values, being an improvement of 9% over the entire class. However, I did not discuss this in terms of its statistical significance.

School A - Lesson 3

From the experience of Lesson 2 I increased the number of questions in the tests in Lesson 3 to 60, and I focused on questions concerning graphical fractions as well as the positioning of fractions onto the number line.

School A - Lesson 4

Within this final lesson I included test items in which I asked the learners to select the closest whole number to a particular common fraction to determine the extent to which they were familiar with fraction notation.

School B

The progress for the four lessons was similar to that for School A. The number of learners was larger than for School A and involved three classes of learners. The classes were

more challenging to administer than School A since there was no teacher present during the assessment periods and on some occasions there was talking among the learners which I had to address. Issues with the computers during the class were addressed by the computer technician and did not negatively impact the lessons.

4.11 Data Collection

The data from the pretests were collected on paper test sheets that were handed out to the learners. Each individual test paper was headed by the user code which had been allocated to the learners. In some cases, when the learners had forgotten their code, they were asked to put their name on the paper, and the teacher later informed me which user code this corresponded to, since this linkage between the user code and the learner name was maintained by the teachers.

For the online tests, the data was collected directly from the Web server, which automatically tracked the following:

- Which learner was logged on (by their user code).
- Which lesson, which test, and which item.
- The start and ending time for the question—in which the start time was the time when this was presented to the learner, and the ending time was when they selected the answer.
- The answer that they selected from the MCQ choices.
- An indication of "Don't Know" if they had selected that option instead of selecting an answer.
- Their indication of whether this question was Easy, Just Right or Difficult.

4.12 Data Capturing

The pretest responses were captured into a Microsoft Access database, and following the data capture, random checks were performed on a few papers to ensure that there were no errors in the data capture.

The data from the online tests was captured automatically by the system and stored in an SQL Server database (School A), and in a MySQL database (School B).

All of the data, for both schools and for both the pretest and the online tests, was merged into single Microsoft Access database which was used as the basis for the detailed data analysis.

4.13 Errors and Limitations in the Research Data Gathering

The data gathering processes were subject to various limitations and issues which may have impacted on the quality of the data collected.

For School B, some of the computers did not work as required, and did not have the correct software, and this resulted in some learners being delayed in starting their tests. In a couple of cases this could not be resolved before the lesson was complete, and there were not enough spare computers.

In School B, I noticed some initial collusion between the learners in sharing information. It was not possible to completely avoid learner-to-learner discussions taking place in the class during the tests, and it was also not possible to prevent the learners using the Web for other purposes. I specifically did not record the names or codes of the learners who were engaging in other activities since I felt that this may risk the requirement for anonymity. I also did not record who was sitting close to one another, and as a result all learners are considered as working independently for my analysis. However, in my opinion, these issues were kept under control and did not impact the online lessons or the data collection.

4.14 Rasch Analysis

Having collected the raw data, these needed to be analyzed to assist in answering the research questions. For this analysis I have chosen to use the Rasch method, which is highly suited to educational measurement (Bond & Fox, 2012). I have adapted the Rasch method to suit my requirement to measure the diagnostic properties of test items, and I outline this below. IRT also meets my requirements but is better suited for larger data sets, whereas Rasch was better suited for this study (Wright, [2005]).

I chose to use the WinSteps program (Linacre, 2013) which provides a wide range of outputs and calculated measures for the analysis of test item responses and learner proficiencies.

The inputs to a WinSteps program execution are:

- the set of test items used, each identified by a code
- the set of the learners, each also identified by a code, and
- the set of learner responses to each test item

The learner responses are coded as 0 (for an incorrect response) or 1 (for a correct response). Missing values are accommodated and are coded using a period (.) and are automatically omitted from the analysis, with the Rasch analysis providing measures of learner ability and item difficulty while accommodating missing values. This approach is detailed in Appendix F.

The primary outputs from the Rasch analysis were the measures of the test items on a scale of difficulty, centered on zero, and the grading of the learners using a calculated ability scale. The scales of item difficulty and learner ability were aligned and able to be compared, which is a unique feature of both the Rasch method and IRT.

Another important output from Rasch analysis is the fit of the learner response data to the Rasch model, which occurs as a by-product of the manner in which the Rasch method finds the best-fit model to the input data as provided. As a result of this process of best-fitting to the totality of the data, there are data that fit the model and other data that do not fit, and which are considered as outliers in the statistical sense. This data included both person and item data. The Rasch method structures test items on a scale of "difficulty" and the learners onto a scale of "ability". The use of misfitting data can distort the Rasch model and it is recommended by Linacre (2013) that these be removed prior to re-running the Rasch model. Misfitting test items are those for which the results are inconsistent with the Rasch model, and these items would have distorted the model to some degree if they were retained. Misfitting items were identified based upon their INFIT or OUTFIT statistics. INFIT is calculated based on an increased weighting of the items which have a measure close to the learner measure, thus measuring the behaviour of items around a learner's ability measure. OUTFIT is unweighted, and is more sensitive to outliers, such as guesses and slips. For both INFIT and OUTFIT, the mean-square value (MNSQ) was used, with values greater than 2.0 having the potential to distort or degrade the measurements made (Linacre, 2002).

Misfitting learners were also removed, for cases where the learner responses did not fit the data results, and such misfits in learner data may have been an indication of

guessing, where the relationship between learner capability and item difficulty cannot be determined due to inconsistency in the responses from these misfitting learners. My justification for removing learners for my analysis was that my research questions are focused on the test items rather than on the learners. I note that the removal of misfitting data is contentious in the case where accuracy in measuring learner ability is required, since such anomalies may point to other factors which are beyond the capability of the Rasch model, and may also point to deficiencies in the item bank. As an example, if a learner scores well on the difficult items, but then scores poorly on the easy items, then this points to an anomaly beyond the capability of the Rasch model.

My approach was to separate the analysis of high-performing learners from the lower-performing learners, and I argue in this study that the assessment methods suited for these two groups are different, with the first group focusing on the measurement of ability and the second group focusing on the measurement of specific conceptual models that cause the lack of performance.

My approach thus used Rasch analysis in two different and sequential analyses. Firstly, I used a traditional Rasch analysis to determine the proficiency of learners on the basis of the test items presented. I then removed these high-performing, proficient, learners from the set of learners that I took into the second analysis, in which I analyzed the incorrect responses in more detail than for the first analysis. My justification for this removal is that if a learner demonstrated proficiency, which is the STABLE stage of my model, then by definition they would not have made mistakes that warrant the analysis of their misconceptions. Any such mistakes made by otherwise proficient learners were treated as slips, as noted by Olivier (1989).

For my analysis of the low-performing learners I applied the Rasch method by addressing particular misconceptions that were linked to test items as the rich distractors in the MCQs. I analyzed the learner responses to determine which test items are good indicators of each codified misconception, and also which learners showed evidence of the usage of the various misconceptions. As explained earlier, the trait I was measuring is not “ability”, in its traditional meaning, but rather the extent to which learners’ responses are accounted for by their usage of a particular way of thinking.

Stacey and Steinle (2006) have argued that Rasch analysis is inapplicable for diagnostic purposes when used in its traditional approach to measure ability. They claim

that “there is nothing to gain in following that [Rasch] approach in this and other cases” (p. 89), and their particular application was the DCT2 diagnostic test which is designed to elicit evidence of decimal number misconceptions. Stacey and Steinle’s claim is based on analysis of their results using a Rasch analysis, and they provide an explanation that the Rasch requirement for trait unidimensionality is not satisfied with tests which are designed to discover multiple traits, as is the case within all diagnostic tests designed to reveal multiple misconceptions. Rather, Stacey and Steinle propose a model based on “mapping learning” rather than “measuring learning” (p. 78), and suggest that “learning as revealed by answers to test items is not always of the type that is best regarded as ‘measurable’” (p. 89). When Rasch is used to measure misconceptions in the same way that it is applied to measure ability then I agree with the conclusions of Stacey and Steinle (2006).

However, I argue that Rasch can be used effectively for diagnostic analysis, and my mitigation for Stacey and Steinle’s (2006) claim of inapplicability is to conduct Rasch analyses in parallel for each identified misconception, so that the analysis then comprises a parallel set of unidimensional analyses. The resulting measures are then provided for learner ability, as well as learner usage of specific misconceptions to account for their responses. On this basis, the most likely misconceptions can be determined for each learner to account for their responses.

To process the data I used the WinSteps program (Linacre, 2013) the details of which are included in Appendix F.

4.15 General Approach to Data Analysis

I outline here the general approach I adopted for the data analysis of the online assessments, to answer the research questions. This general approach consists of a sequential set of steps which are applied for each of the micro-domains, and which was adapted as required for the individual micro-domains.

- I first analyzed the response patterns over all the test items and all the learners, using a frequency table, to see which choices have been selected more than others and to identify if there were any unexpected high frequencies in incorrect choices that were not previously identified as rich distractors. If the misconceptions were well understood, and were reflected in the choices of the

test items, then this step should reveal that there were high frequencies of responses for the correct choices and for the rich distractors, but low frequency responses for the random distractors.

- I then identified and extracted the learners who were proficient in the micro-domain and classified them in the STABLE stage of development. This was conducted using a Rasch analysis on the test items by encoding the learner responses as either correct or incorrect. Learners whose ability score was beyond a specified cutoff measure were placed into the STABLE development stage and were removed from the set of learners for further processing of evidence of misconceptions. During this process any test items that showed poor correlations or were misfits for the measurement of ability were removed and for this, and throughout this analysis I used the point-measure correlation as provided in WinSteps (Linacre, 2013) which shows the extent to which the item responses are in line with the learner measures.
- For those learners in the STABLE stage, there would still be a few test items that were answered incorrectly. These responses were classified as slips, but may be also seen as "late-stage" misconceptions when they are selected consistently by learners who are otherwise proficient. In such cases, the learners were better considered to be in the IMMINENT stage. However, for the purpose of this study decisions were made on a case-by-case basis, since there was insufficient evidence to determine whether a mistake was a slip or a misconception.
- I next identified those learners who were in the IMMINENT stage, being those learners who mostly responded correctly, but who had a sufficiently high number of incorrect responses that they could not be considered to be fully proficient. Those learners in the IMMINENT stage were close to proficiency, and this step explored the nature of their incorrect responses for evidence of known misconceptions. This also helped to determine which misconceptions occurred in the latter stages of cognitive development in this micro-domain. This was similar to the analysis of the proficient learners, but in this case there would be far more data to analyze. These learners were identified based upon their Rasch measure of ability, and those learners with an ability score in the

range 1.0-1.5 were generally selected into the IMMINENT stage. These learners were not removed from the list to be studied further, since their misconceptions would be analyzed together with the remaining learners.

- The set of learners who were not in the STABLE stage were analyzed using parallel Rasch analyses, one analysis for each of the identified misconceptions, to categorize the learners into the ACTIVE, EMERGENT, and ABSENT stages.
- The learners who showed no consistency in their responses in terms of the misconceptions were positioned into the ABSENT stage, since from the data provided I could not detect the presence of either conceptions or misconceptions that could account for their responses. In effect, as far as the test results were concerned, the learner responses were random, resulting from pure guesswork with no systematic basis.
- The remaining learners were those who were within the ACTIVE and EMERGENT stages of development, with the difference being that the ACTIVE stage learners were achieving around 50% proficiency and their incorrect responses could be identified as misconceptions. The EMERGENT stage learners would have some evidence of using misconceptions, as distinct from the ABSENT stage learners.

I now continue with the analysis of the pretests, which provides the lead in to the full analysis of the online test data in Chapter 6.

CHAPTER 5 :
DATA ANALYSIS AND RESULTS - PRETEST

In this chapter I present the test results from the pretest using bar charts for questions with a small number of categories of responses and a tabular format for the constructed-response items. When reference is made to the number of learners who selected a particular choice, I use the notational form of N/M (for example, 23/76), meaning "N learners out of M", and these results are also presented as percentages where appropriate. In situations in which there is reference to a common fraction the notational form of $\frac{6}{10}$ is used. When identifying particular responses from constructed-response questions these are often placed in quotation marks, such as "54.", to clearly distinguish these responses.

The data from the pretest were manually captured into a Microsoft Access database from the paper answer sheets completed by the learners. For each of the questions presented in this chapter, suggestions are made as to how the responses may be used to assess the learner development stage.

The pretests results were used as an aid to understand the nature of the misconceptions used by the learners in classes used in this study, and to help to select the types of test items for the online assessments.

5.1 Pretest Results and Question-Level Analysis

A pretest of 15 test items was conducted with learners in both of the study schools and each item was included for its potential to expose misconceptions.

Each of the test items used in the pretest is now explained, with each representing a potential micro-domain. The full pretest paper is provided in Appendix B, with explanations on the source of the items and the reason for inclusion for each of the items.

I use the term "choice" to indicate a particular choice provided within the MCQ questions and I use the term "response" to indicate a learner's selection of a particular choice.

For each of the questions I present the response counts, and in some cases I supplement this with a summary table of the difficulty levels for each item as indicated by the learners as being Easy, Just Right or Difficult.

Question 1: Common Fraction Ordering

In this Question I asked the learners to select the larger of the unit fractions $\frac{1}{4}$ and $\frac{1}{6}$. This question is below the curriculum level of proficiency expected for Grade 7-8 learners, and the failure of a learner to answer this item correctly may point to a serious lack in conceptual understanding of the common fraction system, but it may also be the result of a guess or a slip. Even with such a simple test item there is a linkage back to the development stages of my model. A learner who answered this correctly may have been in the STABLE stage, even though guessing may account for this result. A learner who made a slip may be in the IMMINENT stage, and a learner who selected the correct response but based upon the use of a misconception, would be in the ACTIVE stage. A learner who guessed would be in the ABSENT stage. Whereas this single question cannot provide sufficient information to determine the development stage of a learner, the response can provide some useful information.

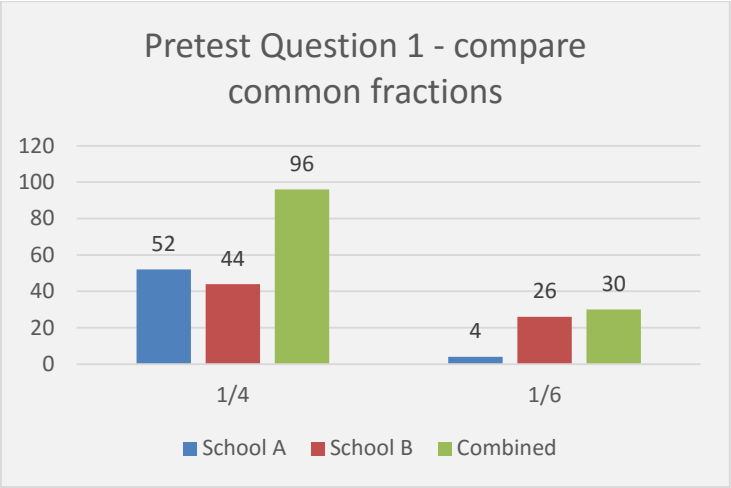


Figure 16. Pretest Question 1 - compare common fractions

Results

One of my observations from these results is the large proportion of the learners from School B (Grade 8) who selected $\frac{1}{6}$ as being the larger (26/70), compared to School A (4/56). However, it was not my purpose in this study to compare these two schools, given that my focus was on the test items themselves and their diagnostic potential.

Analysis

During the design of this study my expectation had been that this item would be answered correctly by most of the learners. However, I had previously discovered that

such an apparently simple item can elicit evidence of misconceptions, arising from the pilot conducted with adults and some Grade 12 learners.

There is one well-known misconception for the selection of an incorrect result for this example—that whole number reasoning is being applied so that $\frac{1}{6}$ is seen as the larger number since $6 > 4$. In the absence of any other explanation, the simplest explanation is likely to be the correct one—which is a method of reasoning known as Occam's Razor (Metaphysics Research Lab, 2014). Even if there are other plausible explanations, a question as simple as this will be an aid in eliciting this misconception. However, with only two choices, this item is also prone to guessing, with a 50-50 chance of a learner selecting a correct choice even if they have no knowledge of common fractions.

The significant values highlighted in Table 11 show that 28 (12+16) of the learners who selected $\frac{1}{6}$ as being the largest also indicated that they found this question either Just Right for their level of ability or that they found this Easy. Only two of the learners who selected this incorrect choice found this item to be Difficult. Also, only one learner who selected $\frac{1}{4}$ as his/her choice indicated this as Difficult, which is an indication of guessing. The term "N/S" indicates the count of learners who did not select the level of difficulty. The learners were asked to write "D/K" to represent "Don't Know", rather than simply resorting to guessing.

Table 11. Pretest Question 1 - difficulty levels

OPTION	N/S	Difficult	Just Right	Easy	TOTAL
D/K	8				8
$\frac{1}{4}$	2	1	14	79	96
$\frac{1}{6}$		2	**12**	**16**	30
TOTAL	10	3	26	95	134

When the learner response is coupled with the indication of how difficult they found the test item, more information is obtained that may improve our inference of the development stage. From the analysis of those who marked this item as Just Right or Easy, it is possible to not only gather evidence that 30 of the learners selected the incorrect option of $\frac{1}{6}$ but also to determine that of these, 28 (12+16) are likely to have answered this

on the basis of a misconception they held about the nature of fractions, when the additional information on difficulty is also considered.

Question 2: Fraction Estimation

The second item asked the learner to estimate the magnitude of a common fraction addition, where the addition was difficult to calculate in full. This item assesses the learners' conceptual understanding of the magnitude of common fractions beyond what can be inferred from the test item used in Question 1.

Question 2

Which is the closest to the sum $\frac{7}{8}+\frac{12}{13}$?

A. 1
B. 2
C. 19
D. 21
E. 40

This question has been cited by Kilpatrick et al. (2001) and it appears to be a good diagnostic item. My research question RQ1 asks why this item is "good" for diagnostic purposes when compared to other items, so that I can better understand the nature of such good diagnostic questions, with a measurable scale of "goodness" in terms of their capability as tools to elicit evidence of misconceptions.

Kilpatrick et al. (2001) note that the responses of 19 (choice C) and 21 (choice D) were selected by learners who lacked a full conceptual understanding of the common fractions, and who resorted to adding either the numerator or the denominator. To fully calculate this sum $\frac{7}{8}+\frac{12}{13}$ requires the learner to determine the common denominator of the mutually-prime 8 and 13, and then to perform the necessary multiplications of the numerators to arrive at a point where they have the same denominator and thus can be added. However, an exact answer for this sum is not required, as suggested by the word "closest" in the question, and which in turn requires that the learner has a sufficient conceptual understanding of the problem statement to plan their solution.

Results

This question produced similar results to those reported by Kilpatrick et al. (2001), with options C, D, and E eliciting 42, 19, and 25 responses respectively from the combined data set of both schools, where choice E (40) is the sum of all of the numerators and denominators in the fractional sum. The correct choice B was selected by only 18 learners—fewer than any of the individual choices C, D, or E.

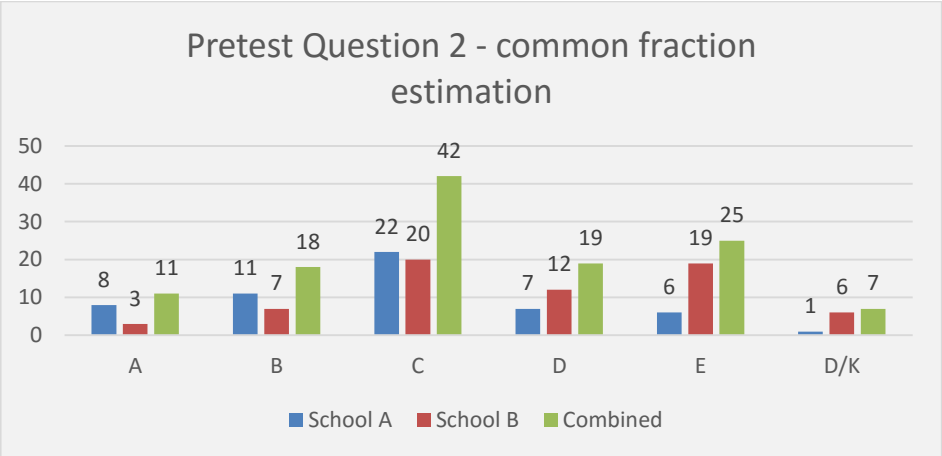


Figure 17. Pretest Question 2 - common fraction estimation

Analysis

Whereas this result confirms prior studies, my interest was to determine to what extent this type of test item could be used for the diagnosis of misconceptions in common fractions. The response distribution in Figure 17 provided an indication of the suitability of this type of item for diagnostic purposes, since the largest choice C is based upon the misconception that learners see the addition of common fractions as the sum of their numerators. I consider that one criterion for item suitability is that learners who held this misconception would select those choices which were introduced as rich distractors linked to this misconception. When this is the case then these items are good indicators of this misconception. Whereas we do not know in advance the learners who hold particular misconceptions, a Rasch analysis can simultaneously calculate both learner and item measures, and thus can provide the basis for estimating the extent to which a learner used this misconception. A qualitative analysis of this question also shows that the introduction of 19 as a choice had the potential to provoke selection by learners who would tend to add the numerators when adding fractions. Thus there were two possible approaches for analysis of suitability, one statistical-quantitative and the other qualitative.

Question 3: Decimal Fraction Place-Value

What is the place value of the digit 7 in the number 0.06758

a. Tens
b. Units
c. Tenths
d. Hundredths
e. Thousandths
f. Ten-thousandths

This question explored misconceptions in decimal number representation, in the extent to which a learner understands the place-value of particular positions in a decimal number.

The results in Figure 18 show two peaks. The first peak is the correct choice "thousandths" (E), and the second, and marginally higher, peak is "hundredths" (choice D) which was selected by the majority of the learners in both schools. This is explained by learners treating the decimal number as a whole number, effectively ignoring the decimal point. This is essentially the same conceptual basis as for Question 1 of this pretest.

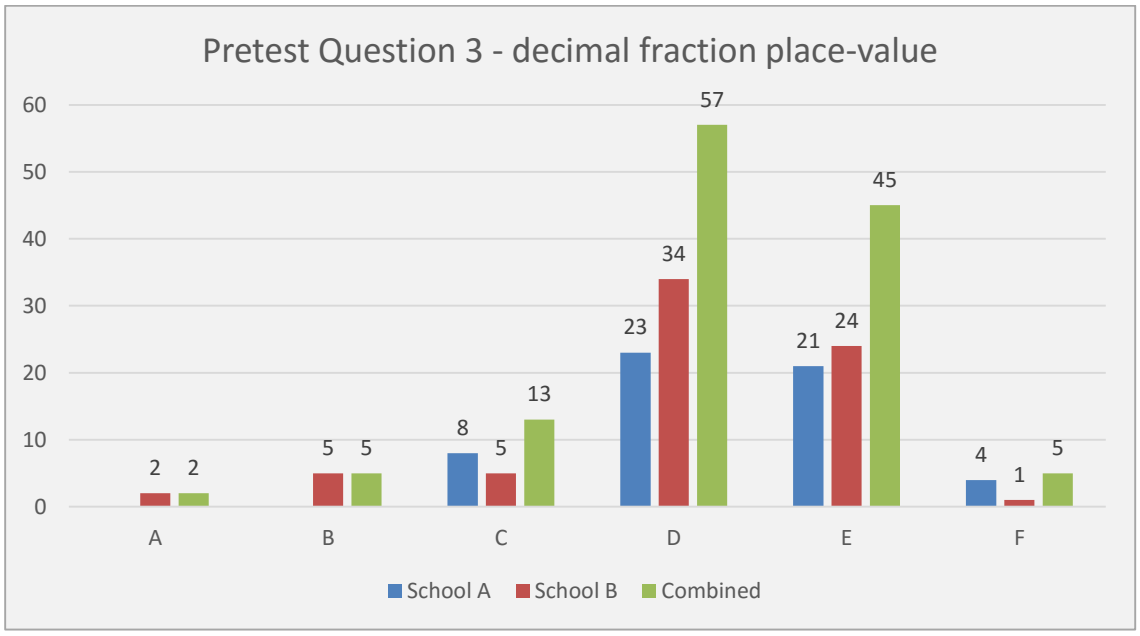


Figure 18. Pretest Question 3 - decimal fraction place-value

This test item can be analyzed in terms of the self-knowledge of the learners in terms of their perceived level of difficulty. Most of the learners responded to this item,

and Table 12 shows how many learners identified this question as being Easy or Just Right, as highlighted in red.

Table 12. Pretest Question 3 – difficulty levels

OPTION	N/S	Difficult	Just Right	Easy	TOTAL
N/S	7				7
A	1			1	2
B		2	1	2	5
C		2	6	5	13
D		2	**23**	**32**	57
E	1	1	9	34	45
F			1	4	5
TOTAL	9	7	40	78	134

Analysis

These results highlight a lack of knowledge of place-value in decimal fractions, where learners revert back to their knowledge of place-value in whole numbers.

The results from this question indicate that this type of item may be suited to elicit evidence of place-value misconceptions for learners who are starting to know the decimal numbers. This is of importance to rational number learning since any incompleteness in place-value knowledge will impact all situations which rely on this knowledge, both in education and in the real world.

Choice D (hundredths) was selected by 57/134 of the learners, most of whom (55) indicted the level of difficulty as either Just Right or Easy. This choice may be selected for two possible reasons. Firstly, that the learner ignores the 0, then sees the 6 as tenths and the 7 as hundredths. Secondly, that the learner ignored the decimal point altogether, and treats the number as a whole number, seeing the 7 as hundreds, and then selecting hundredths as the most likely choice. The challenge is that a single choice may result from many ways of thinking, and this particular item may not be a good diagnostic item to identify place-value misconceptions, considering the qualitative requirements for good diagnostic questions as indicated by Bart et al. (1994).

Question 4: Ordering of 5 Decimal Numbers

Which of the following is the smallest?

A. 0.25
B. 0.125
C. 0.5
D. 0.675
E. 0.375

I have derived Question 4 from item B10 in the TIMSS 1999 study (Mullis et al., 2000; NCES, 2015).

Results

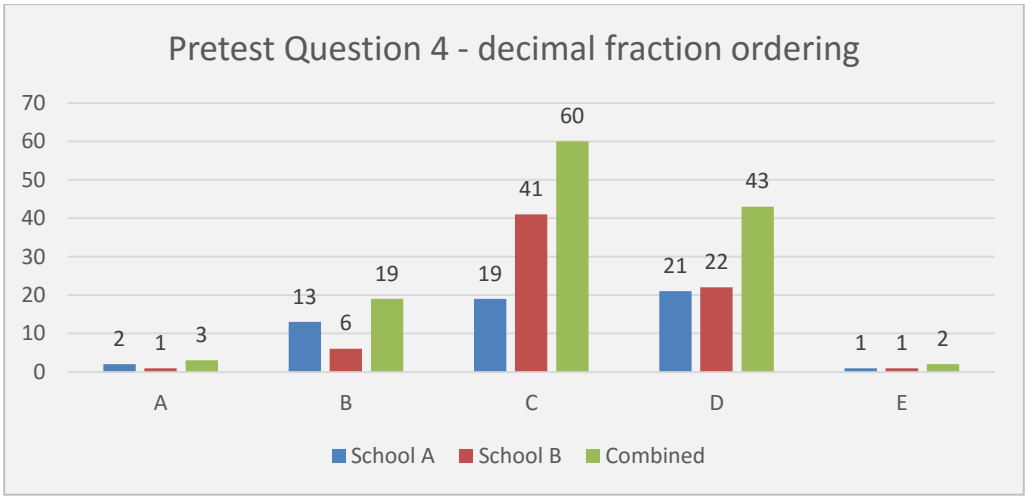


Figure 19. Pretest Question 4 – decimal fraction ordering

Whereas the correct choice is B, there are far more responses for choices C and A. Choice C (0.5) can be accounted for by whole-number reasoning, in which the decimal point is ignored, which is one variation of Steinle's (2004a) L1 misconception. Choice D (0.675) may be accounted for by various misconceptions, with one being where these choices are seen as negative numbers, where -675 is the smallest of the set, and another where these numbers are seen common fractions, where $\frac{1}{675}$ is the smallest fraction. These ways of thinking have been documented by Steinle (2004a) as her S3 code.

Analysis

This type of test item, comparing the relative magnitude of decimal numbers, has proven from prior research to be a useful instrument for identifying and documenting

misconceptions, and thus is useful for further study into their effectiveness for diagnostic purposes.

The challenge is to determine which particular test items are better suited to elicit which of the known misconceptions, so that this information can be used to help position the learners into one of the Development Stages of the model.

Given that choices C and D feature prominently in the responses, it is likely that these represent specific stages in the conceptual development of decimal numbers. Choice C (0.5) is an indication of whole number thinking, essentially ignoring the decimal point altogether, which, like the pretest question above, represents the least-developed conceptual model, whereas option D (0.675) requires something more than prior knowledge of the whole number system, since otherwise choice C would have been more likely. Thus choice D reflects a more developed stage than choice C. I thus position choice C into the stage of EMERGENT and choice D into the ACTIVE stage.

Question 5: Decimal Representation

Which is the decimal representation of the number "two hundred and six and nine tenths"?

A. 206.90
B. 206.910
C. 206.09
D. 206+9/10
E. 206.9
F. 2006 9/10
G. 20069.10

Decimal numbers can be represented in words or by using decimal notation. Question 5 explored the link between these two representations by asking learners to select which decimal number corresponds to a particular word statement. This test item was adapted from item L09 in the TIMSS 1999 released item set (Mullis et al., 2000; NCES, 2015), and two of the four choices in the TIMSS item have been retained as my choices E and C, which are complemented with six other options that I have added which may have the potential to identify other misconceptions.

Results

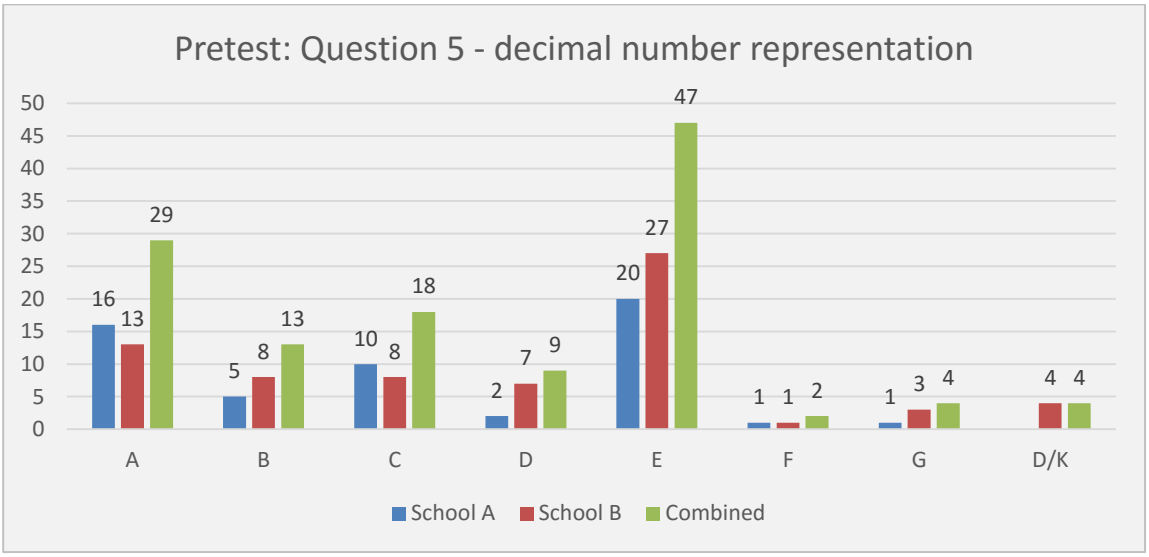


Figure 20. Pretest Question 5 - decimal number representation

Figure 20 shows that whereas the generally-accepted correct choice is E (206.9), which accounts for the highest number of learner responses, choice A (206.90) is also correct, but it is an alternative response which indicates greater precision, and together these two comprise 76/126 responses. However, choices B (206.910) and C (206.09) each had a relatively large number of responses when compared to choices F and G. Most learners indicated that choices B and C were either Just Right or Easy as shown in Table 13.

Table 13. Pretest Question 5 – difficulty levels

Option	N/S	Difficult	Just Right	Easy	TOTAL
A	1	1	8	19	29
B		3	**2**	**8**	13
C		4	**11**	**3**	18
D	1		4	4	9
D/K	2	2			4
E	1	5	13	28	47
F		1	1		2
G		1	3		4
TOTAL	5	17	42	62	126

Analysis

This test item appears to have some diagnostic value given the systematic selection of choices B and C. The results show that when the correct response was not

selected, the incorrect responses were biased towards choices B and C, which supports an inquiry into whether these were based upon common ways of thinking among this learner sample.

The correct responses (choices A and E) are an indication of the STABLE development stage of the learner, and choices B and C can position learners in the ACTIVE stage on the basis of presumed intermediate conceptions which account for these responses. For example, choice B may have been selected on the basis of learners believing that they should put both the 9 and the 10 (from the "nine tenths") into the selected answer, and for which the response of "0.9" may not appear as the correct representation to a learner since it does not explicitly identify the word "tenths" by a "10".

Question 6: Common Fractions to Percentages

What percentage is equivalent to the fraction $\frac{3}{4}$?

A. 3%
B. 4%
C. 34%
D. 50%
E. 75%
F. 100%
G. 133.33%

This question had seven choices, which included a number of rich distractors. The whole numbers for the numerator and denominator were both included as rich distractors, such as including 3% as choice A where 3 is the numerator of the fraction. The numerator and denominator digits combined together as choice C (34%).

Results

There is an indication, from the high frequency of choice E in Figure 21, that many learners understand the conversion from common fractions to percentages. The only significant alternative is choice C (34%).

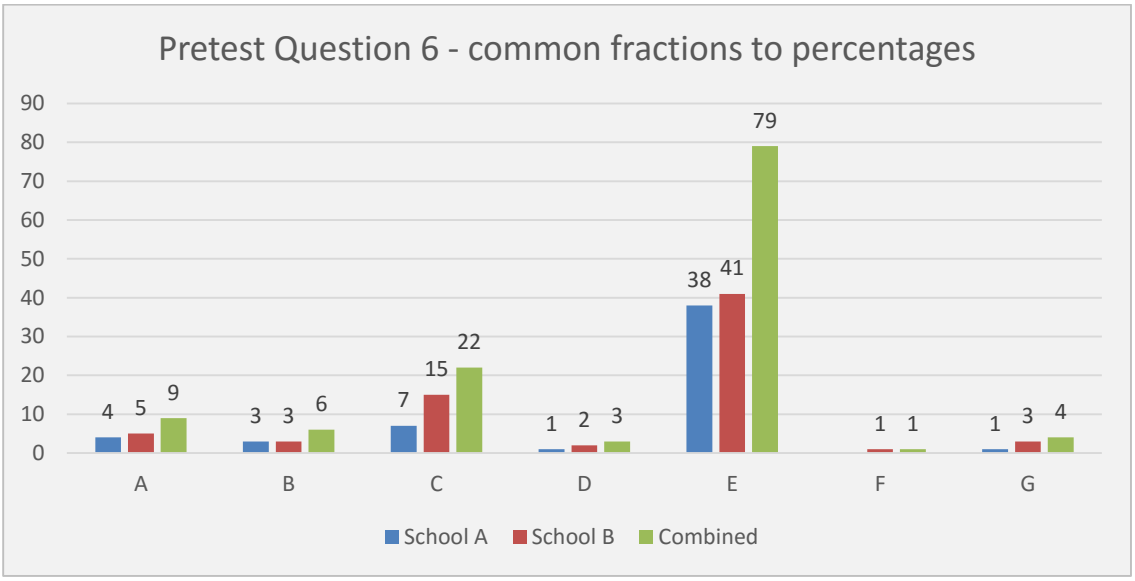


Figure 21. Pretest Question 6 - common fractions to percentages

Analysis

There was insufficient evidence from this example to conclude why individual learners selected rich distractors over the correct choice, and why the frequency of choice C was so much higher than choices A and B, when each of these appeared to support the usage of a single type of misconception.

When these results are explored in terms of the development stages, the correct responses indicate the STABLE stage, and the rich distractors of A, B, and C indicate the ACTIVE stage. A learner could be placed in the ABSENT stage when there is no apparent conception which is evident in the response, such as for choices D, F, and G. However, individual learners may have had other misconceptions which account for these choices where these misconceptions are not as yet known.

Question 7: Equivalent Fractions

What is the value of x if $\frac{3}{4} = \frac{x}{16}$

A. 3
B. 4
C. 7
D. 12
E. 13
F. 15

This question explored the operations and calculations required to make two fractions equivalent to one another. This was presented as a simple equation in which two

common fractions were presented, one of which had a missing value, represented by the variable x, which must be selected from the range of choices provided.

The distractors for this question included a range of plausible choices, including choice F (15), which was created by producing the same additive difference from 3 that 4 is from 16 in the denominator (+12). An alternative and simpler additive reading is that 3 is one less than 4 and thus the number we are looking for is one less than 16, being 15.

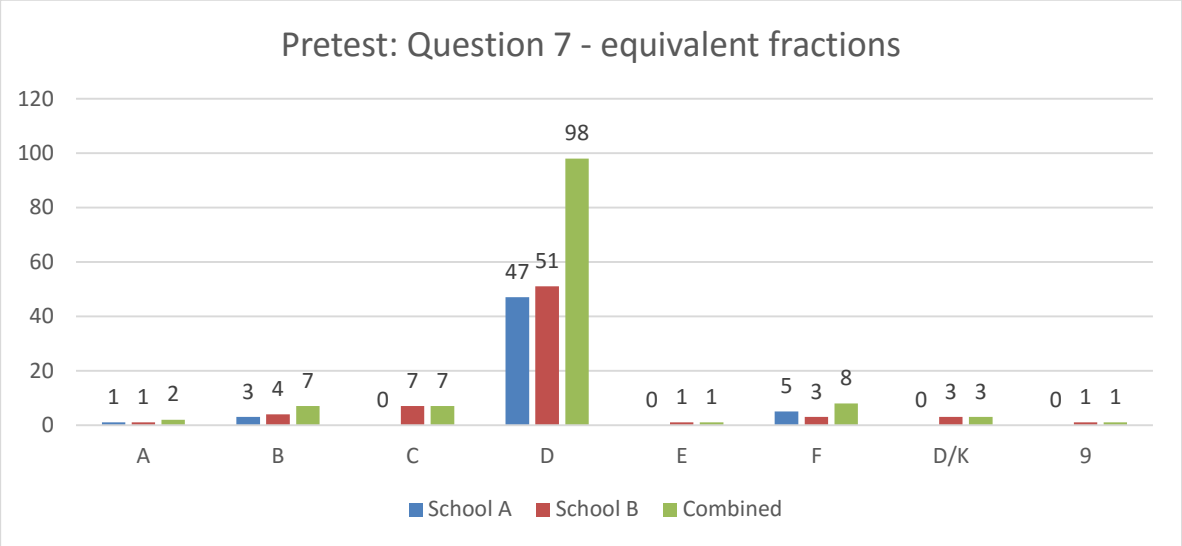


Figure 22. Pretest Question 7 – equivalent fractions

The results in Figure 22 demonstrated significant proficiency of the majority of the learners in both of the schools. One of the learners in school B entered an alternative response "9" which would result from noting that 4x4 = 16, and thus 3x3 = 9. This is an alternative relationship which I had not considered when designing this question, and this may have been more widely selected if this was one of the rich distractors.

Analysis

This question established that this was a proficiency of learners in both schools, with insufficient evidence of systematic errors on the basis of the choices provided. This question addressed procedural competence, and was more arithmetically challenging than the other questions presented up to this point in the pretest.

The results from this question showed that more learners had been assessed as competent on this question than on the apparently simpler questions in this pretest. This points to the possibility that whereas the questions requiring a number of steps are useful

to measure learner proficiency, it is the simple, and often trivial, questions that appear to be better suited for diagnostic purposes.

This question also highlighted a potential problem in the construction of diagnostic items, when a particular way of thinking was not reflected in one of the available choices, as was the case for the learner who wrote "9" as his/her response rather than selecting one of the given responses. This question thus violates the "Exhaustive Rule Set Usage" of Bart et al. (1994) which requires that every way of thinking is reflected in at least one choice. However, this also challenges our understanding of learner thinking, in terms of whether we have discovered all the various ways of thinking that learners employ, which are needed to build good diagnostic questions.

Question 8: Decimal Addition

What is the result of .25 + .4

A. .29
B. .65
C. 4.25
D. 25.4
E. 6.5

This question challenges the understanding of the decimal numbers in terms of how the .4 is treated. Every choice in this question has been designed as a rich distractor with the correct choice being B.

Choice A (.29) results from treating the .25 and .4 as whole numbers, by dropping the decimal point, and then adding them to get .29, and finally inserting the decimal point back into the calculation. Choices C (4.25) and D (25.4) combine the decimal numbers to create the choice, and choice E (6.5) considers the learner multiplying the correct response by 10.

For both schools, the distractor choice A was by far the most widely selected, even though the correct answer is choice B, which is the second-smallest choice, with only choice E having a smaller response. This type of item has potential as a good diagnostic question, since an important misconception is exposed in learner understanding.

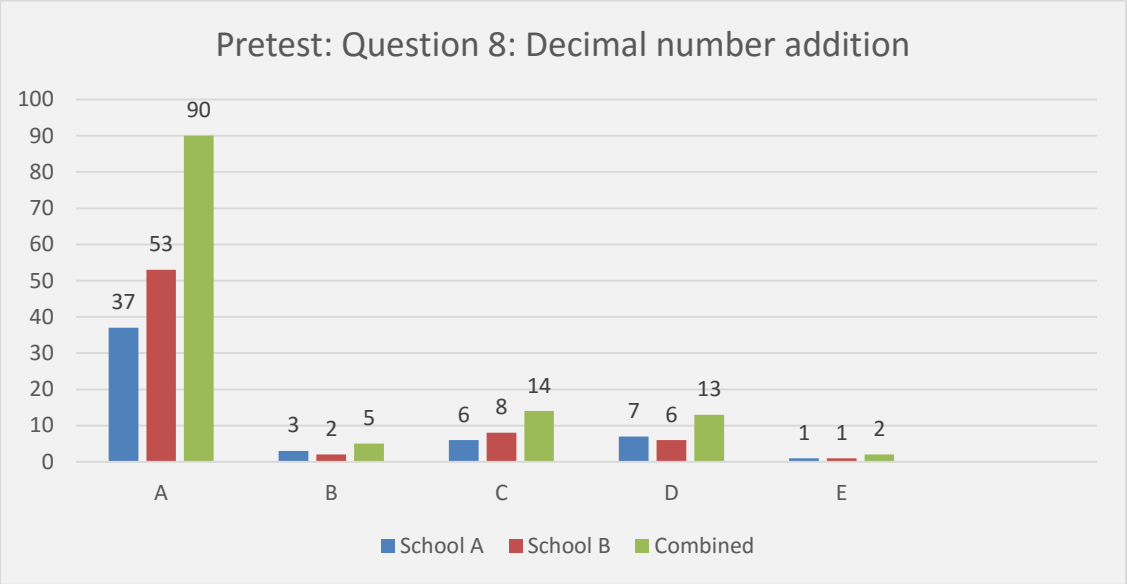


Figure 23. Pretest Question 8 - results

Analysis

The most likely explanation is that choice A was causing the learners to consider these two input numbers as whole numbers, so that 25+4 = 29. Choices C and D also had relatively high responses when compared to choices B and E, and show a way of thinking as observed earlier in which the learners may have believed that they had to include all of the input digits into the answer.

Another consideration is that these decimal numbers were presented without a leading zero, and this may have had an impact on the learners' choices. In the real world we are presented with numbers in many different forms and notations, and the dropping of the initial zero in the whole number part remains a valid decimal number, and examples of the usage of decimal numbers which drop the leading zero are presented by Kilpatrick et al. (2001).

The development stage is ACTIVE for those who selected the distractor choice A which showed a high-level of consistency between the two schools. The correct choices, answered by very few learners, indicates the STABLE stage, with the remainder likely to indicate an ABSENT stage of development. I would consider that choices C and D may point to the EMERGENT stage in which there is some evidence of conceptions, but insufficient to indicate conceptual development.

One advantage of this question for diagnostic purposes is its simplicity, and I continue to argue that good diagnostic questions should be as simple as possible while providing evidence of specific ways of thinking.

Question 9: Decimal Subtraction

What is the result of 7 - .4

This question was a variation of Question 8, focused on subtraction rather than addition, and was presented in a constructed-response format rather than as an MCQ. As for Question 8, the decimal point was used without the zero (0) prefix, and this may have caused confusion with the learners over and above their existing misconceptions.

Table 14. Pretest Question 9 – decimal subtraction

Response	Frequency
	6
.3	16
.9	1
+3	1
0.3	2
11 OR 3	1
3	75
3,0	1
3.	1
3.0	1
3.4	1
4	2
4.	1
4.7	1
5	1
6.6	13
7.3	1
7.4	**1**
D/K	1
DK	1
E	4

The results of the pretest, given in Table 14, showed that there were three significant responses being ".3" (16/132), "3" (75/132), and the correct answer "6.6" (13/132).

By far the largest frequency (75) was the response of "3", which may be explained by the learners ignoring the decimal point. The response of ".3" received 16 responses, in which the numbers were both treated as whole numbers, but in which the decimal point was added back in to the final response. The correct answer "6.6" only received 13 responses.

The difficulty factors for these three highest responses are shown in Table 15 and indicate that 60/75 of the learners considered this question as Easy.

Table 15. Pretest Question 9 – selected difficulty levels

RESPONSE	N/S	Difficult	Just Right	Easy	TOTAL
.3	2	1	2	11	16
3	1	3	11	**60**	75
6.6		2	5	6	13
TOTAL	3	6	18	77	104

Analysis

There were a large number of learners who responded with the incorrect answer "3" when compared to the other alternatives, which may be accounted for by the learners not considering that .4 is a decimal number, and rather treating this as a whole number.

My conclusion is that the subtractive form of the question is unlikely to reveal any further misconceptions within these results. However, the additive form does reveal some evidence of misconceptions that are suited to the identification of the development stage of the learner.

Question 10: Decimal Point Insertion

> The following arithmetic calculations have been worked out, but the decimal place is missing in the answer. Please insert the decimal point correctly in the following two answers:
>
> a. 657 x .7 = 46004
> b. 16.2 ÷ 3 = 54

I made an error in the wording for this question, since it should have indicated "the decimal *point* is missing", rather than "decimal place".

This question was representative of a special class of diagnostic questions that pose a problem together with an answer which is incomplete, where the learner is asked

to fill in the missing element. In this case the answer was missing its decimal point to mark the separation of the whole number part and the decimal fractional part of the decimal number.

After the paper was handed out to the learners, I noted an error, in that the answer presented for part A was incorrect, and it should have been presented as 4599. I thus removed Question 10a from further analysis due to this error and I only analyzed Question 10b.

Results

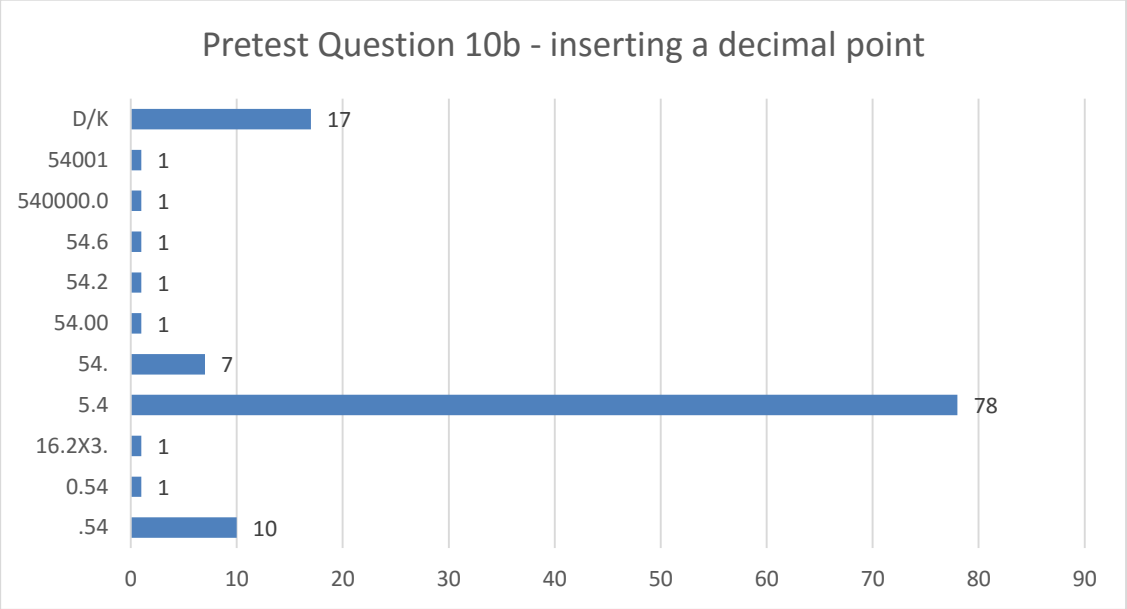


Figure 24. Pretest Question 10b – inserting a decimal point

As seen in Figure 24, for the second of the two questions there were 78 learners who selected the correct response "5.4" with less than 10% responding with either of the alternatives "54." (7) and ".54" (10).

Analysis

Question 10b was relatively easy, since dividing a decimal number by a whole number is well within the capabilities of the learners.

This type of question is not well-studied, and offers little additional support for diagnostic purposes for my needs in this study. I thus did not analyze this in terms of the Development Stages.

Question 11: Decimal Addition Estimation

Which of the following is closest to the sum 6.91 + 4.05

A. 6.00 + 4.00
B. 6.00 + 5.00
C. 7.00 + 4.00
D. 7.00 + 5.00
E. 8.00 + 3.00

In this question, the sum of the two decimal numbers in the item stem is 10.96, and some of the learners performed this calculation in full on their answer sheets. Noting that 10.96 is close to 11, this means that choices B, C, and E are all close, and this question was thus intended to discover what learners considered the term "closest" to mean. The answer I was seeking is choice C in which both the value as well as the form of the answer are close to the question.

This question was adapted from item H09 in the released items set from the TIMSS 1999 study (Mullis et al., 2000; NCES, 2015). The TIMSS item H09 used 3-digit whole numbers, and I modified this to use 3-digit decimal numbers, with a single-digit whole number, and with two decimal fraction digits for the tenths and hundredths positions. Of the 38 countries participating in TIMSS 1999, Singapore was highest with a 97% pass rate, the international average was 80%, and South Africa was in the lowest position with a 38% pass rate. This is essentially identical to the 39% (49/125) results from my study.

Results

The results provided in Figure 25 showed a peak for the correct choice C for both schools combined, while School B had choice A as the highest frequency. Choices B and D had high values, and whereas choice B adds up to the estimated value, choices A and D do not. In this case in there was a significant difference between the Grade 8 class in School B, and the Grade 7 class in School A.

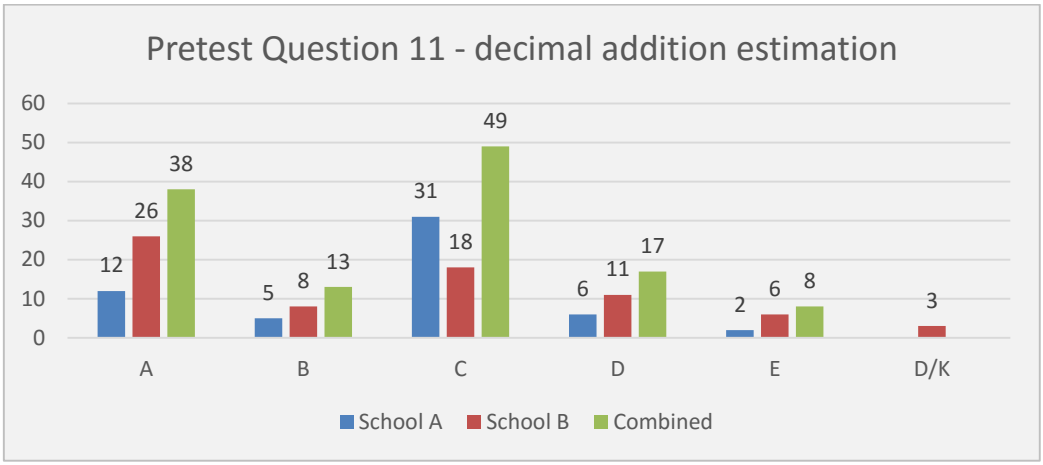


Figure 25. Pretest Question 11 - decimal addition estimation

Analysis

There were multiple "correct" solutions for this question, when this is determined by the magnitude of the answer alone, and this is likely the reason why the responses were distributed over the choices available. Choice A was likely selected because the whole number part of the decimal numbers is 6+4 which is the same as integer parts of the numbers in the question text, even though the sum of 10 is not close to 10.96 as for the other choices.

Without further analysis of the reasons for the learners' responses, it was not possible to position these responses onto the development stages in my model, and thus this particular type of question appears to have limited value for my online diagnostic tests.

Question 12: Common Fraction Equivalence

Write down a fraction that is equivalent to $\frac{3}{8}$

Three related questions were included in the pretest as constructed-response questions, to enable me to examine the patterns of responses. Question 12 explored equivalent fractions, asking the learner to write a fraction equivalent to the fraction provided. I first addressed Question 12 and then I combined the analyses for Question 13 and Question 14 which both address the density property of the common fractions—which is that between any two common fractions it is possible to find another common fraction—and this attribute of the common fractions makes them different from the whole numbers.

Results

Table 16. Pretest Question 12 – common fraction equivalence

Response	Frequency Count
	7
/4	1
0,38	1
1/2	2
1/4	5
1/7	1
12/16	1
12/32	1
2/2	2
2/4	1
2/5	1
2/7	1
22/8	1
24	2
24/64	2
24/8	1
26	1
3/16	1
3/4	2
3/5	1
30/80	1
375/1000	1
38%	1
4/12	1
4/16	1
4/4	3
4/9	1
5/10	1
5/12	1
6/16	53
6/8	1
8/3	3
9/16	5
9/24	8
9/64	1
D/K	16

In Table 16 the responses with a green background indicate the correct responses provided by the learners.

The vast majority of the learners responded with "6/16", created by simply doubling the numerator and the denominator, although a range of other correct answers were given. The most frequent incorrect responses were "1/4" and "9/16" each with only five responses, and with a large number of other responses for which there are only a few responses.

Analysis

From these results I conclude that the constructed-response format is suited to discovery of new ways of thinking, but is not suited to diagnostic assessment, since there are too many possibilities for which there is no apparent conceptual basis. This type of item may be used to identify proficiency when a correct response is provided, but the incorrect responses are generally not identified with any specific misconception. Given the challenges in the lack of common patterns in the frequency of incorrect response, I have not attempted to position the learners into the Development Stages.

Questions 13/14: Common Fraction Density

Question 13 : Write down a fraction that is larger than $\frac{2}{7}$ and less than 1
Question 14 : Write down a fraction that is larger than $\frac{3}{4}$ and less than 1

Question 13 asked the learner to find a new fraction which is between a given fraction and 1, where there are additional numerators available using the same denominator. Question 14 then repeated the form of Question 13 but for which there are no other numerators available which satisfy the requirement and which thus required a change to the denominator for the solution.

Table 17. Pretest Question 13 – common fraction density

Response	Frequency count
0	2
1	1
1 9/7	1
1 AND 6	1
1/1	1
1/14	1
1/2	20

Response	Frequency count
1/4	2
1/4 OR 1/2	1
1/6	1
1/7	5
15	1
2/4	1
2/5	2
2/8	3
2/9	1
3/3	1
3/4	5
3/6	2
3/7	13
-3/7	1
3/8	7
3/9	1
34	1
4/14 (Note 1)	8
4/6	1
4/7	5
5/13	1
5/14	1
5/7	4
-5/7	1
5/8	1
6/14	1
6/7	10
7.0/10	1
7/8	1
D/K	14

Note 1: The “4/14” is equivalent to “2/7” but is not larger, as required.

Table 18. Pretest Question 14 – common fraction density

Response	Frequency count
0.5	1
0.5/6	1
1	1
1/1	1
1-/12	1
1/2	10
1/3	4

Response	Frequency count
1/4	7
1/5	1
1/7	1
1/8	1
1/9	1
12	1
12/12	1
12/16	2
15/16	1
2/3	4
2/4	4
2/4 1/2	1
2/5	1
2/7	1
2/8	1
3	1
3/5	1
3/8	1
3/9	1
4 AND 5	1
4/4	4
4/5	9
46	1
5/10	1
5/5	1
5/6	4
5/8	1
6.4	1
6/4	1
6/7	1
6/8 (Note 1)	9
6/9	1
7/8	20
7/9 (Note 2)	1
8/10	2
9/10	2
9/16	1
9/8	1
D/K	13

Note 1: The response "6/8" has the same value as "3/4" and is not thus larger. This is the same error as noted in Question 13.

Note 2: "7/9" is an interesting response, since it is quite difficult to determine that "7/9" is greater than "3/4". If these are each converted to the denominator 36, being the lowest common multiple of 4 and 9, then "3/4" becomes "27/36" and "7/9" becomes "28/36", which is larger, but only just. This was either a well-informed response or a good guess.

The most common correct response is "7/8" which shows a good understanding of the density of the common fractions. Since there is no further numerator beyond 3 with a denominator of 4 while still remaining below 1, the learners simply doubled the denominator to "6/8" and then added 1 to the numerator, giving "7/8". The other responses were used rarely by only 1-2 learners, and yet some of these represent common misconceptions that can be grouped for analysis purposes.

Analysis

These questions illustrated a range of issues in learner understanding of the common fractions. For Questions 13 and 14, many of the learners entered values that were either equal to the lower fraction or were larger than 1.

Whereas these items have potential for the discovery of misconceptions, their value lies in the constructed-response format and did not suit the MCQ format that I was using for the online tests. The only viable diagnostic response was when the learners were suggesting a value that was equivalent to the lower value given, on the understanding that this is larger. For example, seeing "6/10" as larger than "4/5".

Question 15: Ordering of Decimal Numbers (10 pairs)

This question was a simplified variation of the decimal ordering item type, which has been the subject of extensive research in terms of learner misconceptions (Sackur-Grisvard & Leonard, 1985; Resnick et al, 1989; Steinle, 2004a; Steinle & Stacey, 2005).

When learners are asked to order a set of decimal numbers into increasing or decreasing magnitudes, or to select the largest or smallest value, this has been shown from the previously-cited studies to expose a range of behaviours, including those studied and verified over large-scale, longitudinal studies (Steinle, 2004b).

This small set of ten pairs of decimal numbers has been designed to explore the extent of the misconceptions both at the level of the class as a whole and at the level of the individuals.

Draw a circle around the SMALLEST of EACH of the following pairs of decimal numbers

A.	0.45	0.39
B.	0.4	0.39
C.	0.45	0.3
D.	5.45	5.39
E.	0.453	0.3
F.	0.398	0.3
G.	8.4	8.3
H.	7.45	7.9
I.	3.45	3.33
J.	0.45	0.08

Results

The results are presented in the sequence of the percentage of learners who obtained the correct answer. This does not imply that the higher percentages are easier, and this also does not imply directly that any of these are good indicators of misconceptions. For this question I use the expression 15(B) to indicate the sub-question 15(B).

Table 19. Pretest Question 15 – decimal ordering results

Sub-Question	Count	Number Correct	% Correct
I	87	64	73.6%
G	93	67	72.0%
C	108	77	71.3%
D	94	66	70.2%
A	98	66	67.3%
F	106	71	67.0%
E	105	67	63.8%
J	100	58	58.0%
H	88	40	45.5%
B	107	42	39.3%

The lowest score was from 15(B) that asked the learners to compare 0.4 and 0.39, and which may expose the WHOLE misconception.

The highest scores were from 15(I) and 15(G), which both have the same number of decimal digits in the fractional part.

Analysis

There is value in these types of test items, and they are easy to administer in a MCQ format, and thus were selected for use within the online tests. However, there is a limitation that a learner response on items with only two choices may result in various ways of thinking, and it may not be possible to determine the true conceptual cause without a sufficient number of items which is contrary to my assumed need for efficiency when used in a classroom environment.

The analysis of these ten items over the complete set of learners shows 81 learners who answered all 10 sub-questions, and I analyze the responses from this subset using the total number of the correct responses as an indicator of proficiency. Of these 81 learners, 9 scored 10/10 and another 6 scored 9/10. The learners scoring 9/10 made mistakes on 15(B) and 15(H) only, suggesting that these may represent items which can elicit late-stage misconceptions in otherwise proficient learners. There are 28 learners who scored 8/10 and of the 58 mistakes made by this group, 28 were on 15(B), with another 25 on 15(H), with the remaining 5 mistakes from 15(D) - 2, 15(E) - 1, 15(F) - 1 and 15(J) – 1. I use the notation "15(D) – 2" to indicate that learners made 2 mistakes on 15(D).

Based upon this analysis I position those learners with 10/10 into the STABLE development stage, and those scoring 9/10 and 8/10 into the IMMINENT stage.

Learners scoring 7/10 made the largest number of mistakes on 15(B) – 8, 15(H) – 7, and 15(J) – 8, with no other questions receiving more than three incorrect responses. So whereas the item 15(J) did not feature as a mistake with those learners scoring 8/10 or more, this was predominant in learners scoring 7/10. Thus it appears that the learners scoring 7/10 may have been confused by the leading zero in item 15(J) which did not cause the higher-scoring learners to make the same mistakes.

When analyzing the 7 learners who obtained a score of 6/10, which consists of 28 total mistakes, the distribution of the mistakes changes, with the largest being 15(B) – 5, 15(E) – 5, and 15(J) – 6. I note that 15(H) is answered incorrectly by only 2/7 of these

learners. I also note that there were only two mistakes for 15(F) even though 15(E) and 15(F) are similar in their structure and in their correct choices.

I position those learners scoring 6/10 and 7/10 in the ACTIVE development stage, showing a range of systemic errors.

There were no learners who scored 5/10, and only three learners each scoring 4/10 and 3/10. However, for these learners, it is difficult to distinguish those learners with some understanding from those who are guessing. Of these six learners, there was little consistency in the successful questions, with the major results as 15(B) – 5, 15(G) – 4, 15(H) – 3, and 15(I) – 3 with the other sub-questions having 2 or less correct responses. Whereas these were distributed quite evenly, when compared to learners scoring 6/10 up to 10/10, it is noted that the highest success was achieved on 15(B) which was one of the items in which the IMMINENT stage learners struggled with.

There were two learners who scored 0/10 and one who scored 1/10, but a surprisingly high 12 learners with a score of 2/10. This latter group of learners also achieved success on 15(B) and 15(H) consistently, as the only two items on which they succeeded, which are the same items which those with 8/10 and 9/10 grappled with, and this presents a contradiction, which is emphasized by the results of those learners who obtained 3/10 and 4/10. It is likely that these low-performing learners were using an alternate way of thinking which caused them to fail on all other questions and to only succeed on these two questions.

Using these scores alone it appears that those learners scoring 0/10, 1/10 and 2/10 were using some conceptual knowledge to account for their responses and thus could be in the EMERGENT stage. However, those learners scoring 3/10 and 4/10, for which there is insufficient evidence to determine consistency in the responses, may be in the ABSENT stage, in which their responses may be indistinguishable from guessing.

What is apparent here is that learners who have a very low score, or 0/10, 1/10 and 2/10, which results are unlikely to result from guessing because of their high frequency of a few incorrect responses, may have better conceptual development than learners who scored 3/10 and 4/10. This is a reflection of the observation I have made earlier concerning the TIMSS results and the issues of making inferences on learner abilities from unusually low scores.

The other anomaly is the relatively large number of learners who achieved a score 2/10 when compared to the totality of learners who scored less than 5/10.

The most unexpected response is the inverse correlation of 15(B) and 15(H) between learners achieving a high total score but who then failed on these sub-questions, with learners achieving a low total score but who succeeded on these same sub-questions. Both of these sub-questions have a common structure which is unique to these sub-questions and which is not present in any of the others, in which one choice has a single decimal place and the other has two decimal places, where the decimal number with the single decimal place is the largest. However, I cannot determine why this inverse correlation is so distinctive, and I suggest that this warrants future study.

5.2 Summary of the Pretest Results

I have used the results of the pretest to help determine the suitability of various test items for online assessments, and to identify those that may help to position learners into the development stages.

I conclude that not every type of test item is suited for diagnostic purposes, and the pretest has helped to identify item types that are more likely to be useful in a diagnostic context, where other types of items appeared to be of little diagnostic value. I now summarize these questions and my decisions on inclusion into the online assessments.

Question 1 concerned Common Fraction Ordering and it provided useful results on the inability of some learners to compare two small and simple common fractions. This single question alone demonstrates the power of a simple question to separate learners with different ways of thinking about a competence which is expected to be in place by Grades 7 and 8. I thus selected this type of question for the online assessment with the intention to compare the diagnostic value of different items.

Question 2 concerned Common Fraction Estimation, using an example from prior studies. This question is relatively complex, and I have created a simpler variation of this for the online assessments using a single common fraction rather than an addition. It seems that simpler questions may provide better diagnostic evidence than questions where learners are required to perform more mathematical operations.

Question 3 on Decimal Fraction Place-Value provided an opportunity to elicit misconceptions arising from the form and notations of the decimal system, which may be

overlooked by more complex items such as decimal number ordering. These questions offer the alternative to diagnose a lower-level skill, being the understanding of place-value notation, prior to the use of such numbers in operations such as ordering. This question type was included in the online assessment structures.

Question 4 on the Ordering of 5 Decimal Numbers had the potential to elicit a number of misconceptions in the decimal numbers. My results from the pretest mirrored the results from other studies, and provided a range of possible responses, where each response may point to a different way of thinking. For the online tests I have included decimal number ordering test items which use two decimal numbers, as in Question 15, which addresses the same misconceptions.

Question 5 on Decimal Representation explored the fundamentals of decimal notation and was similar to Question 3. However, Question 3 showed greater promise for the identification of misconceptions and was also simpler. Thus Question 5 was not used for the online assessments.

Question 6 had the potential to elicit misconceptions, but was not used since this introduced the percentages as a new form of rational number representation. It is my preference to limit the online assessments to decimal numbers and to common fractions due to the short time that was available for these online lessons.

Question 7 required arithmetic operations to produce the answer, and was thus focused on operational rather than conceptual knowledge. Whereas this question was more complex than others given the amount of procedural work required to arrive at an answer, this question showed a high proficiency among the learners, with little evidence of systematic misconceptions proportional to the total sample. This may result from schools focusing on procedural knowledge rather than conceptual understanding. Thus for this group of learners this type of question may not be useful.

The experience from analyzing Questions 8, 9 and 10 was such that I consider these too difficult to interpret for diagnostic purposes although they may have value in further studies.

The form of Question 11 has also been used previously in TIMSS, but I am uncertain about the diagnostic value of this type of question, and I have elected not to use this for the online assessments.

Questions 12, 13, and 14 all concerned actively creating new common fractions which meet a specific requirement. Given that these required constructed-responses, these were not included in the online assessments, since I was using MCQ questions exclusively due to their ease of handling and analysis. I also questioned whether the results of these types of test items have diagnostic value, given the complexity in identifying the conceptual base on which the answers were provided in this pretest.

I have included other types of items in the online tests which were not presented in this pretest, including problems which use the number line, geometric areas, and sets.

I now continue with the analysis of the data arising from the online assessments.

CHAPTER 6 : DATA ANALYSIS - ONLINE DATA

6.1 Introduction

This chapter provides the analysis of data collected during four online assessment lessons as conducted with four classes over two study schools (School A and School B). This data analysis uses the approach described in Chapter 4.

I apply my research questions to the data obtained from the online assessments for each of the micro-domains of my study, and in Chapter 7 I combine the results from the individual micro-domains.

The data was captured directly from the learners' interactions using an online, web-based assessment program. This online assessment program gathers data far more efficiently than is possible with manual data capture from learner-completed paper answer sheets, and also collects far more data, at a faster rate, and in real-time if necessary.

Throughout this chapter I make reference to the test items from my item bank, which is presented in full in Appendix D. These test items are identified as "Item NNNNN", where NNNNN is the item number from the Item Bank.

This chapter is arranged as follows:

- An outline of the data obtained from the online tests.
- A detailed analysis of the data, including answering the research questions, for each micro-domain in turn.

6.2 The Data

Data was captured automatically from the online assessment system as part of the learners' interaction with the system. This data was captured into databases on the web server and was merged into a Microsoft Access database for analysis and reporting.

This merged database contains one record for each test item answered by each learner, and each record includes data fields for the learners' responses as well as the timing of their responses. These data records include items which the learners attempted but were unable to answer. However, no data records were stored for test items that were

not presented to the learners due to their time running out in the lesson, or for cases where the learner did not attend on the day of the test.

Each learner took the same set of tests during each lesson, where each test comprised the same sequence of test items, and the learners completed the tests at their own pace. Some learners finished the tests quickly while others did not complete the tests in the allotted time.

The learners responded to each item in turn, and the system checked that their response was acceptable before they could advance to the next item. An acceptable response was either: (1) a selection of one of the multiple choices available in the MCQ, or; (2) an indication that the question was not understood and thus that a selection from the available choices could not be made. The learners indicated whether they found the test item Easy, Just Right, or Difficult, and whereas this was initially included as an optional question it was modified to be mandatory from the second lesson in School A and for all subsequent lessons for School A and School B.

The data fields captured in the combined database are detailed in Appendix A.

Handling Errors in the Test Items

Some errors were noted in the formulation and encoding of the test items, and one of the database fields, named "ERRORNOTE", stores information about such errors. For example, consider Item 10084 as shown in Figure 26. The question stem was incorrectly worded and this was not noticed until after the learners had completed the tests.

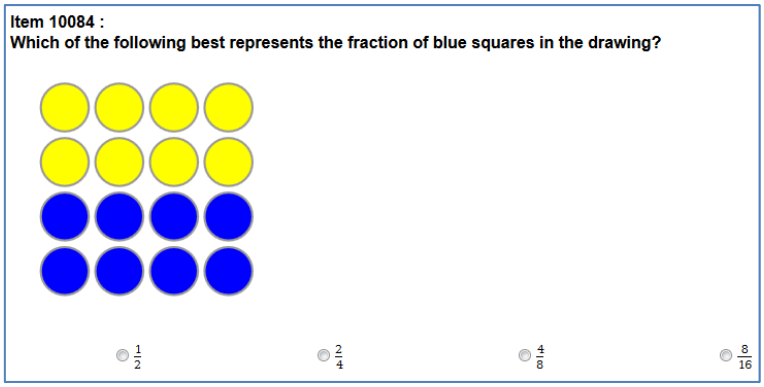


Figure 26. Item 10084

This question asked the learners to select the blue *squares* when the diagram clearly has blue *circles*. This was one of a very small number of cases in which the data analysis must consider the possibility of a misunderstanding by the learners arising from

the incorrect form of the test item. For these cases, the data was checked to ensure that these errors in wording did not impact the learner responses and the ERRORNOTE data entry help to explain misfitting items.

Micro-Domains and Response Counts

The frequency of learner responses by micro-domain is presented in Table 20. The Record Count column is the number of response records on the database for each micro-domain in this study. Some items were used more than once, over different tests, and these duplicates were removed, using only the first response by each learner to each question, to mitigate the case where learners changed their response after seeing the correct result of the duplicated item in the feedback. This is then my data universe.

Table 20. Count of response records by micro-domain

Code	Micro-Domain	Record Count
PV	place-value problems	2599
DO	decimal number ordering	2593
CR	common fraction representation	899
NL	number lines and common fractions	835
CG	common fractions in graphics	936
CO	common fraction ordering	2354
CE	common fraction estimation of values	1252
CA	common fraction add/subtract	1119

The Research Questions and the Data

I first outline how the data supports my research questions, and I repeat the definitions for these research questions as were introduced in Chapter 1. These questions are then addressed in detail for each of the micro-domains, using the Development Stage model which was described on page 85 in Chapter 3.

RQ1 (EFFECTIVENESS): How can we measure test items in terms of their fitness-for-purpose as good diagnostic instruments?

The learner responses are analyzed for each item using the Rasch method, to rank these items by their measured value for detecting misconceptions, with some items then identified as the "best" for this purpose. Selected items are also analyzed qualitatively to

explore whether their diagnostic value can be predicted from the nature and form of the item.

RQ2 (EFFICIENCY): Given a particular diagnostic context, how many good diagnostic questions are sufficient to establish valid and reliable evidence?

RQ2 explores how many items are required to justify a claim that a learner has used a particular misconception. Whereas prior studies have used larger test batteries to improve the reliability of the results, these are not suited for practical use given the time restrictions of classroom settings and I argue for the importance of knowing how many items are sufficient where the ideal is a single test item with the diagnostic power to discriminate between learners who use a particular misconception and those who do not.

This question is answered by considering the "best" test items as identified in RQ1, and by exploring how many such items are needed. This is answered on the basis of the limited items available within this study, with the goal to discover general principles that can support future, more detailed studies.

RQ3 (SELF-KNOWLEDGE): Does access to learner self-knowledge aid the process of diagnosis, in terms of the additional benefit for the added effort in obtaining this information?

Finally, RQ3 explores to what extent learner self-knowledge may help to discover misconceptions, as a qualitative input which is distinct from the quantitative inputs from the learners' responses as analyzed for RQ1 and RQ2. I address RQ3 using the decision matrix presented in Table 10 on page 125, given the learner inputs from the test items and the learners' indication of item difficulty.

6.3 Micro-Domains Summary

For the benefit of the reader, I summarize each of the eight micro-domains used in this study as were introduced in Chapter 4, prior to embarking on the detailed analysis:

- PV - Decimal Place-Value: the knowledge of the place-value system as applied to decimal numbers, and specifically to decimal fractions. Limitations in the knowledge of place-value will hamper conceptual understanding of other micro-domains within the decimal number system such as conversion between various forms of rational numbers.

- DO – Comparing and Ordering Decimal Numbers: the knowledge of the magnitude of decimal numbers and how the numbers are ordered into ascending or descending magnitudes.
- CR – Common Fraction Representation: the ability to match a word description of a fraction with its common fraction notation.
- NL - Number Line and Common Fractions: the conceptual understanding of the numerical magnitude of common fractions by being able to position a fraction on a number line.
- CG - Common Fraction Graphics: the ability to match graphic representations to common fractions.
- CO - Common Fraction Ordering: the ability to rank common fractions by selecting the highest or lowest from a set.
- CE - Common Fraction Estimation: the estimation of the magnitude of common fractions.
- CA - Common Fraction Addition: the addition of common fractions.

These eight micro-domains are concerned primarily with conceptual understanding of the rational numbers and their representations, with some attention to procedural knowledge. These micro-domains include well-researched misconceptions that commonly occur in the development of learners' mathematical proficiency.

I provide a detailed description of the analysis conducted on the first micro-domain of Place-Value, describing how the general approach to analysis, as described in Chapter 4, is applied to this micro-domain. For subsequent micro-domains I reduce the level of detail since much of this approach is repetitive.

Frequency analyses of the learner responses were conducted initially in each micro-domain, to identify whether these met the expected results. The frequency analysis is used to show the number of learners who have selected the correct choice in the MCQ items, compared to the number who selected choices which are based on known misconceptions. These analyses are used to discover evidence of systematic error patterns which I had not pre-identified and which warrant further consideration.

Within this chapter I use the word "lesson" to identify the online lessons conducted as part of this study.

6.4 Micro-Domain PV - Place-Value Knowledge

Initial Analysis of Responses

This micro-domain consists of 25 test items on place-value knowledge, being Items 10001-10020 and 10051-10055 from the Item Bank as presented in Appendix D. Of these, 16 are in the PV1 form and 9 are in the PV2 form, where these different forms are described in Chapter 4 on page 104.

A total of 105 learners provided responses to these items, including 31 learners from School A, for whom these test items were presented over two separate tests in the first lesson, and 74 learners from School B, for whom these items were presented in five tests conducted in the first two lessons. The number of learners attending the lessons differed for each day of testing, resulting in some inconsistency in the total numbers of learners who provided responses.

Table 21 shows the frequency of responses to these items, with the purpose of providing a quick visual inspection of the response patterns. Some patterns warrant explanation, and to support this the cells are colour-coded as follows:

- cells with a blue background are the correct choices, which consistently indicate the highest frequency responses
- cells with a red background are choices are accountable to the WHOLE misconception where the learner sees the decimal number as a single whole number. The choices in this category include those which use the "ths" suffix (such as "hundredths"), as well as choices which do not have this suffix (such as "hundreds") on the assumption that the learner may be confused by the words as well as by the notation
- cells with a yellow background are from choices which have five or more responses and yet are neither correct responses nor accounted for by the WHOLE misconception

In this table, the columns labels are the sequence number of the choices from the MCQs, and are not the actual values of these choices as selected by the learners. For the PV1 test items, where the learners had to select the digit at a given position, the choices are given as the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 as well as the additional choice labelled as "there is no digit in that position". There are 11 possible choices and each is numbered as 1-11 respectively. For the PV2 tests there are 9 choices, which the learner saw as

"thousandths", "hundreds", "tens", "units", "tenths", "hundredths", "thousandths", "ten-thousandths", and "the value cannot be determined", and these are numbered as 1-9.

Table 21. Place-Value: Counts by test item/response

Item#	1	2	3	4	5	6	7	8	9	10	11
10001		22	6	1	51	9					
10002		4	22	1	10	44	2		5		
10003	26	2			6		2	48	1		3
10004	2	18	1	1	3	16	40	4	2		
10005	1		20		9	6	48	3			
10006	1	8	3		8	56	1	23		1	2
10007	30	2		1	3	2	1		57	1	6
10008	5	2	3	2	66	2	3	20			
10009		67	3	2	4	2	22	2			1
10010		21	1	1	8	69	2		1		
10011	3	4		47	1	11	1				
10012	1	10	1		6	47	1				
10013		6	1		4	8	5	40	1		
10014		7			7	49	2				
10015	4	7	2	5	6		1	1	6		33
10016	5		2	11		3	47	1			
10017	1	9	6	1	46	5		1			
10018	1	10		2	1	46	2	3		1	3
10019	5	4	2	51	1	4	2				
10020		4	8	1	7	47	1	1			
10051	1	57	2	2	2	8					2
10052	2	2	5	1	3	13	46	2			
10053	2		7		55	7	2				
10054		7	54	4	2	4	1	1			
10055			2	5		55	3	5	1	2	

Two of the items do not have red cells. Item 10051 asked the learner to select the "hundredths" position in the decimal number 0.3154, for which the correct choice is the digit "1", but this is also the digit that would be selected by the learner if the entire decimal number was considered as a whole number, and also if the learner viewed "hundreds" and "hundredths" as the same. Thus the 57 responses for choice 2 were not distinguishable between proficient learners who used stable conceptions and learners who

used the WHOLE misconception. Item 10054 does not have a specific misconception linked to the choices, but there were 7 learners who selected “hundreds” when asked for the position of the digit “5” in the decimal number “451.678”.

The yellow cells were analyzed to seek an explanation of the patterns and whether they pointed to misconceptions other than the WHOLE misconception. The words used in the item stem and choices are shown in quotation marks, such as “tenths”:

- Dropping the suffix “ths”, such as selecting “tens”, when “tenths” is the correct choice (10001, 10017).
- Identification that the value cannot be determined (10002).
- Counting from the left, ignoring the leading zero, but still considering this as a decimal fraction. Such as considering the 4 in 0.0674 to be in the thousandths position (10003, 10005).
- Looking for a digit when there is none in the position, such as looking for the hundreds position in 0.080001 when “hundredths” was asked for (10007).
- Selection of a choice which is closest to the one needed, when it was not provided as a choice. This included selecting “thousands” for the place-value of 3 in 0.34759, where there was no “ten-thousands” given as an choice, but there was a “ten-thousandths” choice (10008).
- Dropping a leading zero, such as selecting “tenths” for the digit 9 in 117.0905 (10010, 10014).
- High-frequency errors requiring further analysis beyond the scope of this study are:
    - Selecting 4 for the tenths position in 214.579 (10006, 8 learners).
    - Selecting “tenths” for digit 5 in 0.4567 (10012, 6 learners).
    - Item 10015 had three high-frequency responses, which do not fall into the above explanations.
    - Selecting 5 as being in the hundredths position in 0.3154 (10051, 5 learners).
    - Selecting 9 in 241.97 as being in the hundredths position (10053, 7 learners).

- Selecting 5 in 451.678 as being in the hundreds position (10054, 7 learners).
- Selecting 3 or 7 as being in the thousandths position in 23.6759 (10055, 5 learners each).

These unexplained responses are potential misconceptions, due to the high frequency of these responses compared to other choices, but there is insufficient data to warrant a further consideration in this study. These anomalies point the way to a more fine-grained analysis of learner thinking, such as for Items 10015 and 10055, where the frequency of unexplained responses was larger than the choice indicating the WHOLE misconception.

Item Suitability

Proficient learners were isolated and removed from further analysis, on the basis that these learners would not make mistakes and thus their responses could not help to identify items which were better suited for detecting misconceptions. Mistakes made by proficient learners were likely to be slips.

To identify proficient learners, the items used had to be fit-for-purpose, and items with low correlations or which were misfitting would have had a negative impact on the inference process and had to be removed.

In Table 22, which was drawn from TABLE 26.1 in Winsteps (Linacre, 2013), the data is sorted on the column labelled PTMEASURE/CORR—being the "point-measure correlation". This column shows the level of correlation between the item responses and the learner measures, with a range of values between -1 and +1. Negative correlation indicates that learners with high measures tend to fail on this item, and learners with low measures tend to succeed. A similar argument can be made for low positive correlations, such as those below 0.20. Items with negative or low correlation were checked to identify the possible cause, and if these could not be fixed by correcting the coding of the items, then these items were removed as being unrepresentative of the construct being measured. Table 22 shows that all of the Place-Value items had correlation of between 0.48 and 0.78 which are within the acceptable range and thus all test items were retained for the next round of analysis, looking at the fit statistics.

Items 10052, 10053, and 10054 showed high OUTFIT mean squares (MNSQ) values of 2.19, 2.05, and 1.63 respectively. Values larger than 2.00 were attended to and

warranted their consideration for exclusion (Linacre, 2002; Linacre, 2013). The OUTFIT statistic is sensitive to outliers and if this value is high for an item it indicates that learners were making slips or were guessing. Whereas this is fixable, it has the potential to distort the Rasch model.

Table 22. Place-Value: Test item correlation

```
TABLE 26.1 PV - SCHOOLS A+B - CORRECT   PV-AB-CORRECT.out.txt  Nov 17 21:17 2013
INPUT: 105 LEARNER  25 TESTITEM  REPORTED: 105 LEARNER  25 TESTITEM  2 CATS WINSTEPS 3.80.1
-------------------------------------------------------------------------------------------
LEARNER: REAL SEP.: 2.07  REL.: .81 ... TESTITEM: REAL SEP.: 1.84  REL.: .77

         TESTITEM STATISTICS:  CORRELATION ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|         |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| TESTITEM|
|------------------------------------+----------+----------+-----------+-----------+---------|
|    24     54     73    -.80     .36|1.51   2.4|2.19   1.8|  .48   .64| 73.5  83.6| 10054   |
|    22     46     74     .19     .34|1.64   2.9|2.05   2.5|  .49   .69| 69.6  82.3| 10052   |
|    21     57     74   -1.14     .37|1.18   1.0|1.28    .6|  .57   .62| 78.3  83.9| 10051   |
|    14     49     65    -.93     .39|1.31   1.6| .95    .1|  .57   .64| 80.0  83.4| 10014   |
|    25     55     73    -.93     .36|1.20   1.1|1.17    .5|  .57   .64| 82.4  83.8| 10055   |
|    15     33     65    1.23     .35|1.37   1.7|1.50   1.5|  .58   .69| 71.7  81.9| 10015   |
|    12     47     66    -.55     .37|1.15    .8| .91    .0|  .63   .66| 77.0  82.9| 10012   |
|    23     55     73    -.93     .36| .89   -.5|1.63   1.1|  .65   .64| 88.2  83.8| 10053   |
|    10     69    101    -.42     .30|1.03    .2|1.24    .8|  .65   .67| 84.6  81.9| 10010   |
|     7     57    101     .55     .29|1.11    .7|1.23    .9|  .66   .70| 76.7  81.6| 10007   |
|    20     47     68    -.22     .37| .98    .0|1.12    .4|  .67   .68| 86.7  82.5| 10020   |
|    11     47     67    -.42     .37| .85   -.7|1.25    .7|  .70   .67| 88.7  83.0| 10011   |
|    18     46     69    -.09     .36|1.00    .1| .86   -.2|  .70   .69| 81.7  82.1| 10018   |
|     3     48     86     .88     .31| .94   -.3| .89   -.3|  .71   .69| 81.6  81.3| 10003   |
|     6     56    100     .63     .29| .86   -.9|1.02    .2|  .72   .70| 87.6  81.3| 10006   |
|     4     40     86    1.55     .30| .85  -1.0| .66  -1.1|  .72   .67| 84.2  80.0| 10004   |
|     9     66     99    -.22     .30| .85   -.9| .64  -1.1|  .73   .68| 85.4  82.1| 10009   |
|    19     51     67    -.90     .40| .79  -1.0| .49   -.8|  .73   .67| 93.2  85.1| 10019   |
|    13     40     65     .33     .37| .87   -.5| .88   -.2|  .73   .70| 86.7  83.3| 10013   |
|     5     48     86     .79     .31| .88   -.6| .80   -.7|  .74   .70| 85.5  81.7| 10005   |
|    17     46     67    -.15     .37| .80   -.9| .75   -.5|  .74   .69| 86.4  82.8| 10017   |
|    16     47     69    -.22     .37| .82   -.9| .57  -1.1|  .75   .69| 83.3  82.5| 10016   |
|     2     44     87    1.17     .30| .79  -1.3| .65  -1.3|  .75   .69| 85.9  81.2| 10002   |
|     8     65    100    -.12     .29| .72  -1.9| .60  -1.4|  .76   .68| 90.0  82.0| 10008   |
|     1     50     88     .74     .31| .73  -1.6| .62  -1.5|  .78   .71| 91.0  82.2| 10001   |
|------------------------------------+----------+----------+-----------+-----------+---------|
| MEAN    50.5   78.8     .00     .34|1.00    .0|1.04    .0|           | 83.2  82.5|         |
| S.D.     8.1   13.0     .76     .04| .24   1.2| .43   1.0|           |  5.9   1.1|         |
-------------------------------------------------------------------------------------------|
```

These items 10052, 10053, and 10054 were thus removed from the set for further analysis, due to their high OUTFIT values, and the analysis was repeated for a second iteration. This yielded another set of high OUTFIT values from the remaining items, and by continuing this process over two more iterations Items 10015, 10051, and 10055 were also removed. This yielded 19 test items which showed good correlation and acceptable OUTPUT and INFIT statistics. These results are shown in Table 23.

Learner A18 responded to small number of test items, all of which were removed during these iterations. There is thus no response data to compute the proficiency for this learner who was consequently removed from the data set for the computation of proficiency. However, this learner would still play a part in the determination of the

misconceptions when the WHOLE misconception was analyzed. The analysis continues with the remaining 104 learners.

Table 23. Place-Value: Item correlations after iteration 4

```
TABLE 10.1 PV - SCHOOLS A+B - CORRECT PV-AB-CORRECT-4.out.txt  Nov 24 21:52 2013
INPUT: 104 LEARNER  19 TESTITEM  REPORTED: 104 LEARNER  19 TESTITEM  2 CATS WINSTEPS 3.80.1
-------------------------------------------------------------------------------------------
LEARNER: REAL SEP.: 1.92  REL.: .79 ... TESTITEM: REAL SEP.: 1.73  REL.: .75

         TESTITEM STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|         |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| TESTITEM|
|------------------------------------+----------+----------+-----------+-----------+---------|
|    14     49     65   -1.26     .41|1.71   2.9|1.55    .9|A .53   .68| 68.5  83.9| 10014   |
|    12     47     66    -.83     .40|1.51   2.1|1.49   1.0|B .60   .71| 74.5  83.7| 10012   |
|     7     57    101     .45     .31|1.29   1.5|1.42   1.4|C .68   .75| 72.8  82.7| 10007   |
|    11     47     67    -.68     .40| .85   -.6|1.35    .8|D .73   .72| 87.5  84.1| 10011   |
|    10     69    101    -.66     .31|1.07    .4|1.29    .8|E .68   .71| 84.1  82.4| 10010   |
|    19     47     68    -.30     .39|1.10    .5|1.24    .7|F .68   .72| 83.3  83.0| 10020   |
|    13     40     65     .22     .40|1.08    .4|1.15    .5|G .73   .75| 83.3  85.2| 10013   |
|     6     56    100     .54     .31| .96   -.1|1.10    .5|H .74   .74| 83.8  82.4| 10006   |
|     1     50     88     .64     .34| .86   -.6|1.06    .3|I .78   .76| 90.0  83.2| 10001   |
|    17     46     69    -.15     .39|1.05    .3| .94    .0|J .72   .73| 81.5  82.8| 10018   |
|     9     66     99    -.44     .32| .92   -.4|1.04    .2|i .73   .72| 86.3  82.8| 10009   |
|     3     48     86     .83     .33| .93   -.3| .80   -.6|h .76   .74| 82.4  82.1| 10003   |
|    18     51     67   -1.04     .42| .92   -.3| .58   -.6|g .72   .69| 88.7  84.6| 10019   |
|     5     48     86     .73     .34| .91   -.4| .81   -.5|f .77   .75| 85.3  82.5| 10005   |
|    16     46     67    -.22     .39| .87   -.5| .77   -.5|e .75   .72| 84.9  83.4| 10017   |
|     4     40     86    1.61     .32| .81  -1.2| .56  -1.1|d .77   .72| 83.8  80.0| 10004   |
|    15     47     69    -.30     .39| .80   -.9| .58  -1.0|c .77   .72| 83.3  83.0| 10016   |
|     2     44     87    1.16     .33| .79  -1.2| .60  -1.2|b .79   .74| 85.7  81.5| 10002   |
|     8     65    100    -.32     .32| .75  -1.4| .62  -1.2|a .78   .73| 87.7  82.8| 10008   |
|------------------------------------+----------+----------+-----------+-----------+---------|
| MEAN    50.7   80.9     .00     .36|1.01    .0|1.00    .0|           | 83.0  83.0|         |
| S.D.     8.1   14.1     .76     .04| .24   1.1| .32    .8|           |  5.3   1.1|         |
-------------------------------------------------------------------------------------------
```

Identifying STABLE Learners

Learner proficiency was calculated so that high-proficiency learners could be identified and then removed for the second phase of the analysis, which then addressed the less proficient learners. The previous analysis has already performed the calculations and the top performing learners are shown in Table 24 in the sequence of LEARNER MEASURE. The MEASURE column contains the estimated learner ability, for example 4.41 for learner B03.

It was not possible to determine a learner measure which is higher than the most difficult of the items, which in this case is Item 10004 with a difficulty measure of 1.61, as in Table 23. Thus all of the learners with an estimated measure of at least 1.61, being all of those down the list to learner A78, were deemed to have full proficiency in this micro-domain when measured against this set of items. These learners are highlighted with a grey background.

Table 24. Place-Value: Learner measures for ability

TABLE 17.1 PV - SCHOOLS A+B - CORRECT PV-AB-CORRECT-4.out.txt Nov 23 15:02 2013
INPUT: 104 LEARNER 19 TESTITEM REPORTED: 104 LEARNER 19 TESTITEM 2 CATS WINSTEPS 3.80.1

LEARNER: REAL SEP.: 1.92 REL.: .79 ... TESTITEM: REAL SEP.: 1.73 REL.: .75

LEARNER STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASURE-A CORR.	PTMEASURE-A EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	LEARNER
33	19	19	4.41	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B03
55	19	19	4.41	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B25
60	19	19	4.41	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B30
62	19	19	4.41	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B32
77	19	19	4.41	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B47
31	15	15	4.31	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B01
65	15	15	4.31	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B35
76	15	15	4.31	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B46
3	14	14	4.26	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	A03
6	14	14	4.26	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	A07
8	14	14	4.26	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	A10
24	14	14	4.26	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	A68
48	14	14	4.16	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	B18
54	10	10	3.33	1.86	MAXIMUM MEASURE				.00	.00	100.0	100.0	B24
35	18	19	3.15	1.04	1.13	.4	2.47	1.3	-.21	.17	94.7	94.7	B05
37	18	19	3.15	1.04	1.12	.4	1.95	1.0	-.14	.17	94.7	94.7	B07
50	18	19	3.15	1.04	.80	.0	.29	-.4	.50	.17	94.7	94.7	B20
86	18	19	3.15	1.04	1.03	.3	.76	.2	.17	.17	94.7	94.7	B56
90	18	19	3.15	1.04	.98	.3	.59	.0	.26	.17	94.7	94.7	B60
94	18	19	3.15	1.04	1.02	.3	.70	.1	.20	.17	94.7	94.7	B64
95	18	19	3.15	1.04	1.04	.3	.83	.3	.14	.17	94.7	94.7	B65
96	18	19	3.15	1.04	1.03	.3	.76	.2	.17	.17	94.7	94.7	B66
39	14	15	3.03	1.05	1.05	.3	.87	.3	.14	.18	93.3	93.3	B09
43	14	15	3.03	1.05	1.05	.3	.87	.3	.14	.18	93.3	93.3	B13
14	13	14	2.97	1.06	.92	.2	.50	-.1	.35	.19	92.9	92.8	A31
15	13	14	2.97	1.06	1.18	.5	2.78	1.4	-.27	.19	92.9	92.8	A32
18	13	14	2.97	1.06	1.07	.4	.96	.4	.11	.19	92.9	92.8	A43
22	13	14	2.97	1.06	.92	.2	.50	-.1	.35	.19	92.9	92.8	A66
4	9	10	2.82	1.07	1.05	.3	.91	.3	.14	.18	90.0	89.9	A05
56	12	13	2.66	1.05	.92	.2	.55	-.1	.38	.17	92.3	92.3	B26
61	14	15	2.48	1.05	1.06	.4	1.14	.5	.00	.13	93.3	93.3	B31
49	17	19	2.36	.77	1.16	.5	1.26	.6	.00	.23	89.5	89.4	B19
51	17	19	2.36	.77	1.24	.6	2.05	1.3	-.22	.23	89.5	89.4	B21
63	17	19	2.36	.77	1.02	.2	1.03	.3	.20	.23	89.5	89.4	B33
66	17	19	2.36	.77	.97	.1	.62	-.3	.35	.23	89.5	89.4	B36
83	17	19	2.36	.77	1.04	.3	.76	-.1	.25	.23	89.5	89.4	B53
89	17	19	2.36	.77	.87	-.1	.65	-.2	.41	.23	89.5	89.4	B59
102	17	19	2.36	.77	.78	-.2	.45	-.6	.55	.23	89.5	89.4	B72
44	14	16	2.25	.78	.93	.1	1.06	.3	.27	.24	87.5	87.4	B14
34	13	15	2.22	.78	1.30	.7	2.18	1.5	-.36	.24	86.7	86.6	B04
74	13	15	2.22	.78	.80	-.2	.53	-.5	.55	.24	86.7	86.6	B44
7	12	14	2.15	.79	.71	-.5	.43	-.7	.62	.26	85.7	85.6	A08
27	12	14	2.15	.79	.91	.0	.60	-.3	.43	.26	85.7	85.6	A79
72	9	10	2.04	1.07	1.04	.3	.99	.3	.07	.14	90.0	90.0	B42
93	16	19	1.86	.65	1.22	.7	1.27	.6	-.01	.27	84.2	84.2	B63
73	12	15	1.70	.67	1.06	.3	1.08	.3	.20	.28	80.0	79.9	B43
26	11	14	1.62	.68	.78	-.5	.58	-.7	.59	.30	78.6	78.5	A78
84	15	19	1.48	.59	1.13	.5	1.76	1.5	-.01	.30	84.2	79.2	B54
100	15	19	1.48	.59	.73	-.8	.55	-1.0	.66	.30	84.2	79.2	B70
46	14	18	1.37	.60	1.35	1.1	1.85	1.7	-.21	.31	72.2	78.4	B16
… (ROWS REMOVED)													
98	0	2	-1.95	1.98	MINIMUM MEASURE				.00	.00	100.0	100.0	B68
59	2	16	-2.20	.77	1.02	.2	.83	.0	.22	.21	87.5	87.5	B29
38	1	10	-2.51	1.06	1.15	.4	2.05	1.2	-.52	.13	90.0	90.0	B08
20	1	13	-2.68	1.06	.81	.0	.38	-.3	.51	.19	92.3	92.3	A46
13	1	14	-2.69	1.06	1.13	.4	1.39	.7	-.03	.19	92.9	92.8	A30
19	1	14	-2.69	1.06	1.16	.5	2.07	1.1	-.17	.19	92.9	92.8	A44
29	1	14	-2.69	1.06	.81	.0	.36	-.3	.48	.19	92.9	92.8	A82
97	1	18	-3.10	1.04	.99	.3	.68	.1	.20	.14	94.4	94.4	B67
80	0	5	-3.20	1.89	MINIMUM MEASURE				.00	.00	100.0	100.0	B50
45	0	10	-3.24	1.87	MINIMUM MEASURE				.00	.00	100.0	100.0	B15
81	0	10	-3.24	1.87	MINIMUM MEASURE				.00	.00	100.0	100.0	B51
99	0	8	-3.56	1.86	MINIMUM MEASURE				.00	.00	100.0	100.0	B69
68	0	14	-3.81	1.85	MINIMUM MEASURE				.00	.00	100.0	100.0	B38
11	0	14	-3.98	1.85	MINIMUM MEASURE				.00	.00	100.0	100.0	A27
52	0	19	-4.38	1.84	MINIMUM MEASURE				.00	.00	100.0	100.0	B22
MEAN	9.3	14.8	.77	1.02	.99	.1	.99	.1			82.6	82.1	
S.D.	6.4	3.6	2.46	.47	.18	.7	.52	.8			12.5	10.7	

19 items remain from the original 25, since 6 of the items were removed due to high OUTFIT, and thus the maximum score is 19/19. Five learners achieved this maximum measure—being learners B03, B25, B30, B32, and B47. Another nine learners are indicated as MAXIMUM MEASURE, but they have lower ability measures, since they answered fewer items but still answered all of their items correctly. This includes the learners from B01 to B24 in Table 24.

Other learners almost achieved 100% success, including scores of 18/19 and 14/15. Learners B09 and B13 both have a measure of 3.03 and learner B31 had a measure of 2.48, yet all three of these learners scored 14/15. The reason for this difference is that they answered different test items correctly, which results in different calculated measures. So whereas these learners all achieved the same score of 14/15 when measured using classical test theory (CTT), where the score is based only upon the number of items answered correctly, Rasch analysis considers the difficulty of the items which were answered correctly when computing the learner measure.

The proficient learners were initially identified as having ability measures of at least 2.00, as suggested by Linacre (2013) as a suitable starting cut-point. This means that the learners from B03 to B42 in Table 24 were proficient and which I then removed before the analysis progressed to the next step of this analysis. However, many learners had ability estimates greater than 1.61, which is the highest measured difficulty of any of the 19 test items, and thus it was better to set the cut-point at this item measure. This cut-point resulted in 47 learners who were identified as having full proficiency and who have now been removed. However, many learners were at the borderline of proficiency and who demanded special attention.

Using my Development Stage model, these 47 learners were positioned into the STABLE development stage. These learners had evidence of internal schemas which were sufficient to obtain consistent success on place-value, and they showed no evidence of systematic errors in their responses. Some of these proficient learners have made mistakes, such as B19 with 17/18, and these mistakes were more likely "slips". Such slips result from non-cognitive causes, such as rushing to answer a question before reading it properly, or making a mistake when entering their answer into the system, or perhaps having a short break in their concentration in the class. However, it may be possible that some proficient learners continued to hold some misconceptions and these are addressed

later in this analysis when I seek to identify those learners who were in the IMMINENT stage of development who were near, but not quite at, the STABLE level of proficiency.

At the bottom of Table 24 some learners were identified as "MINIMUM MEASURE" since their calculated ability scores were below the level of the easiest items, meaning that these learners could not be measured on this test. These learners obtained no correct responses, such as B50 with 0/5 and B15 with 0/10. However, they were not automatically classified as ABSENT in my model, since their mistakes may have arisen from misconceptions, which then points to an active stage of learning. Thus the traditional CTT approaches to scoring of low-ability learners does not show the true conceptual development of these learners.

Learners were positioned into the ABSENT development stage based on a lack of evidence of systematic errors, and not through low ability scores, and to identify these learners I shifted my approach and examined those incorrect responses which pointed to the learners who were using the WHOLE misconception.

In summary, I started with 105 learners and then lost learner B18 when his/her response data was no longer available after removing test items during the initial iterations. From the remaining 104 learners I removed 47 whose ability measures were greater than 1.60, indicating that they were in the STABLE development stage in this micro-domain of place-value knowledge. I was left with 58 learners that I carried forward to the next step (58=105-47, considering that learner B18, removed previously, could now be reincorporated for the following steps).

Analyzing the WHOLE Misconception

I had previously eliminated 6 misfitting items when calculating the STABLE stage learners and I subsequently reintroduced these and used all of the original 25 items. For each item, the choices were coded with the WHOLE misconception when the choice would be selected by considering the decimal number as a whole number rather than as a decimal number.

The learners' responses were coded with the value 1 (success) if they selected this choice which is linked to the WHOLE misconception. All other responses were coded with 0 (fail), including the correct choices, and missing values were coded with the period (.). In this context the terms "success" and "fail" are misnomers, as previously discussed on page 89 concerning my reconceptualization of item difficulty, since this scoring is now

based on whether the learners selected a choice which reflects the WHOLE misconception.

Firstly, the items were checked for fit against the learners' usage of the WHOLE misconception, and the results of this are shown in Table 25, where the header of this table shows that this was conducted using 58 learners and 25 items.

Item 10055 shows negative correlation (-.08) for Item 10055, and there was also a low correlation for Items 10053 and 10051. These three items also had high OUTFIT values, greater than 1.5, and for the second iteration I removed these items, which yielded a results table in which additional items were then calculated as being out of fit with the consensus of the other items.

Table 25. Place-Value: WHOLE misconception correlation

TABLE 26.1 PV - SCHOOLS A+B - WHOLENUMBER PV-AB-WHOLE.out.txt Jun 20 22:19 2015
INPUT: 58 LEARNER 25 TESTITEM REPORTED: 58 LEARNER 25 TESTITEM 2 CATS WINSTEPS 3.80.1

LEARNER: REAL SEP.: 1.43 REL.: .67 ... TESTITEM: REAL SEP.: 2.82 REL.: .89

TESTITEM STATISTICS: CORRELATION ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASURE-A CORR.	PTMEASURE-A EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	TESTITEM
25	3	38	2.50	.64	1.26	.7	5.32	2.6	-.08	.31	92.1	92.0	10055
23	28	38	-1.96	.41	1.34	1.8	1.80	1.7	.11	.41	76.3	75.1	10053
21	23	39	-1.18	.38	1.52	2.8	1.88	2.7	.13	.50	51.3	73.0	10051
24	1	38	3.74	1.03	1.00	.3	.43	.0	.21	.18	97.4	97.3	10054
12	24	33	-1.72	.45	1.37	1.7	1.31	.8	.25	.46	66.7	76.7	10012
10	44	54	-2.56	.39	1.08	.5	1.50	1.0	.31	.40	81.1	81.8	10010
14	7	32	1.43	.49	1.11	.5	1.50	.9	.36	.46	81.3	80.7	10014
19	6	32	1.35	.52	1.08	.4	1.41	.8	.37	.47	90.6	83.7	10019
15	5	32	1.95	.54	.99	.1	1.46	.8	.37	.40	81.3	85.0	10015
7	26	54	-.42	.33	1.32	1.9	1.32	1.4	.38	.55	67.9	74.2	10007
3	24	42	-.81	.37	1.02	.2	1.48	1.8	.46	.51	75.6	73.0	10003
11	11	33	.61	.44	1.14	.7	.90	-.2	.48	.52	63.6	76.7	10011
13	14	32	.03	.43	1.04	.3	1.24	.9	.50	.55	78.1	74.9	10013
16	11	34	.24	.44	.99	.0	1.33	.9	.53	.55	78.8	77.5	10016
18	10	34	.43	.45	1.02	.1	.92	.0	.53	.54	78.8	78.3	10018
22	8	39	1.15	.45	.83	-.7	.52	-1.0	.59	.45	82.1	81.5	10052
9	21	52	-.05	.34	.86	-.9	.77	-.9	.63	.54	74.5	74.4	10009
1	30	43	-1.42	.38	.75	-1.6	.58	-1.3	.63	.46	81.0	75.0	10001
6	27	53	-.62	.33	.83	-1.1	.78	-1.0	.64	.54	78.8	74.1	10006
2	28	44	-1.11	.36	.74	-1.7	.63	-1.4	.66	.50	83.7	74.3	10002
17	9	32	.63	.47	.78	-.9	.54	-1.0	.66	.52	81.3	79.0	10017
8	18	53	.39	.35	.78	-1.4	.61	-1.4	.67	.54	84.6	76.3	10008
5	28	43	-1.23	.37	.68	-2.1	.59	-1.4	.68	.49	90.5	74.2	10005
4	28	43	-1.21	.37	.68	-2.1	.53	-1.7	.70	.49	81.0	74.5	10004
20	13	33	-.14	.43	.63	-1.9	.49	-1.8	.77	.55	84.8	76.3	10020
MEAN	17.9	40.0	.00	.44	.99	-.1	1.19	.1			79.3	78.4	
S.D.	10.6	7.7	1.45	.14	.24	1.3	.95	1.3			9.4	5.8	

After two more iterations I eventually removed eight items, being Items 10010, 10012, 10014, 10015, 10051, 10053, 10054, 10055. After these four iterations, this yielded the table of results shown in Table 26, which showed good correlation in the PTMEASURE column, and also generally good INFIT and OUTFIT values, with the exception of Items 10007 and 10013. Other than these cases the test items were quite

well-behaved, meaning that there was a good correlation between the item response data and the learner measures in terms of this WHOLE misconception and that there was a good fit to the Rasch model.

Table 26. Place-Value: WHOLE misconception iteration 4

```
TABLE 26.1 PV - SCHOOLS A+B - WHOLENUMB PV-AB-WHOLE-4.out.txt  Jun 20 23:05 2015
INPUT: 58 LEARNER  17 TESTITEM  REPORTED: 58 LEARNER  17 TESTITEM  2 CATS WINSTEPS 3.80.1
-------------------------------------------------------------------------------
LEARNER: REAL SEP.: 1.44  REL.: .67 ... TESTITEM: REAL SEP.: 1.79  REL.: .76

         TESTITEM STATISTICS:  CORRELATION ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL          MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|         |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| TESTITEM|
|------------------------------------+----------+----------+-----------+-----------+---------|
|     7     26     54    -.25     .36|1.72   3.3|1.97   2.7|  .38   .65| 60.4  77.4| 10007   |
|    15      6     32    1.45     .54|1.16    .6|1.54    .9|  .38   .48| 88.9  81.3| 10019   |
|    11     14     32     .40     .46|1.35   1.5|2.22   2.3|  .45   .62| 70.0  77.3| 10013   |
|    17      8     39    1.66     .48|1.01    .1| .59   -.4|  .51   .49| 74.3  81.4| 10052   |
|    14     10     34     .44     .48|1.07    .4| .99    .1|  .56   .58| 77.8  75.9| 10018   |
|    10     11     33    1.05     .46|1.02    .2| .78   -.3|  .57   .56| 67.7  76.0| 10011   |
|    12     11     34     .22     .47|1.05    .3|1.05    .3|  .58   .60| 74.1  75.5| 10016   |
|     3     24     42    -.72     .41|1.11    .6|1.76   1.8|  .58   .64| 74.4  78.6| 10003   |
|    13      9     32     .67     .49| .83   -.7| .61   -.8|  .64   .56| 77.8  77.1| 10017   |
|     8     18     53     .71     .37| .91   -.5| .71   -.8|  .65   .60| 74.5  77.0| 10008   |
|     1     30     43   -1.48     .43| .93   -.2| .69   -.5|  .66   .62| 77.5  80.6| 10001   |
|     6     27     53    -.50     .37| .97   -.1| .97    .0|  .66   .65| 76.6  77.5| 10006   |
|     4     28     43   -1.25     .42| .85   -.6| .64   -.7|  .70   .63| 80.0  79.7| 10004   |
|     9     21     52     .19     .37| .82  -1.0| .64  -1.2|  .70   .63| 82.6  76.8| 10009   |
|     2     28     44   -1.10     .41| .76  -1.1| .68   -.7|  .72   .64| 87.8  79.7| 10002   |
|    16     13     33    -.22     .47| .69  -1.5| .56  -1.4|  .75   .63| 85.2  75.2| 10020   |
|     5     28     43   -1.27     .42| .58  -2.2| .45  -1.4|  .78   .63| 95.0  79.5| 10005   |
|------------------------------------+----------+----------+-----------+-----------+---------|
| MEAN    18.4   40.9     .00     .44| .99   -.1| .99    .0|           | 77.9  78.0|         |
| S.D.     8.3    7.9     .93     .05| .26   1.2| .53   1.2|           |  8.1   2.0|         |
```

I now had a data set with responses from 58 learners and for the remaining 17 items. My task was to determine the extent to which these 58 learners exhibited evidence of using the WHOLE misconception in their responses to these test items. This information was provided using WinSteps output in Table 27, which shows the learners in the sequence of their estimated "MEASURE" which for the learner measure was the extent to which their responses were accountable to the WHOLE misconception.

In the first row of this table, learner B51 had 9/9 responses which were accountable to the WHOLE misconception with a measure of 2.98, and learner B22 had achieved 15/16 responses which were accountable to the WHOLE misconception with a measure of 2.90. The seven learners from B51 down to B50 had no more than one response which was not accounted for by the WHOLE misconception.

At the bottom of this table, the learner responses showed no indication of the WHOLE misconception, however, some of these learners had achieved a relatively high ability mark but their calculated ability measure was insufficient to position them in the STABLE stage.

Table 27. Place-Value: WHOLE misconception

TABLE 17.1 PV - SCHOOLS A+B - WHOLENUMB PV-AB-WHOLE-4.out.txt Jun 20 23:05 2015
INPUT: 58 LEARNER 17 TESTITEM REPORTED: 58 LEARNER 17 TESTITEM 2 CATS WINSTEPS 3.80.1

LEARNER: REAL SEP.: 1.44 REL.: .67 ... TESTITEM: REAL SEP.: 1.79 REL.: .76

LEARNER STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASURE-A CORR.	PTMEASURE-A EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	LEARNER
44	9	9	2.98	1.88		MAXIMUM MEASURE			.00	.00	100.0	100.0	B51
29	15	16	2.90	1.05	1.16	.5	2.42	1.2	-.19	.18	93.8	93.7	B22
4	11	12	2.57	1.09	.69	-.2	.27	-.3	.58	.27	91.7	91.6	A14
7	11	12	2.57	1.09	1.05	.3	.63	.1	.28	.27	91.7	91.6	A27
11	10	11	2.55	1.09	1.13	.4	.88	.4	.17	.26	90.9	90.8	A42
49	9	10	2.18	1.12	.58	-.4	.24	-.5	.72	.33	90.0	90.0	B58
43	4	5	1.99	1.15	1.17	.5	1.16	.5	-.06	.23	80.0	80.0	B50
15	10	12	1.69	.83	.61	-.8	.36	-.8	.72	.35	83.3	83.3	A67
33	11	14	1.63	.69	1.15	.5	1.03	.3	.22	.34	71.4	78.6	B29
57	13	17	1.39	.61	.77	-.7	.61	-.7	.59	.35	88.2	77.4	B76
10	9	11	1.33	.82	.90	-.1	.73	-.2	.44	.31	81.8	81.8	A41
54	5	7	1.25	.87	.74	-.6	.66	-.6	.70	.28	85.7	72.8	B69
35	11	15	1.10	.63	.96	.0	.76	-.4	.47	.38	73.3	76.3	B37
2	9	12	1.10	.73	1.30	.9	1.25	.6	.14	.40	66.7	77.2	A02
6	9	12	1.10	.73	.64	-1.0	.47	-1.0	.74	.40	83.3	77.2	A26
9	9	12	1.10	.73	.73	-.7	.53	-.8	.67	.40	83.3	77.2	A30
13	8	11	1.05	.74	1.47	1.3	1.64	1.2	-.06	.39	54.5	75.4	A46
52	11	16	.99	.58	.95	-.1	.83	-.3	.45	.38	75.0	72.3	B67
5	0	1	.80	2.19		MINIMUM MEASURE			.00	.00	100.0	100.0	A18
12	8	12	.61	.68	.99	.1	1.19	.6	.39	.42	75.0	72.9	A44
17	8	12	.61	.68	1.18	.7	1.23	.7	.26	.42	58.3	72.9	A80
18	8	12	.61	.68	1.47	1.4	1.47	1.1	.01	.42	58.3	72.9	A82
45	6	11	.57	.63	1.20	1.2	1.23	1.3	-.10	.26	54.5	61.5	B52
37	9	14	.46	.60	.76	-1.0	.67	-1.1	.65	.37	71.4	69.2	B39
39	9	14	.46	.60	1.17	.8	1.23	.8	.17	.37	57.1	69.2	B41
34	8	14	.11	.58	.72	-1.4	.66	-1.4	.70	.38	85.7	67.3	B34
23	4	9	.05	.69	.98	.0	.96	-.1	.31	.26	66.7	61.8	B10
36	8	15	.05	.57	.87	-.5	.80	-.7	.55	.41	73.3	68.2	B38
21	6	11	-.18	.66	.98	.0	.93	-.2	.42	.38	54.5	67.6	B06
58	7	17	-.43	.54	.65	-1.8	.59	-1.7	.75	.40	82.4	69.2	B85
40	3	9	-.45	.73	1.32	1.2	1.63	1.7	-.42	.24	55.6	67.0	B45
1	5	12	-.65	.64	.83	-.6	.76	-.6	.57	.41	83.3	68.7	A01
48	6	17	-.73	.55	1.30	1.3	1.31	1.0	.11	.39	52.9	70.9	B57
31	5	15	-.93	.59	.56	-2.1	.48	-1.5	.80	.38	93.3	71.1	B27
22	2	9	-1.04	.82	1.34	.9	2.09	1.7	-.61	.21	77.8	77.8	B08
8	4	12	-1.07	.66	1.04	.2	.92	.0	.37	.38	66.7	70.0	A28
30	4	15	-1.30	.63	1.63	2.0	2.32	2.1	-.36	.35	66.7	73.8	B23
20	4	16	-1.34	.62	.94	-.1	.73	-.4	.45	.36	68.8	75.4	B02
50	4	17	-1.39	.61	.85	-.4	.65	-.6	.53	.35	82.4	76.7	B61
53	0	2	-1.41	1.98		MINIMUM MEASURE			.00	.00	100.0	100.0	B68
26	3	10	-1.43	.73	.98	.0	.86	-.1	.36	.33	60.0	70.1	B15
41	1	9	-1.89	1.07	.85	.1	.55	-.2	.50	.15	88.9	88.9	B48
24	1	10	-1.93	1.07	1.05	.3	.96	.3	.11	.17	90.0	90.0	B11
51	0	3	-2.08	1.93		MINIMUM MEASURE			.00	.00	100.0	100.0	B62
42	2	11	-2.09	.81	1.06	.3	.97	.2	.21	.27	81.8	81.7	B49
14	2	12	-2.10	.81	1.07	.3	1.35	.7	.15	.28	83.3	83.2	A50
16	2	12	-2.10	.81	1.05	.3	.90	.2	.25	.28	83.3	83.2	A77
32	2	15	-2.26	.79	.88	-.1	.58	-.3	.42	.26	86.7	86.6	B28
25	2	17	-2.32	.78	1.14	.4	.96	.3	.15	.26	88.2	88.2	B12
28	2	17	-2.32	.78	.94	.1	.71	-.1	.34	.26	88.2	88.2	B17
19	1	10	-2.88	1.07	1.12	.4	1.19	.6	.04	.19	90.0	89.9	A86
27	1	16	-2.93	1.05	1.14	.4	1.82	.9	-.09	.19	93.8	93.7	B16
3	1	12	-2.95	1.07	1.14	.4	1.32	.7	.01	.20	91.7	91.6	A06
38	0	9	-3.20	1.86		MINIMUM MEASURE			.00	.00	100.0	100.0	B40
47	0	9	-3.20	1.86		MINIMUM MEASURE			.00	.00	100.0	100.0	B55
56	0	10	-3.23	1.86		MINIMUM MEASURE			.00	.00	100.0	100.0	B71
46	0	17	-4.40	1.85		MINIMUM MEASURE			.00	.00	100.0	100.0	B54
55	0	17	-4.40	1.85		MINIMUM MEASURE			.00	.00	100.0	100.0	B70
MEAN	5.4	12.0	-.39	.96	1.00	.1	.99	.1			77.5	78.1	
S.D.	4.0	3.7	1.89	.45	.24	.8	.48	.8			12.6	8.9	

I then differentiated learners who used the WHOLE misconception exclusively from learners whose responses were accounted for by other misconceptions or by

guessing. I chose a learner measure (in the MEASURE column) of 1.00 as the cut-point, which included 18/58 learners, from learner B22 down to learner B67—with a measure of 0.99, which I treated as sufficiently close to 1.00—in Table 27. The cut-points were chosen to best differentiate the learners into the various stages.

The resulting selection included learner A02, for whom only 9/12 were attributable to the WHOLE misconception which may seem to be too low compared to other learners selected. This is explained by the manner in which Rasch calculates measures for the items, and the specific 9 items which were answered by learner A02, since not all combinations of 9/12 items will produce the same learner measure. Some test items were better indicators of the WHOLE misconception than others, and this is the reason why learners A44 and A80, both of whom scored 8/12 test items, calculated measures of 0.61 which was far lower than the measure of 1.10 for learner A82. The test items which they answered had different levels of suitability as an indicator of the WHOLE misconception.

Returning to my Development Stages, I initially removed 47 learners who were in the STABLE stage. The remaining learners than had to be positioned into the remaining stages of ABSENT, EMERGENT, ACTIVE and IMMINENT, since my model required that every learner was placed into one, and only one, of these stages. The three primary stages—ABSENT, ACTIVE and STABLE—represent the Zones of Competence, Learning and Incompetence. The other two stages—EMERGENT and IMMINENT—are transition stages between the primary stages.

There were 18/58 learners who used the WHOLE misconception to account for their responses, I positioned these learners in the ACTIVE stage, since their responses indicated active use of misconceptions. These 18 learners in the ACTIVE stage complemented the 47 learners who were positioned into the STABLE stage, which was 65 out of the original 105 learners, thus leaving 40 learners who were spread over the three remaining Development Stages.

The IMMINENT stage is for learners who were achieving a high level of success, but whose knowledge could not yet be considered STABLE. These learners were making mistakes that may have resulted from some use of misconceptions or they may have been making slips for which there is no systematic explanation. I thus moved learners into the IMMINENT stage if they achieved a high ability measure, but which was insufficient to

be considered as STABLE. To differentiate the IMMINENT stage learners, I selected a cut-point of 0.60 on the learner ability measures, and this resulted in 12 learners with ability measures between 0.60 and 1.60, all of whom scored low on the WHOLE misconception measure, meaning that the errors which they made were generally not accountable to the use of the WHOLE misconception.

To summarize, I have positioned 47 of the original 105 learners into the STABLE stage, 12 into the IMMINENT stage, and 18 into the ACTIVE stage, which left 28 learners to be classified into the EMERGENT and ABSENT stages. The difference between these two Development Stages is that the EMERGENT stage learners showed some evidence of using misconceptions to account for their incorrect responses, whereas ABSENT stage learners showed little or no such evidence, and thus their responses could only be inferred as random guesses.

I adopted a simple rule for the EMERGENT stage, which selected learners who had higher ability measures, but for whom the incorrect responses did not show sufficient evidence of the systematic usage of the WHOLE misconception to warrant their being in the ACTIVE stage. I initially set a cut-point ability measure of 0.00 as the basis for these learners, which meant that 3 learners were positioned into the EMERGENT state, and the remaining 25 were in the ABSENT state. These results were shown in Table 28 which shows the measures for these remaining 28 learners in decreasing sequence of their WHOLE measure. Each of these learners had an ability measure which was insufficient to allocate them to the STABLE stage or IMMINENT stages, and also a WHOLE measure which was insufficient to position them into the ACTIVE or EMERGENT stages.

The ABSENT stage was used for situations in which I could not account for the learner responses either on the basis of ability or through their use of the WHOLE misconception. This classification had some anomalies, such as learners B67 down to B87 in Table 28, who had more than 50% of their responses which were attributable to the WHOLE misconception, identified by the WScore and WCount columns in the table. For example, learner B67 scored 11/16 on the WHOLE misconception, but the estimated WHOLE measure of 0.99 was insufficient to classify this learner as ACTIVE. In essence, the Rasch measures were not sufficiently high to provide evidence that that these learners were using the WHOLE misconceptions for their responses. The problem is that the

Rasch analysis appears to contradict my rejection of learners who score 11/16 or 8/12 on this WHOLE misconception. This behaviour is explained by the specific items these learners were answering and whether these items were themselves good indicators of this misconception.

Table 28. Place-Value: EMERGENT+ABSENT learners

Learner Code	Correct Measure	WHOLE Measure	AScore	ACount	WScore	WCount
B67	-3.1	0.99	1	18	11	16
A80	-1.87	0.61	2	14	8	12
A44	-2.69	0.61	1	14	8	12
A82	-2.69	0.61	1	14	8	12
B52	-1.32	0.57	4	14	6	11
B41	-1.36	0.46	3	15	9	14
B39	-1.88	0.46	2	15	9	14
B34	-0.29	0.11	6	15	8	14
B10	-0.66	0.05	4	10	4	9
B38	-3.81	0.05	0	14	8	15
B06	-0.19	-0.18	6	14	6	11
B85	-0.88	-0.43	6	19	7	17
B45	-1.12	-0.45	3	10	3	9
A01	-1.87	-0.65	2	14	5	12
B57	-1.48	-0.73	4	19	6	17
B27	0.01	-0.93	7	15	5	15
B08	-2.51	-1.04	1	10	2	9
A28	-0.91	-1.07	4	14	4	12
B02	-0.63	-1.34	6	17	4	16
B61	-1.16	-1.39	5	19	4	17
B68	-1.95	-1.41	0	2	0	2
B15	-3.24	-1.43	0	10	3	10
B11	-1.68	-1.93	2	10	1	10
B62	-0.34	-2.08	2	4	0	3
A50	-0.54	-2.1	5	14	2	12
A86	0	-2.88	4	10	1	10
B40	-0.24	-3.2	5	10	0	9
B71	0.18	-3.23	6	10	0	10

Some of the learners who scored relatively high on ability, such as B71 with 6/10, and B27 with 7/15, were positioned within the EMERGENT or ABSENT stages because of the inability of this model to classify their incorrect responses as the WHOLE misconception with a sufficient level of confidence. Thus, I could not account for why

these learners were making mistakes using my model of misconceptions within place-value knowledge and some other explanation must be sought, which may include the following:

- pure guessing by the learner
- additional, unknown misconceptions
- extraneous factors such as inability to concentrate
- learners helping other learners in a blind-leading-the-blind scenario
- a limitation in the implementation of the WHOLE misconception.

Given these limitations I assume that, in the absence of any other evidence, that these responses were the result of guessing. This conclusion could be refined through improved items which could test for other ways of thinking.

These final 28 learners were now split into the EMERGENT and ABSENT groups based upon their level of usage of the WHOLE misconception. Following this analysis, I selected another cut-point, at -1.00, as the limit above which the learners appeared to be using the WHOLE misconception for some incorrect responses. The top 16 learners had WHOLE measures greater than -1.0, and they showed limited usage of the WHOLE misconception, and were positioned in the EMERGENT stage. The remaining 12 learners showed insufficient evidence of ability and also little or no evidence of using the WHOLE misconception, and were considered as having no schemas that were consistently applied, and were positioned in the ABSENT stage.

Answering the Research Questions

RQ1 (EFFECTIVENESS)

For the WHOLE misconception analysis, 17 items were used, selected from the original 25 for which the correlation and the INFIT and OUTFIT values were acceptable. From Table 26 the highest item measure is 1.66 for Item 10052 and the lowest is -1.48 for Item 10001, which are both measures of the item "difficulty" of the WHOLE misconception.

A low Rasch item measure implies that the item is "easy", meaning that it was answered successfully by most learners, and a high measure implies "difficult", in which the fewest learners answered this item successfully. In the context of measuring WHOLE misconception usage, my interest was in distinguishing test items which were good indicators of the WHOLE misconception. Thus my focus was on the "easy" test items,

and particularly those with scores of less than -1.00. Applying this logic back to the results of Table 26 I infer that items 10001, 10002, 10004, and 10005 are likely better than the others since they detected the WHOLE misconception for more learners than for the other items, and these items are also valid test items for this construct which were not removed due to OUTFIT or INFIT behaviours. However, these items were all among the first test items presented to the learners and the argument for the use of the "easy" items must be considered in the light of the sequence in which the learners were presented with the test items. This yielded another surprising result—that the more test items which were presented to the learners, the less they were inclined to select the WHOLE misconception choice. It thus appears that the learners were learning through merely answering the items, since feedback was not presented to the learners until the end of each test.

My recommendation is to use only those test items that are better suited to the discovery of this misconception. These "easy" items are valid indicators of the WHOLE misconceptions, where "easy" is interpreted that a learner with this misconception is more likely to select the MCQ choice which is linked to this misconception.

These items can be qualitatively analyzed, to identify features of the items which may render these suitable for this diagnosis. Item 10005 asked for the place-value of 6 in the decimal number 20.0067, for which the correct response was "thousandths", and for which both "tenths" and "tens" were treated as WHOLE misconceptions. The frequency count of responses in Table 21 showed that these two choices of "tens" (column 3) and "tenths" (column 5) made up the largest fraction of incorrect responses.

I compared this to Item 10011, which asked for the digit in the "tenths" position of decimal number 200.3154. Of the 19 learners who selected incorrect choices, 11 selected 5 (column 6 in the table), which was the digit which would have been in the "tens" position if the entire number was seen as a whole number, and thus this was a result of using the WHOLE misconception.

However, this item also had a number of responses from the digits 0 and 1 and had a high measure, meaning that the learners who used the WHOLE misconception were less likely to select the WHOLE choice (being the digit 5) when answering this item, when compared to the Item 10005 above.

RQ2 (EFFICIENCY)

Learners were positioned in the ACTIVE stage based upon the evidence from the Rasch analysis of their use of the WHOLE misconceptions, and learners showing less evidence of this usage were positioned in the EMERGENT stage. This process was conducted using the 25 items in this test set, which was reduced to 17 items which were determined to be suitable by considering their correlation and their fit statistics. I now ask which combination of these items can produce similar results, and how many items are needed.

Some of the items were better as diagnostic indicators of the WHOLE misconception, as established by RQ1, and thus one useful place to find a minimal set of items is with these "better" items. Thus, the problem of efficiency is not only concerned with how many items are needed, but is also concerned with which specific items are needed to achieve the maximum benefit in valid inferences arising from the smallest number of items asked.

An alternative, and exhaustive, approach is to determine the learner measures by using subsets of the items—using one, two, three, four, or more items at a time—using the previously calculated item measures, and by comparing these new learner measures to the learner measures already calculated using the entire set of items. However, this approach is computationally challenging given the very large number of combinations over which this calculation must be performed. There were 17 items which were used for this place-value, and the results of the Rasch analysis are shown in Table 26 on page 187. A subset of one item from this set then provides 18 alternatives. Using two items there are 18x17=306 combinations, for three items this is 18*17*16 = 4896 combinations, and for four this rises to 73,440 combinations. For each of these combinations of items the learner measures must be calculated and then compared to their original measures from using the full set of 17 items. The best combination would be that which is the closest to the calculated estimations of the learner WHOLE misconception measure as indicated in Table 27 on page 188, in which the notion of "closest" must be formalized.

This approach is infeasible without incurring considerable computational effort and thus it was my evaluation that a few best-fitting items be selected. I thus selected the items with the lowest measures, which had the greatest propensity to be selected by learners, as identified in the WHOLE misconception analyses in Table 26 above. These are Items 10001, 10005, 10004, and 10002, in that sequence. Future studies may refine

this approach, perhaps using the exhaustive analysis of every possible combination as I contemplate above.

RQ3 (SELF-KNOWLEDGE)

There were 2489 total responses to questions posed in this place-value micro-domain and of these 834 were incorrect responses. These incorrect responses were analyzed to determine whether it was possible to infer the development stage from learners using the decision matrix I presented in Table 10.

Of the 834 incorrect responses, 387 were indicated by the learners as Easy, and an additional 298 as Just Right. These 685 (387+298) responses comprised 82% of the incorrect responses, and were from learners who were likely to have believed they were answering the question correctly.

As an example, Item 10001 was previously identified as being a good candidate for usage in a diagnostic environment, and it exhibited the maximum value of potential evidence for the WHOLE misconception. This test item asked the learner to identify the place-value of the digit 7 in the number 36.748. There were 38 (22+6+1+9) learners who selected an incorrect response of which 22 selected “hundreds” and another 9 learners selected “tenths”. Of the 22 who selected “hundreds”, which is an indication of the WHOLE misconception by treating the decimal number as through it is a whole number by disregarding the decimal point, 18/22 learners identified this question as Easy or Just Right, and only one learner identified this test item as Difficult. Three of the learners did not answer this question on difficulty.

For those learners who used this misconception, there was thus no additional value to be obtained from asking whether they found this item Easy, Just Right, or Difficult, since the Rasch analysis of the results was itself sufficient for 18/22 of the cases.

The learners who marked this item as Difficult may have been in the ABSENT development stage, and this could not be directly inferred from an examination of the responses. As a result, there is some value in knowing the learners’ perception of the difficulty, even if this is being used solely to isolate those learners who found the test item Difficult and for whom I could then infer the lack of a suitable schema and consequently the likelihood of a guessed response. The alternative is the presence of a schema and hence evidence of a misconception. However, this must be balanced against the additional time and effort in gathering and the analysis of the data.

Summary

The place-value micro-domain has been analyzed using the standard approach I introduced in Chapter 4.

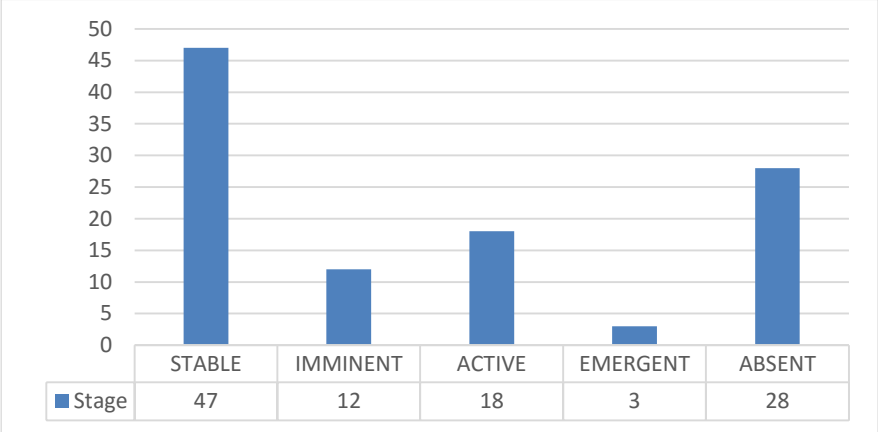


Figure 27. Place-Value : Learners by Stage

The distribution of the learners over the five Development Stages is given in Figure 27. This shows the expected pattern that IMMINENT and EMERGENT transition stages have less learners than those in the primary stages of STABLE, ACTIVE, and ABSENT.

My diagnostic model is incomplete for place-value knowledge but can improve over time as other ways of thinking are incorporated into the model from prior studies. The model will also improve as new, better diagnostic test items are added to the item bank. By improving this diagnostic model, it can help to identify various ways of thinking, and has the potential to improve the effectiveness of diagnostic assessment by accounting for an increasing number of misconceptions which are exposed through learner responses.

6.5 Micro-Domain DO - Decimal Number Ordering

Initial Analysis of Responses

This micro-domain can be reduced to the problem of understanding the relative magnitude of decimal numbers. The item bank contains 30 items, some of which were drawn from prior studies, and others which were created to address specific misconceptions.

Two tests were conducted in the second online lesson, with Test 1 comprising Items 10021-10040 and Test 2 comprising Items 10041-10050. A short instructional sequence was included between these tests. The items were structured mostly as the problem of finding the largest or smallest of two decimal numbers. Five of the items also included a choice concerning whether the two decimal numbers were equal. Three of the items comprised five decimal numbers as choices. A total of 97 learners responded to these tests.

The Items 10029 and 10049 presented choices which were whole numbers rather than decimal fractions and could be considered as "trick" questions. These were both removed from the detailed analysis, but they warrant a short analysis. Their choices did not contain a decimal point and were presented vertically in a staggered form, left-aligned, in which the same size place-values were not in line.

Item 10029, shown in Figure 28, was presented around halfway through Test 1, and 6/91 learners selected 1111 as the smallest. This would be correct if the choices were decimal fractions with a leading decimal point. This item was introduced to explore problems with recognizing differences between whole numbers and decimals. Given than only six learners responded in this way, one explanation is that these few learners did not read the question properly.

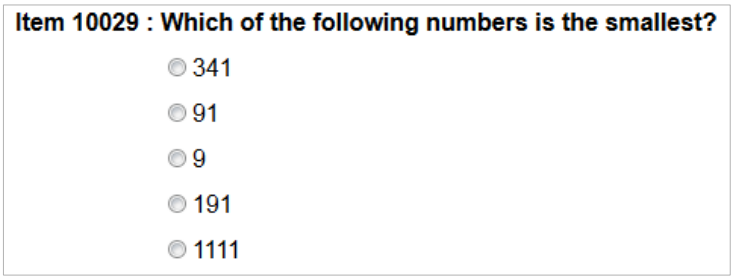


Figure 28. Item 10029

Item 10049, shown in Figure 29, used the specific wording "decimal numbers" rather than simply "numbers". 58/78 learners selected the correct response (9), and 18/78 selected the last option "55555" as being the smallest, which was a far higher frequency than for the similar choice in Item 10029. These items differed in the sequencing of the choices, and in Item 10049's use of the term "decimal numbers". Of these 18/78 learners selecting "55555", five were positioned in the STABLE stage and these were among very few mistakes made by these STABLE stage learners, and whereas these should be considered as slips they may have had a consistent basis which would cause them to be

late-stage misconceptions. In such cases, these learners could then be considered to be in the IMMINENT stage, given that they had lingering misconceptions. Without further data it was not possible to determine whether these were slips or misconceptions.

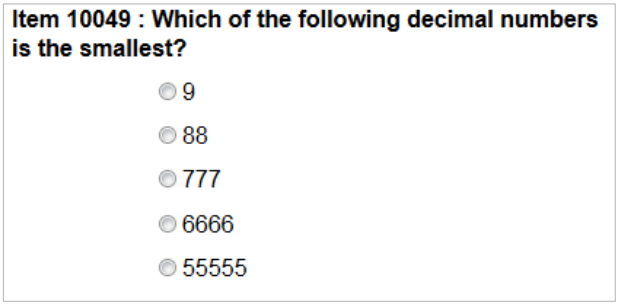


Figure 29. Item 10049

A correct response may result from knowing the subject matter, from guessing, or from the use of a misconception. Some items had the same choice which was both the correct response and which was also indicative of one or more misconceptions. This complicated the determination of the conceptual basis of the responses, and I have chosen to conduct two analyses based on different approaches to selecting the data for analysis.

For the first analysis, the total set of responses was used. This produced a larger data set for analysis, but correct responses may have been selected based on a misconception. My second analysis used only error responses and removed all correct responses. This was a reduced data set which avoided the issue of a correct response also resulting from a misconception. This resulted in smaller datasets of between three and seven test items for each of the misconceptions.

Identifying STABLE Learners

This analysis follows a similar structure to the previous micro-domain, but omits much of the explanations and associated tables.

Firstly, STABLE stage learners were identified and removed from further analysis. This analysis was performed on the basis of the correct choices obtained from each of the 28 test items in this micro-domain, with Item 10029 and Item 10049 of the original 30 items removed as not being a true part of this micro-domain.

The cutoff ability measure was set at 2.00 for STABLE stage learners, resulting in the selection of 16/97 learners being positioned as STABLE. These learners obtained relatively high correct scores, such as 28/28, 19/19, and 27/28.

The remaining 81 (=97-16) learners were filtered to identify those in the IMMINENT stage, based on an ability measure in the range 1.00-2.00, resulting in the selection of 20 learners into this stage. Thus, from the 97 learners, 16 were classified as STABLE and another 20 as IMMINENT, leaving 61 positioned in the ACTIVE, EMERGENT, and ABSENT stages. The IMMINENT stage learners were retained for the analysis of the misconceptions, while STABLE stage learners were removed as per the standard approach to analysis. Thus 81 learners were retained for the analysis of the misconceptions.

To continue the arguments from the previous micro-domain, the STABLE learners were removed since they made too few mistakes to add value to the analysis of items suited to diagnose misconceptions. Also, the approach to measuring low-proficiency must be different from traditional ability measurement, to account for their responses in terms of misconceptions.

Analysis of Decimal Number Misconceptions

This analysis used a subset of Steinle's coding structure for ways of thinking in decimal numbers (Steinle, 2004a, Chapter 3), which was presented as Table 4 on page 111. Steinle's codes used in this analysis were A1, A2, U1, U2, L1, L2, L3, and S3 and these were applied to each of the 28 test items in this micro-domain. The test items differed from Steinle's in three ways. Firstly, the use of three items which have five choices rather than only two choices. Secondly, three test items had whole number parts which were different between the choices, such as 2.414 vs. 3.001 in Item 10035. Finally, in the use of leading and trailing zeros in some items. Steinle's S1 was not used in this study, due to the lack of items which target this code.

For items with only two choices it was a challenge to elicit evidence of misconceptions, since a learner's response may have been based on true proficiency or may have resulted from one of many misconceptions. Each misconception was analyzed using a separate Rasch analysis, effectively performing these in parallel. These analyses were performed only where there was sufficient data available, since some misconceptions occurred less frequently than others as the conceptual basis for the rich distractors in my items.

Steinle's (2004a) code A1 (task expert) was established by positioning learners into the STABLE stage, and code A2 (money thinking) was aligned to the IMMINENT

stage, considering the types of errors which were made by otherwise expert learners. This analysis examined patterns of response errors which indicated misconceptions that occurred in otherwise proficient learners before reaching full proficiency. The A2 code was interpreted for this study to comprise learners who had a proficiency measure of between 1.00 and 2.00, and who were positioned in the IMMINENT stage. This approach is distinct from Steinle's definition of the A2 code which is applied to learners who have a partial set of rules derived from money thinking.

The U1 and U2 codes represent unclassified errors, and apply to learner responses which did not fit within the coding system, and which could not be further analyzed. These responses were treated as random errors, in the absence of any further knowledge about the ways of thinking. Whereas U1 was treated as unclassified, U2 was used to indicate misread, misrule, or mischievous, and thus may indicate potential patterns of behaviour. As an example, U2 could have been used to classify an otherwise expert learner who misread smallest for largest and solved the wrong problem; however if this was to be applied, this behaviour should then have been applied consistently. These U1/U2 codes have been used to position learners into the ABSENT stage, since it is beyond the diagnostic capability of the test items and the inferential processes to identify and to uncover misconceptions which were the basis for these responses. Over time, such unknown but systematic patterns of errors could provide the basis for research into more specialized misconceptions. For my purposes, expert learners who are making a few mistakes, such as misreading, would be classified into my IMMINENT stage.

The codes L1, L2 L3, and S3 were analyzed using parallel Rasch analyses, and the resulting learner measures were an indication of the extent to which these ways of thinking accounted for the responses to the test items. Some of the items were not suited to measure specific codes and where the evidence of the learner use of these ways of thinking did not correlate with an item then this item was removed for further analysis. Items with no positive results were also removed, in which no learner selected the rich distractor linked to the code.

Poor correlations may indicate a miscoding of the items, resulting in incorrect calculations of the measures if these items are retained. When miscoding is ruled out as a cause then the item's construct validity should be questioned. This process of eliminating

misfitting items is more complex when an individual choice can elicit evidence of more than one way of thinking.

The choices from each of the 28 test items were coded with the particular misconceptions that they address—essentially, which of Steinle's (2004a) codes can account for a learner's selection of this choice.

The Rasch analyses are now presented for each of the codes used for this micro-domain and the results are then presented in Table 29.

L1: whole number thinking and decimal point ignored

All 28 test items were analyzed to determine which of these items elicited evidence of L1 thinking. Five of the test items showed a low correlation to the learner responses and were unsuitable as indicators. One of these test items, Item 10028, had a correlation value of -0.1, which indicated a lack of correlation between the learners' pattern of responses and the calculated learner measures. This item asked the learner to select the smaller of 0.05 and 0.9, for which the misconception and the correct choices were the same and this item had no use as a predictor of this behaviour. However, the Item 10027, which asked for the smaller of 0.09 and 0.5, had an excellent correlation of 0.56. Thus these two items had significantly different value for diagnostic purposes even though they appeared to be similar in form

Four other items produced low correlations and all had exactly three decimal places in each of their choices, such as Item 10035 to find the smaller of 2.414, and 3.001. From this small sample it appears that for the L1 misconception to be elicited more effectively, the decimal fractions for the choices should not have the same number of decimal places. However, this argument does not hold for Item 10022, asking for the smallest of 2.39 and 2.40, which showed good correlation to the learner measures but which was different from other items in having a trailing zero in one of the choices. The five test items with poor correlation were removed from the data set, and the analysis was repeated.

Following the Rasch calculation, Item 10034 showed a high OUTFIT value and was also removed and the calculation again repeated, producing an acceptable fit of the test items to the learners' propensity to use the L1 misconception.

L2: column overflow

For this way of thinking, 15 items were removed due to the lack of correlation and fit to the data, leaving 13 test items which were good indicators, but with varying degrees of discrimination between learners who used the L2 misconception and learners who did not.

In total, 17 learners had high measures on the L2 code including learners B57 and B85 who both selected 12 out of 13 choices and for whom this provided evidence of a highly consistent usage of the L2 way of thinking.

L3: reverse thinking

The L3 way of thinking was exemplified by 0.79 being read as 97 with the "ths" dropped, and thus 9 hundreds and 7 tens.

Learners B10 and B40 scored 14/14 on this analysis, meaning that all 14 of their responses were accounted for by this way of thinking. In total, five learners achieved calculated measures on L3 greater than 1.5.

S3: reciprocal thinking

For S3 thinking, the decimal point was seen as being equivalent to the common fraction symbol so that 3.4 was seen as the same as $\frac{3}{4}$.

22 learners had responses which fit this model, and 9 of these learners showed a high usage of this misconception to account for their measures. These included learners B12 and B14 with 20/22 responses accountable to this way of thinking.

Identifying Learners by Stage - all Responses

Parallel Rasch analyses were conducted for each of the four codes. Assumptions were made on how the coding should address leading and trailing zeroes, which are not explicitly included in Steinle's (2004a) model, and also on whether to ignore or consider differences in the whole number part of the choices.

These Rasch analyses have produced sets of learner measures for each of these codes, where each measure is an indicator of the propensity of a learner to select a choice based upon a particular way of thinking. These analyses also produced item measures, where a larger value indicates that less learners who used this misconception actually

selected the coded choice from these items, whereas lower item measures were selected by more learners, and thus were considered as better to meet the requirements for RQ2.

The calculated learner measures are summarized in Table 29, with learners in the STABLE stage shaded in grey. These learners were not analyzed for the coded misconceptions.

In the L1, L2, L3 and S3 columns, measures greater than 1.50 are highlighted with a light orange background, and with a light yellow background for measures between 1.00 and 1.50. These are indicators of learners' use of these misconceptions to select the choices. In the ABILITY column, the IMMINENT stage learners are shaded with a light orange background when the measure is in the range 1.00-2.00.

Learners were positioned in the ACTIVE stage when they were not classified as STABLE or IMMINENT but had a measure of 1.50 or more on at least one of the coded misconceptions. Learners were positioned in the EMERGENT stage if they had no evidence of misconceptions from these measures, but had an ABILITY measure of at least 0.50, and are identified by light-green color in the ABILITY column. The remaining learners were positioned in the ABSENT stage and are indicated by the absence of shading.

Table 29. Decimal Ordering: Results of Rasch analysis

Learner	ABILITY	L1	L2	L3	S3
A01	1.42	-2.08	-1.59	-2.45	1.18
A02	0.45	0.64	1.58	0.32	-1.21
A05	1.42	-1.70	-1.09	-1.97	1.18
A06	0.29	-0.21	-0.21	0.11	-0.26
A07	4.86				
A08	-0.51	0.87	1.58	1.02	-1.49
A10	0.80	-1.39	-2.18	-1.29	0.91
A14	1.67	-1.70	-0.21	-1.97	0.91
A18	0.45	0.64	2.18	0.32	-0.71
A26	2.82				
A27	-0.03	0.64	0.21	0.55	-1.21
A28	0.62	0.64	2.18	0.11	-0.95
A30	0.62	-0.64	-0.21	-0.32	0.19
A31	3.59				
A32	4.86				
A42	-0.03	0.21	0.64	0.11	-0.48
A43	2.33				

Learner	ABILITY	L1	L2	L3	S3
A44	1.67	-1.70	-1.09	-1.97	0.19
A46	0.13	-0.21	-1.09	-0.32	-0.48
A50	0.99	-0.21	-	-0.54	-0.48
A66	-0.34	0.87	1.09	0.78	-1.21
A67	1.42	-0.87	-0.64	-1.60	0.42
A68	0.29	-1.39	-2.18	-0.77	0.91
A77	2.33				
A78	0.13	-0.87	-1.59	-1.02	0.91
A79	0.44	-4.04	-3.85	-3.82	2.90
A80	-0.34	0.42	1.09	0.32	-0.71
A82	1.97	-0.21	0.64	-0.54	-0.48
A86	1.19	-1.12	-1.09	-1.29	0.42
B01	3.59				
B02	-0.51	-0.87	-0.21	-0.77	0.91
B03	4.86				
B04	0.29	-0.87	-1.59	-0.54	-0.03
B05	1.97	-0.42	-0.21	-0.77	-0.48
B07	0.29	0.42	-0.21	0.55	-0.95
B08	-0.27	0.63	-0.57	0.58	-1.10
B09	-0.11	2.77	3.58	1.93	-2.95
B10	-1.04	2.72	0.97	3.95	-2.87
B11	-0.51	1.70	0.64	1.02	-1.81
B12	0.45	-4.59	-3.06	-3.22	2.75
B13	1.42	-0.21	0.64	-1.02	-0.26
B14	0.99	-3.34	-2.18	-3.22	2.75
B16	-0.67	1.12	0.64	0.78	-1.21
B17	0.99	-3.34	-2.18	-4.48	2.22
B19	1.04	0.02	0.27	-0.16	0.14
B20	0.99	-3.34	-2.18	-3.22	2.75
B21	1.67	-0.42	0.64	-1.02	-0.26
B23	0.29	0.42	2.18	0.32	-0.71
B24	1.19	-2.08	-2.18	-3.22	1.81
B25	0.62	-0.87	0.21	-0.77	0.42
B26	2.52				
B27	2.33				
B28	0.45	-0.64	-0.21	-0.77	0.42
B29	0.13	1.39	2.18	1.60	-2.69
B30	4.86				
B32	0.31	-0.71	-1.22	-0.58	0.23
B33	3.59				

Learner	ABILITY	L1	L2	L3	S3
B34	-0.51	0.98	0.17	1.35	-2.06
B35	1.42	-1.39	-1.09	-1.97	0.66
B36	0.21	1.39	2.02	0.58	-1.53
B37	1.15	-0.85	-1.57	-1.29	0.42
B38	-0.19	0.87	1.58	0.78	-1.21
B39	1.22	-1.04	-	-1.18	0.51
B40	-0.76	4.00	2.02	3.95	-4.16
B41	0.45	1.39	2.02	1.35	-2.06
B42	0.08	-3.75	-2.03	-3.42	4.24
B43	0.31	1.06	1.09	0.71	-1.79
B44	4.63				
B45	1.14	1.10	-	1.36	-2.13
B46	-0.27	-1.71	-3.69	-1.08	1.18
B47	1.04	-1.68	-0.74	-0.87	0.14
B48	2.30				
B49	2.00				
B51	1.60	-1.30	-0.57	-2.03	0.52
B52	-0.19	1.12	0.64	0.55	-1.49
B53	-0.34	-0.74	-0.76	-0.86	0.91
B54	-0.03	1.39	1.58	0.78	-2.19
B55	0.45	0.64	0.21	0.32	-0.95
B56	1.97	-2.08	-1.59	-2.45	1.18
B57	-0.03	2.56	3.06	1.29	-2.69
B58	-0.19	0.64	0.21	0.78	-1.21
B59	1.67	-2.08	-1.59	-2.45	1.18
B60	3.59				
B61	0.45	-3.34	-4.38	-1.97	2.22
B62	0.90	0.99	-	1.36	-1.96
B63	0.99	-0.87	0.21	-0.77	-0.71
B64	-0.51	1.39	1.09	1.29	-1.81
B65	0.80	-1.70	-1.59	-1.60	0.91
B66	-0.19	-1.70	-3.06	-1.02	1.47
B67	0.29	0.64	0.64	0.32	-1.81
B68	-1.43	3.02	3.36	1.17	-1.86
B69	0.45	-1.71	-2.32	-1.50	1.18
B70	-1.02	0.42	0.21	0.32	-0.71
B71	0.62	0.00	1.09	0.11	-0.48
B72	0.62	-1.12	-1.59	-1.02	0.91
B76	0.62	0.87	1.09	0.55	-1.49
B85	-0.19	2.56	3.06	1.60	-2.19

A visual analysis of Table 29 yields the following observations:

- There were 24 learners in the ACTIVE stage—who had insufficient ability measures to be STABLE, and who had at least one misconception measure of 1.50
- There remained 37/97 learners for whom no conclusion could be reached in terms of their responses, being learners with no scores highlighted with shading. These 37 learners had measures which did not show proficiency and also did not show use of one or more of the individual misconceptions. These learners may have been using misconceptions but these cannot be determined validly using this approach within the scope of the current set of misconceptions and the current item bank. These were positioned in the ABSENT or EMERGENT stages with EMERGENT learners differentiated by having an ability measure of at least 0.50. This splits off 10 learners as EMERGENT, leaving 27 in the ABSENT stage.
- There were no L1 or L3 measures greater than 1.50 with the learners in School A, and only one case of S3 which was greater than 1.50.
- For most learners who showed evidence of a misconception, there was overlap between the measures, so that a high score in one measure often indicated a high score in another, and this was a by-product of the structure of this set of items.
- For each of the codes, there were a number of learners who measured higher than 1.00:
    - L1 : 16 learners
    - L2 : 21 learners
    - L3 : 13 learners
    - S3 : 15 learners

A total of 115 errors were recorded for the 26 learners who were classified as STABLE or IMMINENT, with these being from all but one of the 28 test items. This item, for which all of these top 26 learners scored 100%, was Item 10035, asking for the smaller of the decimal numbers 2.414 and 3.001, in which all learners correctly selected 2.414.

However, 11 of these 26 highest proficiency learners selected 0.79 as being smaller than 0.6 in Item 10025. Two other test items received a relatively high frequency of incorrect choices among these groups of 26 top learners, being Item 10030, asking for the smallest of 0.09 and 0.500, for which 7 learners selected 0.500, and Item 10045, asking for the smallest of 10.0 and 10.125, with 6 learners selecting 10.125. Thus these three items, 10025, 10030, and 10045, are candidates to be considered as suited to detect late-stage misconceptions.

Identifying Learners by Stage – Error Responses only

This analysis differs from the previous analysis in that only errors were analyzed so that the correct responses were removed before the analysis. This analysis was only conducted for this micro-domain and not for others, due to its relatively large number of documented misconceptions, and the challenge that many item choices could have been selected from both stable conceptions as well as from misconceptions, thus complicating the identification of the most likely conceptual base of the learner.

The results are summarized in Table 30, which is presented in the same structure as the previous analysis. This table shows, for each learner, the measures which indicate to what extent the learner was likely to have used a particular misconception in responding to the test item. In some cases, there were overlaps in the misconceptions used, with many misconceptions being possible explanations for a learner response.

This table only shows results for misconception measures in the cells where these are larger than 1.50, and as for the previous analysis shows the full sample of learners, with their assessed ability as measured from the CORRECT choices. As a result, these are not shaded with a background colour like Table 29.

The results are shown in the sequence of high-to-low learner proficiency, as was calculated in the previous analysis, and which remains unchanged. The STABLE learners were included as part of the analysis of the misconceptions for this variation of the analysis, due to the need to have as much data as possible on the errors. This shows high measures for some codes, such as learner A26 for codes L1 and L3, which offer alternative explanations for their errors.

Table 30. Decimal Ordering: Learner measures

Learner	CORRECT	L1	L2	L3	S3
A32	4.86				

Learner	CORRECT	L1	L2	L3	S3
B03	4.86				
B30	4.86				
A07	4.86				
B44	4.63				
B60	3.59				
B01	3.59			1.81	
B33	3.59				
A31	3.59				
A26	2.82	2.08		2.48	
B26	2.52				
B27	2.33				
A77	2.33				
A43	2.33				
B48	2.30				
B49	2.00	2.05		2.44	
B05	1.97				
B56	1.97				2.30
A82	1.97	2.08		2.48	
B21	1.67	2.65			
B59	1.67				2.30
A44	1.67				
A14	1.67		2.15		2.93
B51	1.60	2.05		2.44	
A67	1.42	1.51			
A05	1.42				3.20
A01	1.42				
B35	1.42				2.35
B13	1.42	2.85			
B39	1.22				
B24	1.19				2.69
A86	1.19			1.88	
B37	1.15				2.30
B45	1.14				
B19	1.04				
B47	1.04				
B14	0.99				3.59
B63	0.99				
B20	0.99				3.59
A50	0.99	2.05		2.44	
B17	0.99				3.69

Learner	CORRECT	L1	L2	L3	S3
B62	0.90				
B65	0.80				
A10	0.80				3.58
A30	0.62				3.05
A28	0.62	2.79		3.07	
B25	0.62				1.51
B76	0.62	2.61		3.07	
B71	0.62	2.61		3.07	
B72	0.62				2.10
B12	0.45				3.20
B55	0.45				
B61	0.45				2.10
A18	0.45		2.37	3.07	
B41	0.45	2.61		3.07	
A02	0.45	1.91		3.07	
B28	0.45			3.05	
B69	0.45				2.07
A79	0.44				
B32	0.31			1.88	
B43	0.31	2.08		2.48	
A06	0.29	2.61		3.07	
B67	0.29	2.38		2.48	
B07	0.29	2.61		3.07	
B04	0.29				2.34
A68	0.29				2.10
B23	0.29	3.01		1.81	
B36	0.21				
A78	0.13				1.75
A46	0.13				
B29	0.13	2.61		3.07	
B42	0.08		2.15		2.93
A27	-0.03	2.04		3.07	
B54	-0.03	2.04			
A42	-0.03	1.91			
B57	-0.03	3.55		3.07	
B09	-0.11	2.79		3.07	
B38	-0.19	2.60		3.05	
B66	-0.19				2.10
B85	-0.19	3.55		3.07	
B52	-0.19				

Learner	CORRECT	L1	L2	L3	S3
B58	-0.19	2.61		3.07	
B08	-0.27				
B46	-0.27				2.07
A66	-0.34	2.79		3.07	
A80	-0.34				2.68
B53	-0.34				4.11
B11	-0.51	2.04			
B02	-0.51	2.03			
B34	-0.51				
B64	-0.51	2.04		3.07	
A08	-0.51	3.46		3.05	
B16	-0.67	2.79		3.07	
B40	-0.76	2.79		3.07	
B70	-1.02	2.04			
B10	-1.04	2.79		3.07	
B68	-1.43	2.65			

This table is read as follows: the ABILITY column shows the measured ability of the learner, using all of the responses; the L1, L2, L3 and S3 columns shows the likely reasons for mistakes made by the learners, using only error responses, even if the correct responses is also accountable to the same way of thinking.

In this table, learners A14 and B11 are shaded to provide ease of identification for the following explanations.

Learner A14 scored 23/28 correct with a measure of 1.67 and also scored a high measure of 2.93 on the S3 misconception. This is interpreted that for two test items for which the learner made a mistake and which are also linked to the S3 misconception, there is evidence that this learner used this misconception as the basis for selecting these incorrect responses. The additional explanation is that they may have used the S3 way of thinking to account for some of their correct responses. However, 23/28 is around an 82% pass mark, which appears to be very high by CTT standards of measurement, and yet the L2 and S3 misconceptions have far higher measures which makes them more likely explanations for this learner's responses.

Learner B11, who was at the lower end of the proficiency scale, received an estimated measure of -0.51 on the basis of 11/28, and a measure of 2.04 on the basis of a

score 4/5 on the L1 misconception test. From this it can be inferred that this learner used the L1 misconception to account for 4/5 of the test items which measure L1.

There were only three learners who obtained a measure > 1.50 for L2 when using the incorrect responses only, without also including the correct responses.

This analysis was performed on a relatively small number of items and thus its reliability is questionable, given that some of these results may also be attributed to guessing. Thus this approach requires more data to arrive at conclusive inferences.

Answering the Research Questions

RQ1 (EFFECTIVENESS)

For L1, 22 items had good fit to the learner responses, but these were not all equal in their value. Item 10022 had the lowest measure (-1.64), and thus has the potential to catch the largest number of learners. Items 10036, 10045, and 10025 also had low measures and are also candidates for being good diagnostic items.

However, most of these items had only two choices for learner selection of the smallest or largest, and there was an overlap concerning which way of thinking accounts for the learner responses. The best test items are those for which no other explanation was possible—which follows the guidelines on the "semi-dense" items of Bart et al. (1994), In this case, Items 10023 and 10038 are likely to be the best diagnostic items, since for both of these items the selection of the L1 option does not have any other explanation in terms of the misconception being measured, so that L1 is the only possible conceptual explanation after considering guessing and slips.

The L2 code did not have individual items for which there was a single reason why a learner would choose the L2 choice over another. However, a suitable set of items can provide evidence that L2 thinking was used consistently. Of the items which provide a good fit to the Rasch model, those which were answered by the majority of learners who held this misconception are Items 10026 (-2.93), 10042 (-1.59), 10039 (-1.42), and 10048 (-1.41). These four items are likely to be the best set to elicit evidence of this misconception.

The L3 code was identified as the conception used by a number of learners, with two learners, B10 and B40, scoring 14/14 on this code. This means that these learners used this misconception, and only this misconception, in answering the items. However,

there were a few items which provided consistent results across the sample of learners, and those which are more likely to elicit this evidence are Items 10036, 10025, and 10045.

Finally, Items 10047, 10031, and 10040 had the lowest measures for S3, and thus will catch the largest number of learners using this misconception. However, Items 10037 and 10044 also had low scores and should be considered as potential diagnostic indicators of this misconception.

RQ2 (EFFICIENCY)

RQ2 asks how many diagnostic test items are needed to provide sufficient evidence of misconceptions used by the learners. The focus is not on test items which provide evidence of ability and proficiency, but rather those which expose particular misconceptions and ways of thinking. This is complicated by the large number of misconceptions associated with decimal number ordering and the need to distinguish between these misconceptions.

The assumption is made that ideal items do exist—for which the individual choices each point to a specific misconception with no overlap between the choices and the conceptions which they elicit. For such ideal items the choice selected by a learner would then point directly and exactly to the misconception which accounts for the response. This approach only fails in the case of the ABSENT stage learners, who were likely to be guessing and thus did not have systematic patterns of responses.

For each of the codes L1, L2, L3, and S3, there were some test items that can be used to elicit evidence of the usage of the misconceptions, expressed in decreasing levels of suitability (best first):

- L1 : Items 10022, 10036, 10045, 10025
- L2 : Items 10026, 10042, 10039, 10048
- L3 : Items 10036, 10025, 10045
- S3 : Items 10047, 10031, 10040, 10037, 10044

There was little consistency between the items identified for the individual ways of thinking, and thus RQ2 was a challenge to answer unless all of these items are used. My conclusion is that test items with only two choices are not suited for efficient diagnostic usage, since a large number will be required to be used to identify the conceptions being used to select the responses. An alternative is to use items with more than two choices which can elicit more types of misconceptions. The use of items with

only two choices is also aggravated by the number of misconceptions which are included for this micro-domain. If there was only a single misconception being assessed then there would be fewer alternatives for explanations of the learners' conceptual base for their responses than if there were four separate misconceptions, as in this micro-domain. One recommendation for classroom practice is that misconceptions are addressed by the teacher individually so that the diagnostic assessments can be more targeted.

RQ3 (SELF-KNOWLEDGE)

Item 10021 asked for the smaller of (1) 2.36 and (2) 2.4 and 40 incorrect responses were given for choice (2), which was coded with both L1 and L3. Of these 40 incorrect responses, only four learners identified this test item as Difficult and the others identified this as either Easy or Just Right.

Similar findings can be made for each of the other test items in this micro-domain, and it appears that this self-knowledge is primarily useful for identifying those learners who identified items as Difficult, and who could then be placed into the ABSENT stage of development, since the Easy and Just Right learners are already positioned into the Development Stages quantitatively using the Rasch analysis alone.

Summary

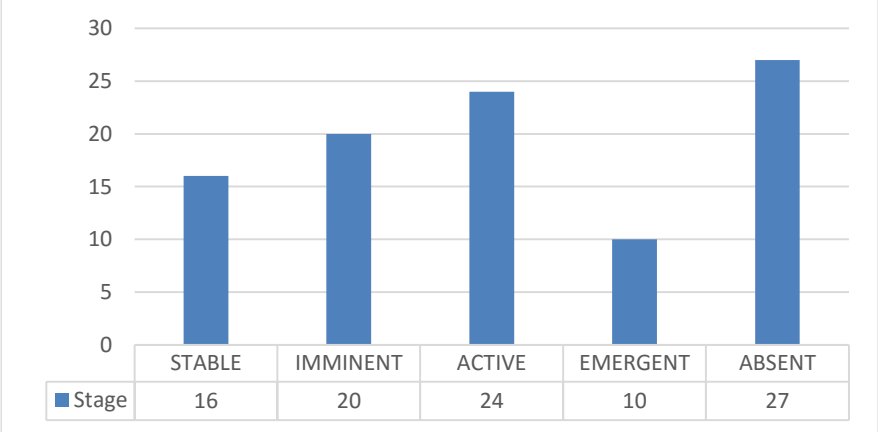


Figure 30. Decimal Ordering : Learners by Stage

The learners were quite evenly distributed over the five stages, as shown in Figure 30. This distribution involved setting a number of cut-points for both the ability measure and for the individual coded misconceptions. These cut-points were set on the basis of a visual inspection of the values rather than using a more formalized approach, and by manipulating cut-points the learners were positioned differently. These stages indicated

relative learner development in a micro-domain rather than being an exact measure, and future research is needed to improve the nature of this positioning into these Development Stages. I contend that knowing the learner Development Stages is useful for the teacher, and I suggest that a single vocabulary of stages will assist learners in better understanding these trajectories.

Test items with only two choices were problematic for assessing proficiency, since pure guessing resulted in an expected score of 50%. These test items elicited evidence of learners' understanding of decimal magnitudes using problems which involved the ordering of two numbers. These have proven utility for rational number research, but their usage for efficient classroom application should be questioned. However, the misconceptions used in this micro-domain may be overlooked when using traditional assessments, unless these are specifically targeted.

Of the original 97 learners, 16 were identified as being in the STABLE stage due to their proficiency over a range of the items. These fell into the A1 code.

Of the remaining 81 learners, 37 learners could not be classified using the codes representing the constructs being measured when the correct items were included using the existing set of test items. However, more learners could be classified, under less valid conditions, by omitting the correct responses from the analysis and analyzing only the error responses. This second analysis provided additional information and the number of learners whose error responses could not be classified was reduced to 10.

6.6 Micro-Domain CR - Common Fraction Representation

Initial Analysis of Responses

This micro-domain consisted of ten items, each of which asked for the best fractional representation for a given verbal statement. A total of 88 learners provided responses to one or more of the items.

The number of responses were counted for the ten test items and for each choice, to both determine the general level of fraction knowledge among the learners, and also to identify patterns of error responses.

Table 31. Common Fraction Representation: Counts by test item/response

Item#	1	2	3	4	5
10060	6	7	71	4	
10061	2	6	4	73	2
10062	1	3	76	2	5
10063	76	7	1		3
10064		8		76	2
10065	7	3	2	73	
10066		76	6	3	
10067	79	5		1	
10068	6	23	6	50	
10069	41	35	2	6	

Table 31 shows that Items 10060 to 10067 each had a single dominant choice which was selected by the majority of the learners, and these were the correct choices for each item. These items had a relatively small number of errors, with some representing known misconceptions. Items 10068 and 10069 had two possible correct answers, due to the ambiguity in the wording that was introduced into the test items.

For Item 10064, 8/86 learners selected 9.12 for "nine twelfths" in item 10064, representing a misunderstanding between common fractions and decimal numbers—which I refer to as the DECIMAL misconception, and which is the same as Steinle's (2004a) S3 code used for the decimal number ordering. Two learners selected $\frac{12}{9}$ in which the numerator and denominator were reversed—which I have called the REVERSAL misconception.

For Item 10068 learners were asked to select the closest representation of "three hundred and sixteen fortieths" as a common fraction, with 23 learners selecting $\frac{316}{40}$ and another 50 learners selecting $300\frac{16}{40}$. Both are correct in terms of the different readings of the fraction description, as outlined at the start of this section. This is a limitation of our ability to express fractions in words rather than a limitation of the learners' knowledge.

For Item 10069 the highest response came from a choice which I had predicted to be the most likely to be selected. This was similar to the situation for Item 10068 in which the wording of the fractions was itself ambiguous. The primary difference was that the word "and" was omitted from Item 10068, and I had predicted that the wording "sixty

one sixty thirds" would cause learners to select $\frac{61}{63}$ but only 35 made this choice against 41 selecting $61\frac{60}{3}$. I note that in this case, with the word "and" omitted, more learners selected the mixed fraction, as opposed to Item 10067. Another 6 learners selected the other reading of this "(sixty) (one sixty) (thirds)", being $60\frac{160}{3}$.

These initial observations challenged the validity of using questions involving word fractions in a diagnostic test, since it appears to be the limitations of the wording, rather than the learners' conceptual knowledge, which influenced the selection of the options.

Two misconceptions were identified in Chapter 4, being DECIMAL, when the decimal representation choice is selected in preference to the common fraction (for example selecting 1.4 for "one quarter"), and RECIPROCAL, when the numerator and denominator are interchanged. These accounted for the majority of the errors, and no other errors were coded for this micro-domain.

Identifying STABLE Learners

A total of 47/88 achieved 100% success on these test items, of which 45 learners achieved 10/10, with one learner each achieving 5/5 and 6/6 respectively.

Of the remaining learners, 21 made a single mistake, of which 20 achieved a score of 9/10 and one obtained a score of 3/4. Of these 21 learners most of the responses were accountable for by the DECIMAL and RECIPROCAL misconceptions and there were two cases where the learners selected other choices, of which one, selecting 1/25 for "one quarter" was a surprising result from a learner who otherwise obtained a clean sheet. This left 19 errors to be explained, which was then reduced by another 10 for the errors in the Items 10068 and 10069 which were not true errors due to their ambiguous wording. Of the remaining errors, five consisted of selecting $5\frac{9}{6}$ for "five and six ninths", which showed a high consistency among the errors caused by the learners who made only a single mistake, and which is a variation of the RECIPROCAL misconception. Whereas slips were likely to be random in nature, a systematic error used by a number of learners may have been an indication of a way of thinking which was not identified in advance or uncovered from the patterns of responses. The discovery of these new ways of thinking is beyond the scope of this study and these results point to future studies.

Analyzing the DECIMAL Misconception

All of the items except for 10067 and 10069 provided a choice which highlighted the DECIMAL misconception. These items were highly correlated to the responses received from the learners, and two items had a 100% correlation, being Items 10061 and 10065. This means that these items were perfectly aligned with the results of learner behaviour, and thus are perfect diagnostic instruments. In other words, those learners who showed a propensity to use the DECIMAL misconception have always selected these choices, and those who did not have this propensity did not select these choices. All of the remaining eight test items in this set had good correlation and also acceptable fit statistics.

There were five learners with measures of 1.5 and above for the DECIMAL misconception and another seven learners who showed a strong propensity to use this misconception. Learner B49, for example, selected 6/6 choices on the basis of this misconception.

When there are fewer items for analysis, as for this misconception, then Rasch analysis may be an overkill for practical usage, but it can offer good support when selecting which of the items are better for diagnosis.

Analyzing the RECIPROCAL Misconceptions

Only 2/10 of the test items were useful for the analysis of the RECIPROCAL case. These were Item 10069, which had an issue with the wording as outlined previously, but which has proven useful for eliciting this misconception, as well as Item 10065. No other items which had the RECIPROCAL misconception as a choice received sufficient responses for analysis.

To perform this analysis all error responses were included, but none of the correct responses, since there were no cases in which a choice was both correct and RECRIPROCAL at the same time. Given that the number of error responses in this micro-domain was small, there was very little data available to determine with sufficient validity that this misconception was the cause of the errors.

Learner B16 selected 3/5 choices which were accountable to the RECIPROCAL misconception and learner B66 selected 4/9 of his/her errors on the same basis, and for both of these learners there was a strong indication of RECIPROCAL thinking. The total number of items counted included all test items for which the learner responded in error,

including responses which were not an indicator for this misconception. However, this did not impact on the results, which have shown that only two of the items were valid for the determination of this misconception.

I conclude that RECIPROCAL thinking has value and warrants attention for future studies. The results raise another concern, that a particular way of thinking may only be used by a small number of learners in a class, and that the consensus of the class proficiency may drive the class forwards if these misconceptions are not detected and dealt with appropriately.

Answering the Research Questions

RQ1 (EFFECTIVENESS)

The DECIMAL misconception was selected by 8/10 learners in Item 10064, and by 6/9 learners with Item 10066. These items also correlated with learner measures and were responsible for many of the errors.

For the RECIPROCAL misconception, Item 10065 provided better evidence then others with 7/12 responses being directly attributed to this way of thinking.

RQ2 (EFFICIENCY)

RQ2 cannot be answered for this micro-domain, since there was insufficient data to split up the error responses into smaller subsets. The question is whether a single item may suffice, if this could cause the learner to select the rich distractor rather than the correct response.

RQ3 (SELF-KNOWLEDGE)

For RQ3 there were a total of 213 error responses from the 88 learners involved in the tests for this micro-domain. Of these, learners only provided 14 responses where an item was marked as Difficult, and these were spread over the entire set of test items. The result was that almost all of the learners considered these test items as either Easy or Just Right for their level of proficiency, and thus there was no way to distinguish between the learners who were proficient from those that thought that they were proficient but in reality were not.

As a result, learner difficulty appears to have little or no value for diagnostic purposes on the basis of these misconceptions and the data available.

Summary

A large number of the learners achieved success, but the errors encountered were insufficient to enable the positioning of learners into Development Stages. More research is needed to identify whether and to what extent, this micro-domain is useful for diagnostic purposes. Thus I do not present a summary of the distribution of the learners over the stages as I have done for the previous micro-domains.

However, a few learners used the DECIMAL and the RECIPROCAL misconceptions as identified in advance of the analysis, and even if this helps a few learners it could be beneficial.

6.7 Micro-Domain NL - Number Line for Common Fractions

Initial Analysis of Responses

This micro-domain consists of a range of problems which use the number line to position common fractions. For example, Item 10071, as shown in Figure 31, has a stem with a number line on range 0-1 which has 7 ticks on this range, with a red arrow pointing to the 4th tick. The learner was asked to select the common fraction represented by the point on the number line where the red arrow is pointing.

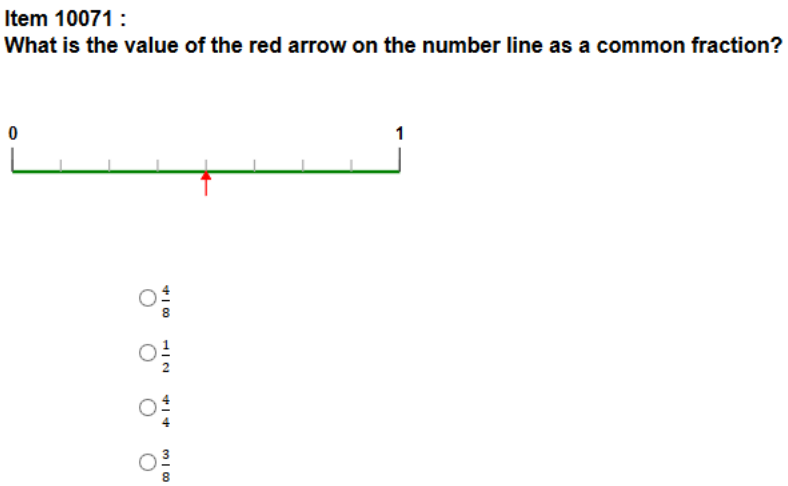


Figure 31. Number line example test item

Three misconceptions were identified in Chapter 4 concerning the number line. The TICKSPARTS misconception concerns confusion between the number of ticks and

the number of parts. The WHOLELINE misconception is where a learner treats the entire number line as a range 0-1 irrespective of how many whole numbers on the line. The DECIMAL misconception occurs when a learner selects a decimal number instead of the common fraction requested. There were few choices which exhibited these individual misconceptions and thus a simpler approach was used for analysis— positioning the learners into the Development Stages in a single pass based on the frequency of responses observed. Thus the Rasch method was not used, and the analysis was based on visual inspection, which may reflect human expertise in seeing patterns in data that machines are unable to identify, and this approach may hold promise for a future AI-based approach to data analysis.

Table 32 shows the number of learners who selected the different choices for the items in this micro-domain. Correct responses are shown with a blue background and both Items 10071 and 10078 have two correct responses.

The test items can be divided into two groups. Firstly, those with little evidence of systematic errors—where there was a dominant correct response and the remaining responses divided among the choices with no other choices. This group consists of the Items 10071, 10073, 10074, 10075, and 10079.

Table 32. Number Line: Counts by test items/response

Item#	1	2	3	4
10070	7	26	42	4
10071	41	25	10	3
10072	4	26	3	47
10073	4	61	6	7
10074	7		1	71
10075	65	7	5	1
10076		5	36	37
10077	2	37	35	2
10078	9	44	25	
10079	64	8	5	1

The second group had a distinct pattern of systematic errors, in which one of the incorrect responses predominated, with a frequency of 26 or more, and these are indicated on the table with a red background. Whereas there were a number of choices which received up to 10 responses, there were no choices which had between 11 and 25

responses. Thus there was a clear distinction between these dominant choices and others. However, the pattern of the other choices was not random, such as Item 10074 for which 7/8 of the errors were attributable to choice 1. These other possible misconceptions were not included in this study, and should be the subject of future studies.

For Item 10070, the number line had five ticks, which were over the range 0-1, with the arrow pointing to the 5th tick. Thus each tick represented the value "one-sixth" and the correct response was $\frac{5}{6}$. However, 26/79 of the learners selected the common fraction $\frac{1}{5}$. One explanation for this high frequency is that the arrow points to the 5th tick, with the digit 5 being consistent between this tick and the choice. This is a variation of the TICKSPARTS misconception.

For Item 10072, one of the choices was the decimal equivalent of the correct answer, in this case 1.5 being provided as a choice in addition to the common fraction $1\frac{1}{2}$. However, the response 1.5 was incorrect for this test item, given that a common fraction was requested. 26/80 learners selected this decimal number response, which was almost identical to the results from Item 10070, of which only 10 learners were common to both items. This is an example of both the DECIMAL and the WHOLELINE misconception.

For Item 10076, there were 4 ticks between each of the whole numbers, 0-1-2-3-4-5, which split each whole number range into 5 parts. The arrow points at the midpoint between the whole number labels 4 and 5, and is positioned midway between the 2nd and 3rd ticks. The correct response of $4\frac{1}{2}$ was selected by 37 learners, with 36 learners selecting $4\frac{2}{5}$ which was close but which was not correct. This was evidence of a misconception in the mathematical reality of the learners which was not considered in advance, in which the learners treated the arrow as pointing at $\frac{2}{5}$ even though it was pointing to a space after this tick. This could have been due to the learners believing that the diagram was wrong and that the arrow was misplaced. This should be explored as a potential misconception in the future, with reference to prior work on the empty number line. Five learners selected the second choice ($5\frac{1}{2}$), which was an error reflecting the process of counting back from the right of the number line rather than from the left. As a contrast, 65 learners selected the correct choice of $2\frac{3}{4}$ in Item 10075, with almost no

learners selecting $2\frac{1}{2}$, where the number line was over the range 0-5 with only the whole numbers marked, and with the arrow pointing to an empty place between the 2 and the 3, but with no guiding tick marks as in Item 10076. Thus the tick marks in Item 10076 may have influenced the learners in their choice over the lack of tick marks in Item 10075. Since this error occurred only on this single item, I suggest that this is included into a future study on learner misconceptions between number lines with ticks and empty number lines in terms of their suitability for diagnosis.

The TICKSPARTS misconception was evident in Item 10077 in which more learners selected choice 2 ($1\frac{5}{11}$) than the correct choice 3 ($1\frac{5}{12}$). The learners likely selected choice 2 ($1\frac{5}{11}$) based upon counting the number of ticks, there being 11 ticks. This misconception was also evident in Item 10078, in which 44 learners selected choice 2 ($2\frac{2}{5}$) as opposed to a mere 25 learners selecting the correct choice 3 ($2\frac{2}{6}$) together with a smaller number selecting the other correct choice 1 ($2\frac{1}{3}$).

Identifying Learners by Stage

For this micro-domain there were 81 learners and 10 test items. Using an ability cutoff at 1.50 there were 28 learners positioned in the STABLE development stage which included those learners having between 8 and 10 correct responses.

These STABLE stage learners collectively made 40 errors, and some of these were common to this group of learners:

- Item 10076: 7 learners selected $4\frac{2}{5}$ rather than $4\frac{1}{2}$ as the answer given that there were two ticks prior to the arrow on the diagram.
- Item 10077: 7 learners selected $1\frac{5}{11}$ using the number of ticks as the denominator rather than the correct $1\frac{5}{12}$ which was the correct choice, using the number of divisions rather than the number of ticks.
- Items 10078: 10 learners selected the incorrect choice with the same reasoning as for Item 10077.

These three cases accounted for 24/40 errors made by the STABLE learners, and this suggests that those who made these errors should be downgraded to the IMMINENT stage since they had lingering misconceptions which may have been hidden by the high

scores that these learners obtained. It is thus my conclusion that these were not slips but were misconceptions.

The ACTIVE stage learners were those when there was a predominant usage of one of more of the misconceptions, and the remaining learners were then positioned in the ABSENT or EMERGENT stages.

Given that the majority of the test items highlighted the TICKSPARTS misconception, this was the only misconception taken further for analysis.

Answering the Research Questions

RQ1 (EFFECTIVENESS)

RQ1 is answered by the success in identifying the TICKSPARTS misconception from both of the items 10077 and 10078. These can be used in conjunction with the following:

- Item 10070 exposed a misconception in the understanding of the fractions $\frac{1}{5}$ vs $\frac{5}{6}$ in terms of the parts and the number line.
- Item 10072 offered the learners the alternative of selecting a decimal number even though the common fraction was requested.
- Item 10076 provided the situation in which the arrow was not positioned on a tick and this helped to identify how learners responded.

The structure of these test items is thus that each test item exposed a different misconception, and the responses from five of the ten items did not provide evidence of systematic errors and misconceptions.

RQ2 (EFFICIENCY)

For RQ2, a single item is likely sufficient to obtain a quick response. For example, to detect the TICKSPARTS misconception either Item 10077 or 10078 alone is sufficient to provide evidence of the misconception.

RQ3 (SELF-KNOWLEDGE)

For RQ3, there were 783 responses in total covering all test items and all learners, with 360 of these indicated as Easy, 228 as Just Right, and 160 responses as Difficult. This frequency of items identified Difficult was far larger than all the other micro-domains in this study. 13 learners identified more than 7 of the items as Difficult with 4

identifying all of the 10 items as Difficult. Learner B34 identified all 10 items as Difficult, and yet scored 6/10, but most of those who identified a high number of items as Difficult did not achieve success, and thus these learners were potential candidates for being in the ABSENT stage. Thus, for this micro-domain there is some potential for using the learner-indicated difficulty as an aid to detection of misconceptions .

Overall, the results of this micro-domain provided potential for discovery of misconceptions, and this should be explored through follow-on studies.

Summary

Whereas some learners were positioned into STABLE and IMMINENT Development Stages, there was insufficient data available to position the remaining learners into the other Development Stages, and thus the summary of the Development Stages is not presented for this micro-domain.

6.8 Micro-Domain CG - Common Fraction Graphics

Initial Analysis of Responses

Table 33 shows the count of the responses over the 12 items in this micro-domain. A blue background denotes the correct responses, identifying multiple correct responses where applicable. A red background highlights incorrect responses with high frequencies which point to possible misconceptions. This micro-domain was not analyzed in advance, in Chapter 4, for any previously identified misconceptions, rather the approach adopted was to analyze these results to find potential misconceptions.

Table 33. Fraction Diagram: Counts by test item/response

Item#	1	2	3	4
10080	1	75	2	
10081	14		62	1
10082	29	14	22	7
10083	50	6	10	5
10084	24	2	7	39
10085	69		1	2
10086	13		59	
10087	1	15	54	2
10088	2	7	57	5
10089	1	52	11	7
10090	1	44	21	5
10091	4	55	1	11

Five of the 12 items show a pattern of systematic errors. Each of these was analyzed qualitatively to understand the possible misconceptions which may have given rise to the learners' responses. The response times were used for this analysis, which were not used for any other previous analyses.

Item 10081 is a 10x10 grid of squares, with the bottom right 6x6 being blue and the remaining squares being red. The learners had to count the squares to obtain the answer, and the average time taken by the learners to answer this item was only slightly longer than Item 10080 which had significantly less squares to count. The challenge is to explain why 14/76 of the learners selected the incorrect choice $\frac{64}{100}$, which is the fraction of the red squares rather than that of the blue squares, and also since the majority of these learners did not make a similar error Item 10080. One possible explanation is that the learners did not read the question and followed the example of Item 10080 which asked for the red squares and then simply applied this to the next item 10081 without reading the stem question properly. However, there is evidence from the responses to Item 10080 of general proficiency in understanding the relationship between the diagram and common fraction notations. This explanation for the high incidence of learners choosing $\frac{64}{100}$ is perhaps not a misconception, but rather more operational, in a rush to answer the question. This is more likely due to a lack of experience in how to read a question before

answering. Even though this explanation is not conceptual in nature, it may be commonly experienced as learner errors, and thus also requires diagnosis and remediation.

Item 10082 was the most difficult of the test items in this micro-domain, requiring both reasoning as well as common fraction knowledge. This test item was derived from the TIMSS 2003 study, item M012001 (Mullis et al., 2004). The highest count of responses was for "1", indicating that only one more circle was needed to be yellow to make up $\frac{4}{5}$ of the total. This response was likely derived from the three yellow circles in the diagram and the target fraction of $\frac{4}{5}$, so that adding "1" would seem like an appropriate response. This is an example of whole number arithmetic in which the denominator was ignored in the calculation.

Item 10083 asked for the best representation of the blue squares—which should have been labelled as circles, but no learner reported this error, nor did any of those in my pilot group—where one diagonal consists of 6 yellow circles, with the rest of the 6x6 circles being blue. There were 10/71 of the learners who selected the choice $\frac{1}{6}$, which is the fraction of yellow circles and not the blue circles. This could have been caused by the learner remembering the previous question, since the colour yellow was the target of that question. This is similar to the issue that arose in Item 10081, further justifying that this notion of "item memory" is possibly a common cause of errors and potentially worthy of further examination.

Item 10087 had a high number of learners (15/72) who selected $2\frac{1}{4}$ as the answer rather than the correct $2\frac{3}{4}$, but there was insufficient evidence to uncover the reason for this. Future research into such fraction diagram representations would be needed to understand the nature of this misconception.

For item 10091, 11 learners selected $3\frac{2}{3}$ as the response, which was a similar error to the selection of $2\frac{2}{6}$ in Item 10088, since they both had more whole units than existed in the diagrams, and also that the wrong color was chosen for the fractional part, being the white segments in Item 10088 and the yellow segments in Item 10091.

Answering the Research Questions

Due to the small number of items within this set, it was not possible to answer the research questions in full, but the analysis of the responses has pointed to future directions for this research.

RQ1 (EFFECTIVENESS)

For RQ1, I have not found prior diagnostic studies that used fraction diagrams in a diagnostic setting, and also no studies that have dealt with misconceptions within such fraction diagrams. As a result, this part of my study is providing initial results which can point to further studies.

RQ2 (EFFICIENCY)

This is not covered for this micro-domain due to limited data available.

RQ3 (SELF-KNOWLEDGE)

There were a total of 870 responses to these 12 test items, with 521 marked as Easy, 267 as Just Right, and 49 as Difficult. However, only 9 of the 93 responses identified as systematic errors were indicated as Difficult. There is thus potential to learn more about this relationship between learner self-reflection and the possible misconceptions associated with fractions represented as diagrams.

Summary

No summary is provided here for how the learners are positioned into the micro-domains. Further research will be needed to identify specific misconceptions and the nature of the best types of problems which are suited for these as diagnostic instruments.

6.9 Micro-Domain CO - Common Fraction Ordering

Initial Analysis of Responses

The full set of 20 items in this micro-domain were presented to the learners in School A, but only 10 of these items were presented to the learners in School B. Whereas Rasch analysis can handle such differences in the number of items, there may be some impact on the items for which there is less data. There were 93 learners in total, with 29 learners from School A, and 64 learners from School B, with 1143 responses recorded.

The 20 items were categorized in advance in terms of an estimated difficulty level based on the nature of the denominators. These are presented in Table 34, which combines the estimated difficulty with the measured item difficulty from the Rasch analysis. This also includes the average duration in seconds that learners spent answering these items, which is extracted separately for both correct and incorrect responses.

The analysis of the response time was not performed for the other micro-domains, and this was also the only micro-domain for which there was an estimated difficulty level.

Table 34. Common Fraction: Test items by difficulty and duration

Item#	Measured Difficulty	Estimated Difficulty	Duration All	Duration Correct	Duration Incorrect
10092	-1.47	1	33.01	31.99	37.63
10102	-1.35	1	30.02	25.97	46.71
10105	-0.70	3	19.44	21.21	15.12
10096	-0.61	4	24.56	27.27	18.75
10104	-0.61	3	26.60	22.28	35.24
10107	-0.45	5	31.46	31.67	30.86
10095	-0.11	3	16.34	18.32	13.74
10097	-0.09	5	22.31	25.67	14.75
10106	-0.07	4	24.66	28.18	20.00
10094	-0.05	3	21.78	21.14	22.59
10100	0.12	5	22.62	22.18	23.44
10108	0.17	5	24.18	23.33	25.70
10093	0.29	2	29.15	33.02	25.44
10101	0.31	5	26.21	26.28	26.14
10099	0.33	5	20.46	16.06	27.50
10098	0.33	5	18.69	19.19	17.90
10103	0.55	4	30.00	31.25	28.33
10111	0.74	5	21.32	20.80	21.92
10110	0.74	5	21.89	25.07	18.23
10109	1.95	4	23.86	29.89	21.00

This table is presented in the sequence of the calculated Rasch difficulty of the items. This Rasch difficulty is an indication of how many of the learners answered the items correctly, so that an item with a low measure was easier to answer correctly than an item with a high measure.

Some observations from this table are:

- Whereas Item 10092 was identified as the easiest, it also had the longest duration. This was the first question presented to the learners in the online assessment, so the learners may have taken the extra time to become familiar with test operations. For the future I suggest that learners are given an initial trial question to help them to become familiar with the types of questions before the test commences.
- The most difficult was Item 10109 and was answered correctly by only 9/28 learners. The correct responses had a far longer duration (29.89s) than the incorrect responses (21.00s).
- Item 10102 was anomalous considering that it was both one of the easiest, and also had the longest duration for incorrect responses, of 47 seconds. In analyzing this in further detail, there were 36 responses of 10 seconds or less, of which only 6 were incorrect, and the longest response measured was 277 seconds which may have impacted the average response time. To obtain a more accurate result, outliers in the response times could be removed. However, for this study, I have included all response times, given that such outliers occurred throughout the set of items and that my concern is not with accuracy but rather with these duration measures as indications of learner activity.
- For many items there was a significant difference between the durations of learners who selected the correct option from those who selected the incorrect option. Items 10102, 10104 and 10099 had the largest differences in the durations of correct responses over the incorrect responses, and likewise Items 10106, 10096, 10109 and 10097 had large differences in incorrect over correct response durations.
- For example, Item 10104 was answered correctly by 57/83 learners, but incorrect responses took 13 seconds longer on average than correct responses.

There is thus some indication that the duration taken by the learners on some items may be a predictor of whether they will succeed or fail, but 12/20 items had relatively similar durations for success and failure, in which the absolute difference was less than 7 seconds on responses which were typically between 20-30 seconds for other items.

Identifying STABLE Learners

The results are presented in "map" form in Table 35 using TABLE 1.0 of WinSteps that shows both the learners and the items on the same scaled diagram. I have not used this map form on previous micro-domains since the previous results benefit from the additional information in the form of the tables. However, for this micro-domain there is no additional information to display, and thus this "map" form is preferred, being shorter and more concise.

This "map" is a qualitative picture of the data, illustrating the goodness-of-fit of the data to the Rasch model. The left panel of the map shows the measured ability of learners from the highest at the top (learner B30), to the lowest at the bottom (learner B62), and compares this to the equivalent measures of item difficulty on the right side from the most difficult at the top (Item 10109), down to the easiest at the bottom (Item 10092). The prefix '10' is dropped from the item numbers to enable these to fit better onto the diagram, so Item 10110 is represented by the label 110. This map diagram shows a bell curve structure for both the learners and the items, with a concentration around the median point, as indicated by the "M" on the central axis, with the learners and items then spreading out to the top and bottom.

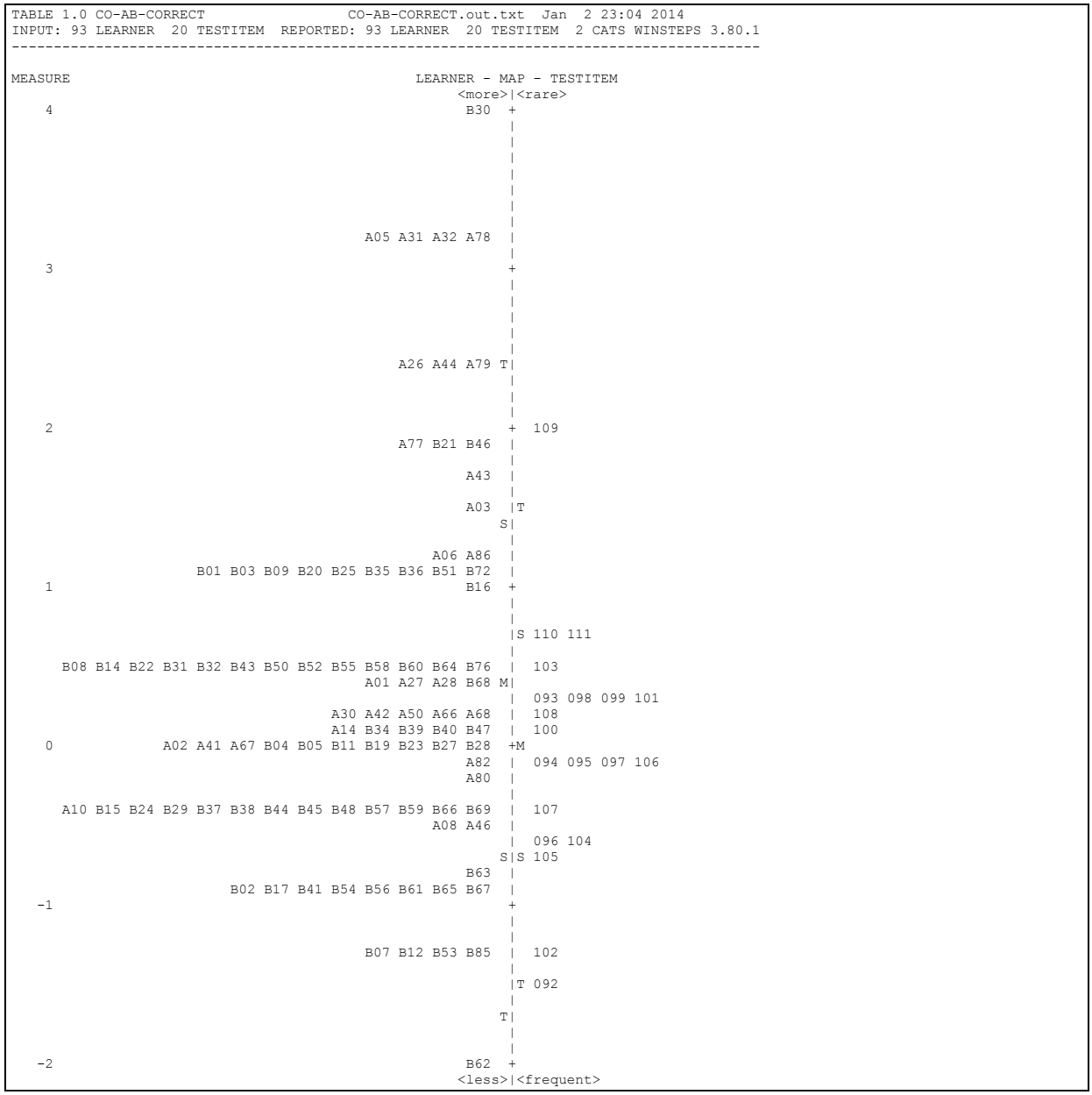
Learner B62 only responded to one test item, and thus cannot be considered as the true worst performer.

The learners around the median measure each scored around 9-10/20 or 5-6/10, which means that they selected an incorrect option for around 50% of the test items. However, given that each of the test items offered only two choices, getting 50% correct could also be achieved by guessing.

The process to determine the learners in the STABLE stage was the same as in previous micro-domains, using a cut-point of 1.5. This selected 12 learners in the STABLE stage, being learners B30 down to A43 on the map.

Below this STABLE group are 13 learners A03 to B16 who had measures of ability from 1.50 down to 1.00 and each obtained high absolute scores such as 18/20, 8/10 and 4/5. Based upon this Rasch measure, these learners were not positioned in the STABLE stage, and need further analysis on their errors.

Table 35. Common Fraction Comparison: Map of learners and items



From this point I continue my analysis with the entire set of 93 learners for the identification of the NUMERATOR and DENOMINATOR misconceptions and am not separating out the STABLE stage learners as I have done previously so that I can explore an alternative approach with fewer steps and with more data.

Analyzing the NUMERATOR Misconception

This misconception is based upon the following rule to find the smallest common fraction:

NUMERATOR rule: Select the smallest numerator, and if the numerators are equal then select the smallest denominator. Similar for the largest.

This rule was applied using a Rasch analysis, and the results of the top few learners are shown Table 36:

Table 36. Common Fraction Ordering: NUMERATOR misconception

```
TABLE 17.1 CO-AB-NUMERATOR            CO-AB-NUMERATOR.out.txt  Jan  3 13:56 2014
INPUT: 93 LEARNER  20 TESTITEM  REPORTED: 93 LEARNER  20 TESTITEM  2 CATS WINSTEPS 3.80.1
-------------------------------------------------------------------------------------------
LEARNER: REAL SEP.: 1.20  REL.: .59 ... TESTITEM: REAL SEP.: 2.20  REL.: .83

         LEARNER STATISTICS:  MEASURE ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|        |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| LEARNER|
|------------------------------------+----------+----------+-----------+-----------+--------|
|    10     19     20    4.51    1.85|      MAXIMUM MEASURE|  .75   .75|100.0 100.0| A27    |
|    11     19     20    4.51    1.85|      MAXIMUM MEASURE|  .75   .75|100.0 100.0| A28    |
|    41     10     10    3.87    1.87|      MAXIMUM MEASURE|  .00   .00|100.0 100.0| B15    |
|    53     10     10    3.87    1.87|      MAXIMUM MEASURE|  .00   .00|100.0 100.0| B29    |
|    67     10     10    3.87    1.87|      MAXIMUM MEASURE|  .00   .00|100.0 100.0| B45    |
|    62      5      5    2.87    1.90|      MAXIMUM MEASURE|  .00   .00|100.0 100.0| B39    |
|    63      5      5    2.87    1.90|      MAXIMUM MEASURE|  .00   .00|100.0 100.0| B40    |
|    38      9     10    2.53    1.09|1.12   .4| .92   .4|  .15   .24| 90.0  89.9| B11    |
|    51      9     10    2.53    1.09|1.12   .4| .92   .4|  .15   .24| 90.0  89.9| B27    |
|    75      9     10    2.53    1.09|1.28   .6|2.39  1.2| -.19   .24| 90.0  89.9| B54    |
|    77      9     10    2.53    1.09|1.29   .6|2.66  1.3| -.22   .24| 90.0  89.9| B56    |
|    86      9     10    2.53    1.09|1.29   .6|2.66  1.3| -.22   .24| 90.0  89.9| B65    |
…
```

As for the case of the Decimal Number Ordering micro-domain, there are some items for which the choices do not conclusively point to a single way of thinking, such as being both correct and also the outcome of applying the NUMERATOR rule. A learner's response to a single item cannot be used to infer which rule the learner was using when they responded. However, inference is improved when using a number of items in a test to identify consistent responses.

Learner A27 scored 19/20 on this NUMERATOR rule, but only scored 12/20 against the correct responses, and consequently their responses are better accounted for by the NUMERATOR rule. Similarly, learner B45 scored 5/10 correct, which is the expected outcome from pure guessing, compared to their score of 10/10 when scored against the NUMERATOR rule.

A cutoff measure of 2.50 resulted in 12 learners who showed evidence of using the NUMERATOR misconception consistently, and included those with raw scores down to 9/10.

Analyzing the DENOMINATOR Misconception

This misconception is structured on the basis of the denominators, where smaller denominators represent larger numbers.

> **DENOMINATOR rule: To find the smallest number select the largest denominator, not considering the numerators. When the denominators are equal then select the smallest numerator.**

Table 37. Common Fraction Ordering: DENOMINATOR misconception results

```
TABLE 17.1 CO-AB-DENOMINATOR        CO-AB-DENOMINATOR.out.txt  Jan  3 13:56 2014
INPUT: 93 LEARNER  20 TESTITEM  REPORTED: 93 LEARNER  20 TESTITEM  2 CATS WINSTEPS 3.80.1
-------------------------------------------------------------------------------------
LEARNER: REAL SEP.: .99  REL.: .50 ... TESTITEM: REAL SEP.: 2.36  REL.: .85

         LEARNER STATISTICS:  MEASURE ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT   |  OUTFIT  |PTMEASURE-A|EXACT MATCH|        |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| LEARNER|
|------------------------------------+----------+----------+-----------+-----------+--------|
|    22     20     20    4.49    1.85|      MAXIMUM MEASURE|  .00   .00|100.0 100.0| A67    |
|    55     10     10    3.49    1.88|      MAXIMUM MEASURE|  .00   .00|100.0 100.0| B31    |
|    28      9     10    2.58    1.09|1.12    .4|1.12    .6|  .13   .20| 88.9  88.9| A82    |
|     1     18     20    2.45     .77|1.02    .2| .77    .1|  .21   .20| 89.5  89.4| A01    |
|    72      9     10    2.16    1.09| .82    .0| .42   -.1|  .46   .23| 90.0  90.1| B51    |
|    21     17     20    1.95     .65|1.24    .7|1.47    .8|  .06   .25| 84.2  84.2| A66    |
|    27     17     20    1.95     .65|1.06    .3|1.06    .3|  .19   .25| 84.2  84.2| A80    |
|     8      8     10    1.42     .83|1.19    .6|2.80   1.9| -.12   .31| 80.0  79.9| A14    |
|    84      4      5    1.40    1.17|1.24    .6|1.13    .5|  .07   .28| 80.0  80.1| B63    |
|    49      8     10    1.28     .84|1.63   1.3|3.91   2.4| -.61   .31| 80.0  80.1| B24    |
|    56      8     10    1.28     .84| .92    .0| .71   -.1|  .42   .31| 80.0  80.1| B32    |
|    68      8     10    1.28     .84| .81   -.3| .56   -.4|  .54   .31| 80.0  80.1| B46    |
|    70      8     10    1.28     .84|1.45   1.0|3.19   2.0| -.34   .31| 80.0  80.1| B48    |
…
```

Table 37shows the top 13 learners where the MEASURE column is the learner measures computed using the DENOMINATOR rule. Learners A67 and B31 each achieved a 100% score on this measure, and A67 also scored 10/20 on the correct choices, and B31 scored 7/10. Learner A01 scored 18/20 using this DENOMINATOR rule, and also scored 12/20 on the correct choice. Thus these 3 learners, A67, B31, and A01, have achieved 50-70% correct, which is an acceptable pass mark in the classroom, and yet the most plausible way of thinking to account for their results is the DENOMINATOR rule.

I set the cut-point at 1.20 resulting in 14 learners who showed evidence of using the DENOMINATOR rule to answer these test items. These were combined with the 12 learners selected as using the NUMERATOR rule to comprise 26 learners for whom there

was evidence of misconception usage and who were positioned within the ACTIVE development stage.

The benefits of using a Rasch model for this data analysis of misconception usage is to validate the responses in terms of a strong statistical model. The other benefit is that these can be scored easily, since the Rasch calculations require automated methods and cannot be calculated by hand.

The vast majority of these items had a good fit to the response data, and consequently the data fits the measures as obtained from the learner responses. There would not have been such a good fit if the constructs being measured were not real and thus this analysis not only provided valid measures of the items and learners, but also provided evidence of these misconceptions being both real and measurable as a reflection of learning thinking on common fractions.

Identifying STABLE, IMMINENT, ACTIVE, and ABSENT Learners

The initial analysis presented in the person-item map in Table 35 shows 25 learners with measures greater than 1.00, of which 12 were positioned in in the STABLE stage with a measure of 1.50 of more. These learners made very few errors, with minimum scores of 9/10 for School A or 19/20 for School B.

The remaining 13 learners from this group of 25 learners made some errors, scoring around 8/10 or 15-18/20. These were classified as IMMINENT stage learners, and an analysis of their errors may help to determine if these are slips or late-stage misconceptions.

From the initial 93 learners, 25 were positioned into the STABLE or IMMINENT stages, with another 26 previously positioned into the ACTIVE stage by reason of their high measures on the NUMERATOR and DENOMINATOR rules. The remaining 42 learners were then in either the ABSENT or EMERGENT stages, however, no rule has been established to separate these and these were retained as a single combined stage.

Answering the Research Questions

RQ1 (EFFECTIVENESS)

For the NUMERATOR rule, the Items *10093, 10094, *10101, *10103, 10105, 10106, and 10107 meet the requirement that the correct choice and the rich distractor are different, implying a clear inference from a learner response. From this qualitative

analysis, the three items identified with an * are best in terms of the quantitative evidence that more of the learners who have used this rule also selected these choices.

For the DENOMINATOR rule, the best items are *10095, 10096, 10097, 10098, 10099, 10100, 10104, 10108, *10109, *10110, and *10111, for which each, like the NUMERATOR rule, has a rich distractor choice which has no other plausible explanation within the limits of this model. Those marked with a * show the best measures from the Rasch analysis and thus could be recommended for diagnostic measurement.

RQ2 (EFFICIENCY)

There is a challenge when using only two choices in a diagnostic test that the responses may be subject to guessing which would result in an expected success rate of 50% per item. However, the items marked with an * provide better indicators of the use of these rules, to distinguish these from proficiency learners, and those who were guessing.

RQ3 (SELF-KNOWLEDGE)

There are some learners who marked an item as Difficult and who also selected the incorrect option. The majority of these responses were for Items 10093, 10095, and 10101, making up around 50% of the items marked as Difficult, with the remainder spread over the other 17 items. Item 10093 was the first item presented, and Item 10095 was relatively easy compared to other items in the set, whereas Item 10101 was more difficult in the sense that the fractions in the stem are very close to each another. However, there was insufficient evidence that the Difficulty Index would help to improve the effectiveness of the diagnostic process.

Summary

This analysis has helped to determine whether there was evidence of specific misconceptions in learner responses, formulated as rules that can be tested on the common fractions.

Two misconceptions were formulated and tested, being the NUMERATOR rule and the DENOMINATOR rule. There was strong evidence that these rules form the basis for the responses of 12 learners for the NUMERATOR rule and 14 learners for the DENOMINATOR rule.

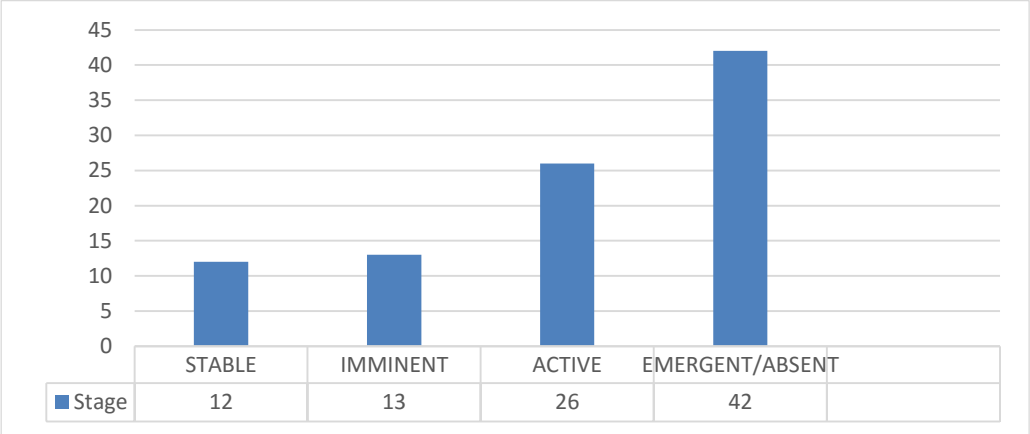


Figure 32. Common Fraction Ordering : Learners by Stage

The distribution of learners for this micro-domain was influenced by the inability to separate the EMERGENT and ABSENT stage learners, but it is clear from Figure 32 there was an overall lack of proficiency among the learners.

6.10 Micro-Domain CE - Common Fraction Estimation

CE1: Initial Analysis of Responses

This micro-domain was structured into two different types of test items, CE1 and CE2, where CE1 items were used for both schools A and B, covering 98 learners with a total of 880 responses, and the CE2 items were used only for school A, with a total of 289 responses from 29 learners.

Items types CE1 and CE2 were analyzed separately for this micro-domain, which was a variation of the standard approach I have adopted.

The results showed a good correlation for each of the items, indicating that each of these items was consistent with the consensus of the others items in determining the learner measures. The INFIT and OUTFIT statistics were within acceptable limits, and thus no items were removed to improve the Rasch modeling.

CE1: Identifying STABLE Learners

A total of 19 learners scored 1.50 or more on the ability measure, where the raw scores were all around 8/10, and these 19 learners were placed into the STABLE stage. These learners collectively made a total of 13 mistakes of which 7 answered Item 10114 incorrectly with 4/7 of these learners choosing 0.12 or 0.5. The choice 0.12 is an indication of the NUMERATOR misconception, which is not expected in learners

achieving high ability scores. This error may provide evidence of late-stage misconceptions which were not explicitly identified by other items in this set.

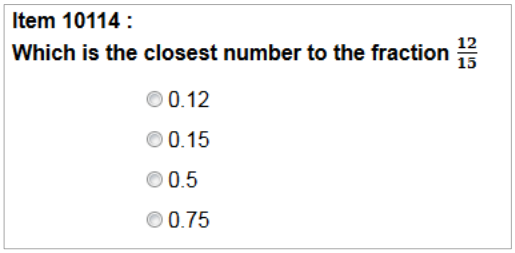


Figure 33. Item 10114

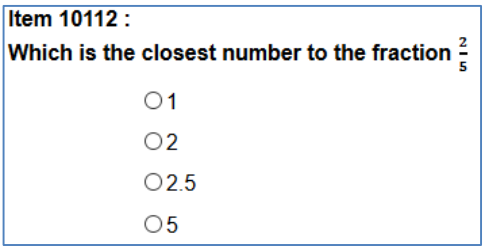


Figure 34. Item 10112

36/92 learners obtained the correct answer on Item 10114, shown in Figure 33, compared to only 28/90 for Item 10112, shown in Figure 34, which was identified by the Rasch calculations as having the highest item difficulty score.

On a visual inspection, Item 10112 appeared to be far simpler than Item 10114, but this observation was not borne out in the results. Item 10112 exposed misconceptions which the most proficient learners did not show evidence of using, except that two of the learners who achieved 9/10 for these item types made their only mistake on Item 10112 with both selecting 2.5, the DECIMAL misconception, as their answer.

CE1: Identifying IMMINENT Learners

Learners with an estimated ability measure of between 1.0 and 1.5, roughly translating to a score of 7/10 were positioned in the IMMINENT stage. This included seven learners who collectively made a total of 23 errors, which were used to identify patterns that indicated a preference for a particular type of error. The first observation is that two items produced no errors at all, being Items 10122 and Item 10126.

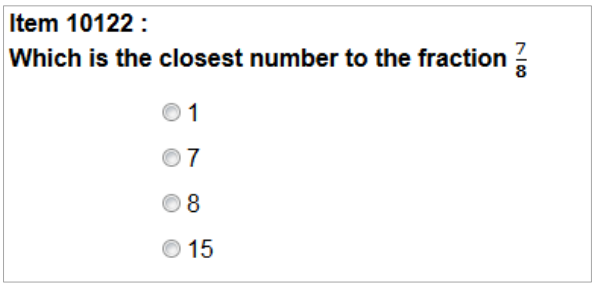


Figure 35. Item 10122

Item 10122, shown in Figure 35, was answered correctly by all of the learners positioned in the STABLE and IMMINENT stages. This item was designed to elicit

responses for the second and third choices, being “7” and “8”, and of the 44 learners who answered this test items incorrectly, 9 responded with “7”, and 22 with “8”.

Thus this item can discriminate between learners in the IMMINENT stage and the other stages, since this appeared to be an early-stage misconception which was more likely to be used by learners in their initial development leading up to the ACTIVE stage.

Item 10112, shown in Figure 34, was answered incorrectly by four of the learners in the IMMINENT stage, and as for the results of the STABLE stage, all of these learners selected the choice “2.5”.

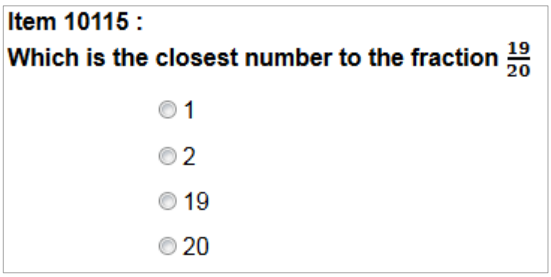


Figure 36. Item 10115

Item 10115 was answered incorrectly by four learners, who selected choices “19” and “20”, which reflected the NUMERATOR and DENOMINATOR misconceptions respectively. This result was predicted by the misconceptions model for this micro-domain, as distinct from the results of Item 10112, for which the highest-frequency response was not predicted. This highlights the challenge in designing the choices for an item, including rich distractors which will elicit misconceptions effectively.

CE1: Identifying ABSENT, EMERGENT and ACTIVE Learners

The learners who were not in the STABLE or IMMINENT stages would have made a number of errors, and it is the extent to which these were accountable to known misconceptions which was important for this analysis.

The ACTIVE stage learners should have achieved around 50% success on the items, with their errors accountable to the known misconceptions. This would indicate active use of some schemas which lead to success, and active schema development to account for their errors. The ACTIVE learners are analyzed in the next section.

The Rasch ability measure has been used to position 19 learners into the STABLE stage and 7 into the IMMINENT stage. It was a general expectation of the Development Stage model that the number of learners in the IMMINENT and EMERGENT stages

would be smaller than the other stages, since these are transition stages between the primary stages of STABLE, ACTIVE and ABSENT.

An additional 39 learners were positioned in the ACTIVE stage, who had ability measures between -0.5 and 1.0. A further 33 learners had measures below -0.5, who needed to be split into those whose responses were showing evidence of some misconception usage (EMERGENT) and those whose responses did not (ABSENT). This requires the identification of the degree to which learners were guessing or were using these known misconceptions.

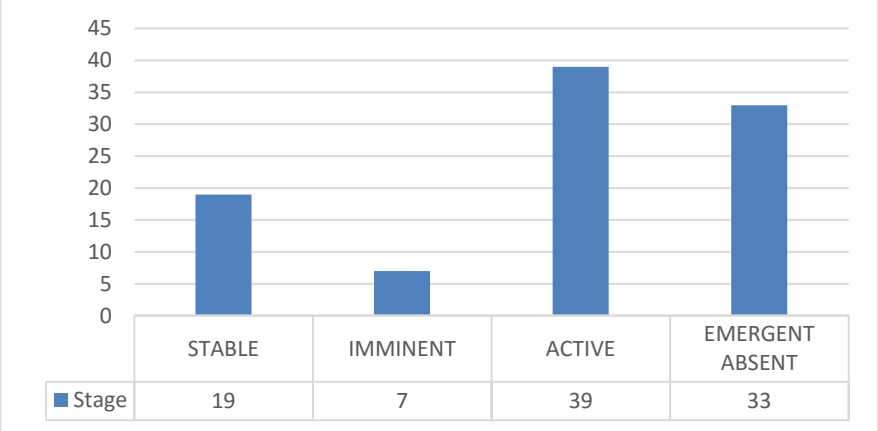


Figure 37. Common Fraction Estimation CE1 : Learners by stage

Learners were positioned into the ABSENT stage when they had an ability measure of less than -0.5, and also showed no indication of having used the NUMERATOR or DENOMINATOR misconceptions. This approach produced one learner, B39, who succeeded on 5/10 with no evidence of pattern in their errors.

In summary, these 71 learners were positioned with 39 in the ACTIVE stage, 32 in the EMERGENT stage, and a single learner in the ABSENT stage.

CE1: Analysis of the misconceptions

The analysis continues by examining the misconceptions themselves and for this three separate Rasch analyses were conducted: for the NUMERATOR misconception; for the DENOMINATOR misconception; and then for the combination of both, for cases where the learner response indicated that either of these misconceptions were used. The analysis commenced with the combined set of misconceptions, and STABLE stage learners were removed to reduce the influence of proficient learners. The correct responses were also

removed, to focus the analysis on the errors, which helped to find an explanation for these errors in terms of either of the two pre-identified misconceptions.

The results showed good correlation and fit, with no items required to be removed for further analysis. However, Item 10126 had a questionable correlation and the analysis was run twice, both including and excluding this item.

The results indicated that all of the items in this analysis were indicators of these misconceptions, but some were better for this purpose. Items 10124 and 10125 yielded no results at all and could not be measured—in other words, no learners selected the rich distractors for these items.

The most unlikely item is Item 10112, for which the rich distractors were selected by only 16/60 learners, and the most likely are Items 10116 and 10123, with 37/41 and 34/39 learners respectively. These two latter items are better as diagnostic indicators for common fraction estimation, since the learners who did not select the correct response were likely to select one of these misconceptions for their response.

Analyzing the items in this micro-domain from the perspective of the misconceptions yielded the following:

- 7 learners predominantly selected the NUMERATOR in their incorrect responses.
- Item 10126 is the best indicator of the NUMERATOR misconception.
- 6 learners predominantly selected the DENOMINATOR misconception in their incorrect responses.
- Items 10122 and 10115 are best indicators of the DENOMINATOR misconception.

CE1: Answering the Research Questions

RQ1 (EFFECTIVENESS)

The NUMERATOR and DENOMINATOR showed insufficient evidence of learner use, so these were analyzed separately. However, the evidence was sufficient when the analyses were combined. The best item to use within the combined analysis are Item 10115 for which the rich distractors were selected by 36/74 learners, and Item 10113, by 39/72 learners. Items 10116 and 10114 also produced good diagnostic results.

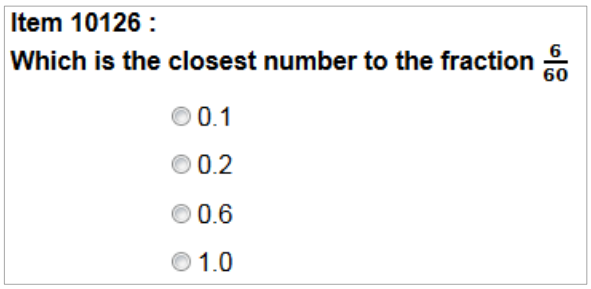


Figure 38. Item 10126

Item 10126 is by far the best item to detect the NUMERATOR misconception with 31/67 of the learners selecting this choice, and the Rasch analysis giving this a measure of -1.39. This low measure means that learners who used this misconception were more likely to select the rich distractor for this item than for any of the other items.

RQ2 (EFFICIENCY)

A single good diagnostic item, such as Item 10115, is sufficient by itself to provide an indication of learners who used either the NUMERATOR or DENOMINATOR misconceptions. Item 10113 can supplement this as required.

RQ3 (SELF-KNOWLEDGE)

A total of 14 learners were identified as almost exclusively using either the NUMERATOR or the DENOMINATOR misconception to account for their incorrect responses, on the basis of the Rasch analyses. Of these the majority also indicated the item difficulty as Easy or Just Right. Thus it is possible to predict that a learner who selected this rich distractor, and who also marked the item as Easy or Just Right, has used this misconception, rather than guessing, since otherwise their Rasch measures would not have indicated this result.

However, there were some exceptions, such as learners B11 and B66 whose responses were split equally between Easy/Just Right and Difficult. For the remainder of the learners no items were marked as Difficult.

Thus for this case, learner self-knowledge is potentially as reliable as the Rasch analysis as an indication of the most prolific usage of misconceptions. This can also be applied very easily without the need to compute the results as required by Rasch analysis.

The analysis of the usage of self-knowledge must also take into account the possibility of guessing on the part of the learners. The learners may not only have guessed the responses to the items, but may also have randomly selected their indication of item

difficulty. Given that there were four choices for each test item, there was a 25% chance of success of selecting a particular misconception, and also given that there may have been two rich distractors within these four choices, there was then a 50% chance of a learner selecting a rich distractor by guessing.

To counter this, I conducted an alternative analysis, using a normalized result in which the learner's indication of difficulty was reduced to account for the possibility of guessing, by reducing the indicated value by 66% to provide a factor on a zero base. This was interpreted as the extent to which the outcomes were directly attributable to some schema usage and not to guessing.

The results showed that there were sets of learners whose responses were more likely to be non-random in nature, and these correlated with the learners whose responses were accountable to these misconceptions from the Rasch analysis. However some learners did not score high on the Rasch analysis and consistently marked the items as Easy or Just Right when selecting these rich distractors. The reason for this was that these learners had selected correct choices or random distractors for other items. However, when these learners did select a rich distractor, they invariably identified these as being Easy or Just Right.

These results point to the potential for using both quantitative and qualitative approaches to measure a learner's use of a misconception. The Rasch analysis provides good evidence when used alone, but there were situations in which the learners may have held these misconceptions and yet this was not indicated in the response pattern. When used in combination with the self-knowledge data there was a greater likelihood of detecting a broader range of learners who were using these misconceptions, and were thus within the ACTIVE or EMERGENT stages of my model.

Whereas the findings from other micro-domains showed no additional value arises from using the self-knowledge of the learners, this is one case in which a positive effect is noted.

CE2: Initial Analysis of Responses

The second analysis for the Common Fraction Estimation micro-domain uses the CE2 test items which asked the learner to find the common fraction which was closest to the given whole number, a decimal number or other common fraction. These items were used

only for school A and were not included within the tests for School B. A total of 29 learners responded to the 10 items in this micro-domain.

Table 38. Common Fraction Estimation: CE2 Counts by test item/response

Item	1	2	3	4
10117	10	18	1	
10118	2	7	11	9
10119	9	6	5	9
10120	7	3	11	8
10121	19	3	3	4
10127	10	2	6	11
10128	13	1	5	10
10129	8	4	13	4
10130	16	11	2	
10131	12	4	4	9

Table 38 provides the frequency analysis of how many learners selected the individual choices for each of the items. This shows a large number of learners who made errors, and the highest frequency errors are highlighted with a red background and were analyzed to identify ways of thinking that may account for these responses.

Item 10128 is shown in Figure 39 and this item was answered correctly by only 5/29 learners. Choice 1 received 13/29 responses, and choice 4 received 10/29 responses, with both having a greater response than the correct choice 3. A qualitative analysis of the choice 1 suggests this may be selected on the basis that $\frac{5}{10}$ consists of the same digits "1" and "5" as found in the item stem "1.5". The fraction $\frac{15}{5}$ in choice 4 also uses these same digits exclusively and perhaps it is a syntactic reading of the choices which influenced the learners in selecting an option, rather than the understanding of which had numeric magnitudes which were close to "1.5".

The result from this frequency analysis, coupled with a similar more detailed analysis of the other items is that these items were beyond the capability of the learners in this Grade 7 class, and that little can be gained from further detailed analysis. As a result, the detailed analysis was kept to a minimum.

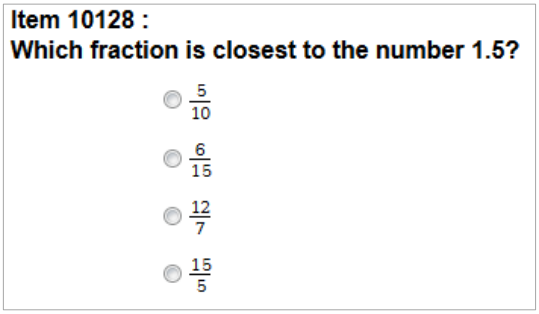


Figure 39. Item 10128

CE2: Identifying STABLE and IMMINENT Learners

The initial Rasch analysis showed that Item 10118 has low correlation and was removed for a second cycle of the Rasch calculations. The results then showed a good correlation and fit and 3 learners were identified with high measured proficiency (7/9 or 8/9 items correct) and were positioned in the STABLE stage.

Another 4 learners were measured at 0.83, achieving 6/9 success and were positioned into the IMMINENT stage. The incorrect responses were concentrated on four items, with Item 10128 being answered incorrectly by four of the learners in the IMMINENT stage, with the other items being Items 10127 and 10131 each with three learners selecting an incorrect option, and Item 10129 with two learners selecting an incorrect option.

CE2: Identifying ABSENT Learners

Nine learners answered no more than two items correctly, of which three learners obtained no correct answers at all. These showed no evident ability, but also showed no indication of using one of the misconceptions identified in this micro-domain.

CE2: Identifying ACTIVE and EMERGENT Learners

The 13 remaining learners obtained some correct answers, and who were analyzed further on their use of the identified misconceptions. However, whereas there were some definite patterns of errors that could contribute to a more detailed analysis, there was insufficient data in this set to warrant this analysis. It is thus recommended that this should be left to a future study in which this type of item can be used within a more extensive quantitative analysis, and which can be coupled with a qualitative analysis of the structure and form of the items, which would ideally be coupled with interviews of the learners to gain an insight into their individual and collective thinking.

The distribution of the learners by stage is summarized in Figure 40

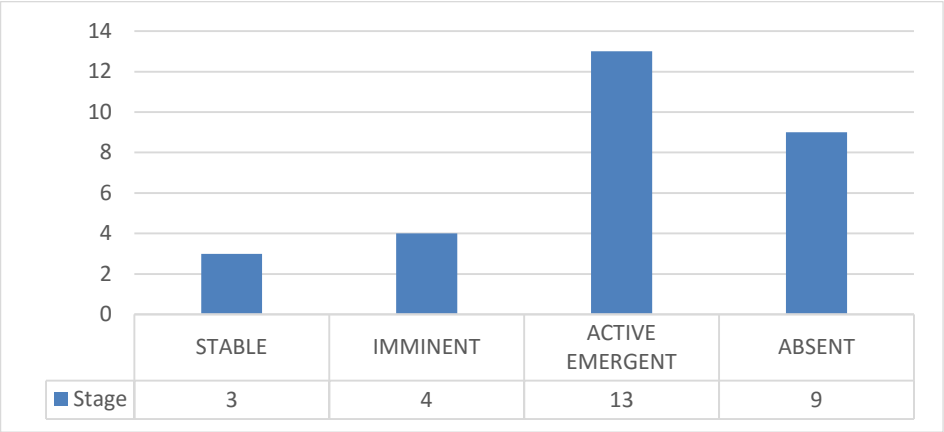


Figure 40. Common Fraction Estimation CE2 : Learners by Stage

CE2: Answering the Research Questions

The research questions were not addressed for the CE2 items due to the limited data available. However, there was clear evidence of systematic selection of errors, which points to misconceptions that may not have been covered by the items in this set.

6.11 Micro-Domain CA - Common Fraction Addition

Initial Analysis of Responses

This micro-domain used 20 items conducted over two tests. Items 10152-10161 were used in Test 1 and Items 10162-10071 for Test 2.

Table 39 shows the count of responses for each of the choices for each of the test items. Those marked with a blue background indicate the correct response and there is only one correct choice for each of the items. Items with a red background indicate a high frequency of responses for incorrect choices which were selected on the basis of having at least 50% of the frequency of the correct choices or for which the frequency was significantly higher than the other incorrect choices.

The distractor choices included both rich distractors identifying common misconceptions and some random distractors with no intended conceptual basis.

For each test, the first five questions were used for both of the participating schools and the second five questions were only used for School A. This accounts for the smaller number of responses for Items 10157-10161 and 10167-10171.

Table 39. Common Fraction Addition: Counts by test item/response

Item#	1	2	3	4
Test 1				
10152	8	3	71	2
10153	33	7	40	3
10154	8	48	21	6
10155	5	11	7	60
10156	38	31	3	11
10157	11	11	3	3
10158		10	15	3
10159	7	11	8	2
10160	6	6	3	13
10161	16	7	4	1
Test 2				
10162	5	2	8	63
10163	6	53	2	17
10164	5	42	17	12
10165	10	6	49	11
10166	15	37	9	15
10167	15	3	5	5
10168	14	4	5	5
10169	3	2	14	9
10170	4	7	2	15
10171	9	14	1	3

Prior studies have shown that the common fraction addition is attempted by adding the numerators and/or denominators—essentially treating the fractions as whole numbers. This is the ADDITION misconception on which this analysis was based and which was used to identify the rich distractors.

Learners selected a significant fraction of the rich distractors for Items 10153, 10156, 10157 and 10158, being where the choice was structured as the sum of either the numerator or the denominator. For instance, Item 10157 included the choice $\frac{11}{15}$ in response to the requested sum $\frac{7}{11} + \frac{4}{15}$. These ADDITION distractors were not provided for all items, and thus were not always available to the learners as a choice. When these rich distractors were not included, then the learners selected other choices.

Item 10154 provided a variation in which the mixed number $1\frac{4}{6}$ was provided as a choice, and this also elicited a high number of learner responses.

In Test 2, Item 10163 requested the sum of $\frac{5}{9} + \frac{4}{9}$ which is the easiest case of common fraction addition since there is a common denominator in place which requires no further work to determine. However, 17/78 learners selected $\frac{9}{18}$, which is the sum of both the numerators and denominators. Items 10169, 10170 and 10171 provided choices that are the sum of numerators, and a significant number of learners selected these choices.

As for previous micro-domains, the proficient learners were initially identified and removed. Following this, the responses from the remaining learners were analyzed to identify their usage of specific misconceptions to account for their responses. The ADDITION misconception is highly relevant, on the basis of the frequency analysis, and this was indicated by rich distractors within Items 10153, 10154, 10156, 10157, 10158, 10163, 10169, 10170, and 10171.

Identifying STABLE Learners

The Rasch analysis showed that many learners had a measured proficiency higher than the measured difficulty of the most difficult item and all of these learners were considered to have total proficiency within the scope and limitations of these 20 test items. This notion of "total proficiency" means that the learners had not made any mistakes, and thus their ability measures could not be accurately assessed, except to infer that within the scope of these tests that they were totally proficient.

A number of learners achieved success on at least 80% of the items they answered, and these learners would be given an "A" grading or perhaps a distinction on a summative test. This analysis resulted in 35 of the learners being identified as proficient and were positioned in the STABLE stage. They were also removed from further analysis. The cut-point measure for those who answered 20 test items was a raw success of around 16/20. Learners succeeding on 15/20 or less were moved to the next step of analysis. These 35 learners were positioned in the STABLE development stage.

Analysis of the ADDITION Misconception

Nine items were identified above as rich distractors for the ADDITION misconception. The responses from these items were analyzed over the non-STABLE learners using a Rasch analysis in which the measures were calculated based on whether the learners selected the ADDITION choice rather than any of the other choices.

Table 40 shows the results of these nine items in terms of their suitability in detecting the ADDITION misconception. The correlations in the PTMEASURE-A CORR column are all within an acceptable range, and some of these, such as Items 10156, 10158, 10163, 10153, and 10154, have high correlations, indicating a strong linkage between these questions—as diagnostic indicators of the ADDITION misconception—with the learner propensity to select these choices.

The INFIT and OUTFIT values were larger than 1.5 for some items, but these remained within an acceptable range, and particularly INFIT, which was the more challenging to address (Linacre, 2002). As a result, all of the items were retained for further analysis.

Item 10158 had a measure of -0.88 and 10/13 learners who answered this item selected the rich distractor for the ADDITION misconception. On this basis this is the most useful discriminator for this misconception.

Table 40. Common Fraction Addition: ADDITION misconception results

TABLE 26.1 CA-AB-ADDITION CA-AB-ADDITION.out.txt Dec 28 14:42 2013
INPUT: 49 LEARNER 9 TESTITEM REPORTED: 47 LEARNER 9 TESTITEM 2 CATS WINSTEPS 3.80.1

LEARNER: REAL SEP.: .00 REL.: .00 ... TESTITEM: REAL SEP.: .00 REL.: .00

TESTITEM STATISTICS: CORRELATION ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASURE-A CORR.	PTMEASURE-A EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	TESTITEM
4	11	16	-.42	.66	1.53	1.6	1.86	1.6	.23	.55	61.5	74.2	10157
8	7	13	.15	.65	1.36	1.6	1.65	2.0	.24	.51	45.5	65.0	10170
7	9	14	.25	.63	1.29	1.2	1.41	1.3	.26	.47	41.7	66.1	10169
9	9	13	.24	.68	.87	-.4	.78	-.5	.56	.45	54.5	67.9	10171
3	27	36	-.25	.50	.82	-.8	.81	-.6	.69	.61	85.7	72.9	10156
5	10	13	-.88	.86	.76	-.4	.61	-.3	.70	.59	90.0	82.3	10158
6	16	32	.81	.49	.92	-.4	.87	-.4	.71	.68	71.4	68.9	10163
1	27	34	-.26	.57	.76	-.9	.66	-.9	.73	.62	82.4	73.9	10153
2	20	30	.35	.55	.82	-.8	.80	-.7	.76	.70	81.3	67.7	10154
MEAN	15.1	22.3	.00	.62	1.01	.1	1.05	.2			68.2	71.0	
S.D.	7.4	9.7	.47	.11	.28	1.0	.44	1.1			17.0	5.1	

However, Item 10158 was relatively difficult compared to the others presented, when analyzed qualitatively, since it required the determination of the lowest common denominator of 9 and 5, being 45. However, 15/28 learners who attempted this test item—

who were all from School A—selected the correct choice $1\frac{1}{45}$, with 10/28 selecting $\frac{6}{14}$ which is the sum of the numerators and denominators.

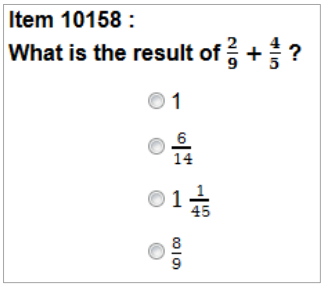


Figure 41. Item 10158

Similar explanations apply to Items 10157, 10153, and 10154.

Identifying ACTIVE and ABSENT Learners

There were 22 learners positioned into the ACTIVE stage, due to having an ADDITION measure of greater than 1.00. Of the original 85 learners, 35 were positioned into the STABLE stage and then another 22 into the ACTIVE stage, and by ignoring the IMMINENT and EMERGENT stages there were 28 learners remaining in the ABSENT stage. The learners were positioned only into the three main stages for this analysis due to the limited number of items which had rich distractors for the ADDITION misconception.

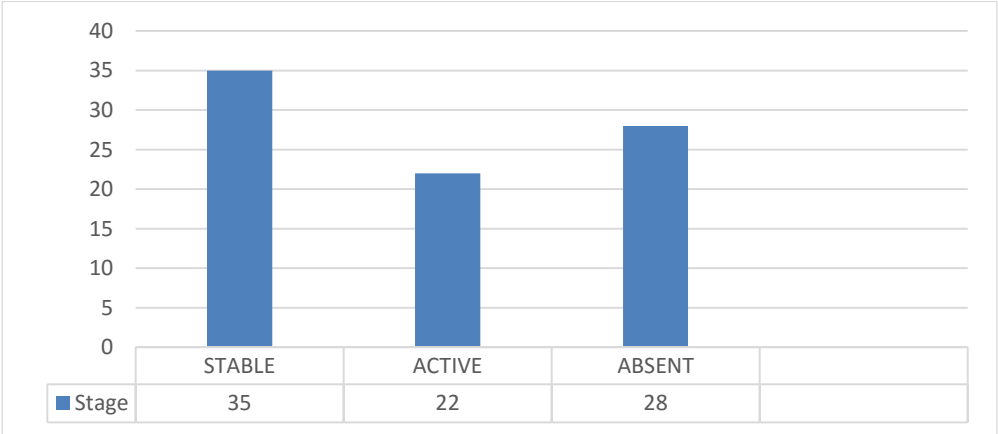


Figure 42. Common Fraction Addition : Learners by Stage

The distribution of learners in Figure 42 still reveals important information about the individual learners, even though this only covers the three primary stages. For example, it shows how many were located in the ABSENT stage and who warranted special attention to commence their path to proficiency in common fraction addition.

However, some learners in the ABSENT stage obtained relatively high marks for this test. For these learners, a lack of conceptual understanding may go unnoticed by the teachers, on the assumption that a good mark is inferred to mean that they know the topic sufficiently to move on in their mathematics studies.

Answering the Research Questions

RQ1 (EFFECTIVENESS)

All nine questions that highlight ADDITION misconception are useful in terms of the Rasch results. Of these, Items 10158 and 10157 were answered using the ADDITION distractor by more learners and thus are more suited as detectors of this misconception.

RQ2 (EFFICIENCY)

Whereas some of the items were used by both of the schools in this study, there was sufficient evidence, from the test items used only by School A, to infer that any of the test items used in this first test would suffice as a good diagnostic item when used by itself. However, within this list, Items 10153 and 10156 are preferable, since they are less complex than other items, and thus can be asked and answered far quicker in a classroom setting.

Thus, my answer to RQ2 is that a single question is likely to be sufficient to provide initial evidence of the usage of this misconception.

RQ3 (SELF-KNOWLEDGE)

There were 139 learner responses for options marked as the ADDITION misconception, and 72 of these were indicated as being Easy, 26 as Just Right, and only 16 as Difficult. Some learners did not answer the question on difficulty, which was optional only for School A, leaving a total of 114 responses. Thus 98/114 of the learners (86%) indicated this as either Easy or Just Right, and yet selected the ADDITION misconception.

6.12 Conclusions

The response data has been analyzed to identify proficient learners, and also to identify misconceptions. The misconceptions resulted from both prior knowledge of specific

misconceptions which had been encoded into the rich distractors, as well as by exploring new misconceptions in micro-domains for which there was little or no prior work. This was conducted within the limits of the size and duration of the tests, and the nature of the item bank.

Most of these micro-domains were analyzed using a standard approach, by first performing a high-level analysis of results, looking for any evidence of systematic errors, and then by matching the error response patterns to the misconceptions which were identified for each micro-domain.

STABLE stage learners were then identified and removed, using a Rasch analysis on the correct choices. Learners who had not reached the STABLE stage but whose ability measures were sufficiently high were identified as IMMINENT and their errors were analyzed to find evidence of common late-stage misconceptions.

All learners other than those in the STABLE stage were analyzed in terms of their propensity to select one or more of the rich distractors, or to select other distractors which identified potential misconceptions uncovered during this study. On this basis, they were positioned into the ACTIVE, EMERGENT and ABSENT stages, with the added information, not presented in the results of this study, that each learner was identified in terms of the specific misconceptions which they were using and which accounted for their positioning into a stage.

Variations to this standard analysis were accommodated to suit the specific nuances for the individual micro-domains, and especially where there was insufficient data to position the learners into the stages. In most of the micro-domains there was an indication of other ways of thinking which were found within the patterns of learner responses but which were not accounted for within the limited test items used, and recommendations are made for future studies to fill these gaps in knowledge.

The next chapter consolidates the findings from this study in terms of the three research questions which this study addresses, and other findings which have arisen from the analysis of the results.

CHAPTER 7 :
RESEARCH RESULTS AND FINDINGS

This chapter summarizes the results obtained from this study and consolidates the findings across the range of micro-domains as reported in Chapter 6.

7.1 Discussion of Results by Research Question

This study began with a vision concerning the potential of the Web to provide help to teachers in mathematics classes. The goal was to both advance the knowledge and practice of diagnostic assessment for mathematics and to identify the potential for systemic introduction of automated tools for diagnostic assessment into classrooms, to coincide with the increased usage of computer technology in secondary education.

My initial analysis led me to treat this as a research problem, to address a gap in how diagnostic assessment is both understood and is used, and to recommend how diagnostic processes could be used in a classroom setting. I selected the rational numbers as my domain of study with a focus on misconceptions in the decimal numbers and the common fractions.

My three research questions were posed in Chapter 1 which are labelled as RQ1 (EFFECTIVENESS), RQ2 (EFFICIENCY), and RQ3 (SELF-KNOWLEDGE). Collectively, these questions address the challenge of obtaining high quality, diagnostic information about learners, where "high-quality" means that the diagnostic information is fit-for-purpose and is valid in terms of what it is measuring. These measurements include diagnostic information about the learners' state of knowledge of mathematics and particularly the conceptual obstacles to developing proficiency in the rational numbers.

In Chapter 1, I motivated the need to answer these three questions as a required step towards providing valid web-based diagnostic tools for mathematics classrooms.

These research questions have guided this project from the start, and have informed the review of prior work; the planning and execution of the initial paper-based pretest; and the content and structure of the online assessments. These online assessments were used to gather the bulk of the data from learners, and were conducted over eight micro-domains in the rational numbers.

The detailed data analysis was conducted over each of these eight micro-domains and was reported in Chapter 6. This analysis is now followed, in this chapter, by a reflection back to these research questions, both individually and across each of the micro-domains. Common threads are identified from the detailed analyses, leading to a number of findings, and to recommendations which extend beyond the scope of these questions. The Development Stage model forms the basis for the analyses, and is assessed in terms of its potential utility, by reviewing its successes and the areas which can be improved. This model is central to this study, and constitutes the primary theoretical vehicle used to both identify conceptual development stages and to position learners into a trajectory based on their stage of development.

RQ1 (EFFECTIVENESS)

The items in this study were designed for the specific purpose to elicit misconceptions. The form and nature of many of these items were drawn from prior studies, and variations of previously reported items were introduced to explore why some items are better than others for diagnostic purposes.

The "semi-dense" criteria for diagnostic items, as identified by Bart et al. (1994), informed the design of some of the items in this study's Item Bank. However, the items do not meet the requirements for semi-density since these criteria are too strict for general application. Rather, semi-density provides a theoretical model to analyze items qualitatively. Many of the items used for this study included choices which were random distractors and which were not based upon any known cognitive rule, thus violating Bart et al.'s "Response Interpretability" criterion. In some cases more than one choice was provided for a given cognitive rule, violating their "Response Discrimination" criterion.

A Rasch analysis was used to measure the diagnostic value of items. This approach is distinct from the qualitative "semi-dense" approach suggested by Bart et al. (1994), yet both approaches have a common purpose of identifying items which are good diagnostic items. The Rasch process was based on learner selection of rich distractors in the MCQ items, each of which was linked to known misconceptions. The items were validated by applying Rasch analyses in parallel, with one analysis for each of the known misconceptions, from which items were identified as better indicators of these misconceptions.

The items were divided into two groups for each misconception, being firstly those suited for diagnosis of the misconception, and secondly those which were not suited. This division was based on the statistical fit of the items with the learners who showed evidence of using this misconception to account for their responses. The Rasch analysis calculated two measures simultaneously, being firstly the construct, as the learners' use of a misconception, and secondly the suitability of the individual items for identifying the misconception. Unsuitable items were identified through the lack of fit to measure this construct, and were removed from the original set of items, leaving a reduced set of items which fit the construct, and were thus suited to measure the learner's use of the misconception. This reduced set of items was used to determine which items are better than others for diagnosing this particular misconception. These "better" items have been described in the section "Item Difficulty Reconceptualized" on page 89 in which these better items are those which are "easier", based on their having low Rasch item measures. "Easier" here means that the rich distractors in these items were selected by a larger group of the learners who used this misconception and which accounted for their responses. These easier items were thus more inclusive to the majority of the learners but being so inclusive they were also subject to an error of measurement in which a learners were indicated as using a misconception when in reality they had not. This was an intentional side-effect and is the lesser of the two evils of measurement error, given the alternative of failing to identify learners who had used this misconception.

The design of the items included the requirement that each choice should not be both a correct response as well as a rich distractor for a specific misconception, however this proved to be unavoidable in some cases. For instance, most of the items in the micro-domain of Decimal Number Ordering asked the learner to select the smaller or larger of two decimal numbers. Given only two choices means that there were a range of ways of thinking which are known to account for learner responses to such items, and thus a single choice may have been selected from more than one way of thinking.

The requirements for good diagnostic items have been discussed in Chapter 4 on page 126, concerning the instrumentation for this study, and are also summarized under the heading of Good Diagnostic Items on page 103. Essentially, my position is that a good diagnostic item must have the primary purpose of identifying a learner's misconceptions, and that it should be usable in conjunction with other diagnostic items to support an

effective and efficient formative assessment classroom structure as proposed by Wiliam (2011b).

RQ1 is now discussed by the individual micro-domains.

PV: Place-Value Knowledge

The Rasch analysis shows that Items 10005, 10001, 10002, and 10004 detect the WHOLE misconception better than others.

DO: Decimal Number Ordering

A number of items were identified as suited for diagnosing each of the four coded ways of thinking—L1, L2, L3, and S3—drawn from Steinle's (2004b) work, but there was almost no overlap between the items suited for each of these codes.

Items 10023 and 10038 are better for the diagnosis of the L1 misconception when these are analyzed qualitatively, since one of their two choices points only to this misconception. These two items were also measured as "easy" by the Rasch analysis, in terms of the number of learners who selected these choices. This dual qualitative/quantitative approach to item validity is preferred over using a single approach where this is possible.

Similar findings on good diagnostic items have been made for each of the ways of thinking in this micro-domain.

CR: Common Fraction Representation

Most learners selected the correct choices, showing general proficiency on the verbal representation of common fractions. As a result, there was insufficient data from the learner errors to identify which items are better for the DECIMAL and the RECIPROCAL misconceptions. The most significant result was that a large number of learners selected the decimal fraction which is similar in its digits to the common fraction, being the DECIMAL misconception. For example, Item 10064, where the decimal number 9.12 was selected as the closest representation to "nine twelfths", when the choice $\frac{9}{12}$ was also available.

Three explanations are suggested for the lack of learner errors: firstly, that the learners already had proficiency in this field; secondly, that the test items were

insufficient to gather evidence of misconceptions; and finally, that the misconceptions in this micro-domain are not well understood and require further study.

NL: Number Line for Common Fractions

Five of the ten test items in this micro-domain resulted in systematic errors made by the learners, and there was evidence of three separate misconceptions. Firstly, the TICKSPARTS misconception which occurs when a learner mistakes the number of ticks as being the same as the number of parts. This misconception was evident from Items 10070 and 10077 and either of these is a good diagnostic item. Secondly, the WHOLELINE misconception, which occurs when learners treat the entire number line as being on the scale 0-1 even though the labels show that the scale is wider. This was evident in Item 10072, where 3/78 learners selected $\frac{3}{4}$ when the arrow pointed at a tick midway between 1 and 2 on the scale of 0-2, and where the majority of learners selected the correct response. Thirdly, the DECIMAL misconception, where the rich distractor is a decimal number, even though a common fraction is requested. This was evidenced from the results of Item 10072 where 26/80 learners selected 1.5 over the correct response of $1\frac{1}{2}$ which was selected by 47/80 learners.

I conclude that the items identified here are all good for isolating these misconceptions even though the data was limited.

CG: Common Fraction Graphics

Misconceptions were not pre-identified for this micro-domain. Rather, the learner responses were analyzed to explore patterns of systematic errors made by the learners and whether these could be formalized as existing or new misconceptions. Some errors appeared to be the result of misreading the item and its choices rather than being a misconception. However, this may be an example of non-cognitive causes of errors which warrant consideration for diagnostic assessment, such as when the previous item influences a learner's reading of the subsequent item.

It is not possible to make further conclusions about the nature of the misconceptions in this micro-domain, and hence no way to determine which items are better than others.

CO: Common Fraction Ordering

The NUMERATOR misconception occurs when the learner uses only the numerator to determine which is larger or smaller, and the DENOMINATOR misconception is when the learner uses only the denominator to make this determination. The Rasch analysis indicated that Items 10093, 10101 and 10103 are better for the NUMERATOR misconception, and Items 10095, 10109, 10110 and 10111 better for the DENOMINATOR misconception; however, when analyzed qualitatively and compared to each other, it was not obvious from the structure and content of these items why one item is better suited than another.

CE: Common Fraction Estimation

Two types of tests were used for this micro-domain, with CE1 asking for the closest number to a common fraction, and CE2 asking for the closest common fraction to a given number. Two misconceptions were considered: NUMERATOR, in which the learner selects a choice based upon the numerator of the common fraction; and DENOMINATOR, in which the learner selects a choice based on the denominator.

For the CE1 items, the analysis was improved when these two misconceptions were combined. This resulted in Items 10013 and 10015 being determined as better for diagnosis; being selected by more learners who had higher measures for the use of these misconceptions.

For the CE2 items, learners were selecting choices based upon the similarity between the digits in the stem and in the choices, and little else can be deduced on the quality of these items for this diagnosis.

CA: Common Fraction Addition

Items 10157 and 10158 are both suited to detect the ADDITION misconception; inducing learners to select a rich distractor for this misconception. This occurs when the common fraction in a choice has either the numerator or the denominator which are the sum of the numerators or denominators of the common fractions in the stem.

RQ2 (EFFICIENCY)

This question concerns how many items are sufficient to produce valid evidence of the use of particular misconceptions by the learners. The economic argument is that the fewer

questions are required, then the faster that informed instruction can progress in the classroom.

Wylie and Wiliam (2006) propose a manual approach which uses a single item, but I have argued against this on the grounds of the validity of the results. In general a single test item is unlikely to offer a sufficient level of validity, due to the impact of slips and guessing. I conclude that Wylie and Wiliam's approach offers a good alternative for classroom diagnosis when technological support is unavailable.

My primary findings from RQ2 are that the number of test items needed for a valid diagnosis is not constant between the micro-domains; and that this number is also not easy to determine. The number of items needed is dependent on the various ways of thinking being diagnosed; the diagnostic quality of the items used; and the level of validity that is required. When diagnostic items are designed to address specific misconceptions, then a single test item can provide an indication of learner thinking. However, to efficiently isolate specific learner conceptions sufficient to position learners into the Development Stages then a single test item is mostly insufficient.

The micro-domains are summarized for this research question, omitting those for which no conclusion could be reached.

PV: Place-Value Knowledge

A single item can provide evidence of a misconception, but only if the item sufficiently discriminates between the correct response and the rich distractor. Such an item should provide sufficient rich distractors to discriminate between words such as "thousands" and "thousandths". However, a valid inference that can discount the impact of guessing and slips will require a small number of items, but far less than the 10-20 used in these tests. As required, different items will be needed for different ways of thinking in place value.

DO: Decimal Number Ordering

For items with only two choices, a single item is insufficient to make a valid inference concerning which way of thinking a learner has used to account for his/her response. Steinle's (2004a) study used a standard battery of 30 test items, which is acceptable for the discovery of misconceptions, but is perhaps too large for efficient usage in a classroom setting.

Thus, items for this micro-domain should rather have more choices, so that a single item can produce a more specific conclusion. This is the case for item B10 from TIMSS 1999, which had five choices, and which was analyzed in section "Micro-Domain DO - Decimal Number Ordering" on page 108.

NL: Number Line Common Fractions

Items 10077 and 10078 were reliable indicators of the TICKSPARTS misconception, and either one of these is sufficient for a quick diagnosis. For the DECIMAL and WHOLELINE misconceptions, there was insufficient data to answer this question.

CO: Common Fraction Ordering

A single item is insufficient for diagnostic purposes when these items are structured by comparing two common fractions to select the larger or smaller. Rather, a number of items are required to establish a valid inference on whether the learner is proficient, is guessing, or is using misconceptions. This is essentially the same situation as encountered with Decimal Number Ordering, and common threads on diagnostic validity should be explored between these micro-domains.

Whereas test items with two choices are easy to administer in a classroom setting, each of the choices has multiple interpretations. This may result in invalid inferences drawn from the learner responses when insufficient items are used; or if the results are not analyzed to isolate the most likely conceptual cause.

CA: Common Fraction Addition

The ADDITION misconception can be isolated quickly using a single test item; but this will not provide sufficient evidence to position the learners into different stages. Items 10153 and 10156 are the most suited and together these are sufficient to tentatively position a learner into one of the STABLE, ACTIVE, or ABSENT stages; but this will apply only if slips and guessing are not considered.

RQ3 (SELF-KNOWLEDGE)

This Difficulty Index uses learner self-knowledge, in conjunction with learner responses, to infer the use of misconceptions. The basis for this approach is that when learners identify an item as Easy or Just Right, and who also select an incorrect choice, then this

is an indication of a misconception. The argument is that by identifying an item as Easy or Just Right, the learner has selected a schema which they believe will solve the problem; thus if the response is wrong then it is likely that that schema is wrong.

However, this self-knowledge data may be influenced by a range of psychological factors, such as learners believing that they will be seen as weak if they indicate an item as Difficult. In some cases, learners who selected Difficult were shown to be relatively proficient.

I repeat here the 3x3 decision matrix as introduced originally in Table 10:

RESPONSE TYPE	LEARNER PERCEPTION OF DIFFICULTY		
	Easy	**Just Right**	**Difficult**
Correct response	*Zone of Competence* STABLE	*Zone of Competence* IMMINENT / ACTIVE	**guessing** *Zone of Incompetence* ABSENT
Incorrect - Rich distractor	**misconception** *Zone of Learning* ACTIVE / EMERGENT	**misconception** *Zone of Learning* ACTIVE / EMERGENT	*Zone of Incompetence* ABSENT
Incorrect - Random distractor	**potential misconception** *Zone of Incompetence* ABSENT	*Zone of Incompetence* ABSENT	*Zone of Incompetence* ABSENT

This Difficulty Index may add value to the data gathered from the learner responses to the test items, when coupled with Rasch analysis measures. However, Rasch analysis needs a computer to carry out the calculations to produce the measures; whereas the Difficulty Index can be applied to individual test items without needing a computer.

RQ3 asks whether this self-knowledge on the perceived difficulty of items adds value by reducing the number of items which are needed in a diagnostic assessment. If this is not the case, then the only impact is that the learner is asked to provide information which does not affect the measurement; achieving the same results as obtained when this information is not used. Given this study's economic focus on diagnostic assessment, this self-knowledge should be avoided unless there is a tangible benefit arising from its usage.

For example, consider learner A67's responses to the 20 questions in the Place-Value micro-domain. These responses produced a measure of 1.52, which is evidence of the WHOLE misconception. On this basis this learner was positioned into the ACTIVE

stage; which means that they had sufficient responses which were accountable to the WHOLE misconception to indicate active learning.

These same responses from learner A67 were also applied to the Difficulty Index matrix structure; by counting the responses to each test item based upon whether the response was, or was not, a rich distractor for the WHOLE misconception; and by coupling this count to the learner-indicated difficulty.

Table 41. Learner A67 Difficulty Index Analysis

LEARNER A67	Easy	Just Right	Difficult
Correct response	**4** STABLE	**1** IMMINENT / ACTIVE	**0** ABSENT
Incorrect - Rich distractor	**13** ACTIVE / EMERGENT	**1** ACTIVE / EMERGENT	**0** ABSENT
Incorrect - Random distractor	**3** ABSENT	**1** ABSENT	**0** ABSENT

This matrix shows a total of 23 responses whereas there were only 20 test items in this set; and this is due to three of the test items having a choice which is both a CORRECT response as well as being a rich distractor for the WHOLE misconception.

The highest frequency of responses (13) was found in the highlighted cell. This cell indicates that the learner selected a rich distractor and also marked the item as Easy. This decision matrix positions the learner in either the ACTIVE or the EMERGENT stage; which is effectively the same result as determined from using the Rasch analysis alone. There is thus no additional benefit to using this information if the Rasch calculations are available. On the other hand, if computers are not available in the classroom to perform the Rasch calculations in real-time, then the Difficulty Index offers a scaled-down alternative to infer learner development stages.

I conclude that the Difficulty Index adds little additional value when the Rasch analysis can be performed on the data. However, there are cases in which this adds value, and the Difficulty Index may be applicable in cases in which the computing requirements of Rasch are not available.

7.2 Key Findings from this Study

Final conclusions are presented now in the form of key findings resulting from this study. Most findings result from the research questions for this study; and other findings are related to the Development Stage model, and to practical issues and implications of Web-based, diagnostic assessments. These key findings are the most significant conclusions resulting from this study.

Finding 1: Learners can be positioned into Fine-Grained Development Stages

My initial literature review identified the lack of a fine-grained model of learner development which was suited as the basis for my study. The purpose of such a model would be to identify the progress of learners through their developmental progress in micro-domains from novice to mastery; while considering the nature of the learners' conceptual model, including their misconceptions.

During daily instruction learners are exposed to new micro-domains, which are often presented as specific problem types. The learners use both prior knowledge and new schemas to follow the teacher and to carry out their work. For example, a learner's prior knowledge may consist of informal whole number knowledge which is applied to new problems in common fractions. A learner's schemas include intermediate conceptions which are developed during learning in a new topic, and which are applied to more complex and challenging problems as introduced by the teacher. As an example, once a knowledge of the addition of common fractions with equal denominators is established, then the teacher may use examples with unequal denominators. This introduces more complexity into the mathematical representation, requiring the construction and reconstruction of schemas, so as to accommodate prior knowledge of common factors.

My concern in this study was to identify where the learners are located in their conceptual development in a micro-domain so as to enable more targeted instruction and possible remediation. The Development Stage model is a response to the practice in which learners are considered to know a topic when they get the right answers; while other learners are considered to not know a topic by getting the wrong answers. Many tests are marked by simply giving a tick or cross, and this provides insufficient information to determine a learner's stage of development in a micro-domain. Two problems arise with

such a simple marking approach. The first, being learners who know what they are doing and who get a wrong answer, being a slip, and secondly, learners who do not know what they are doing and yet achieve success, being the result of a lucky guess.

Learner knowledge in a micro-domain is not a simple dichotomy which separates those who know from those who don't know; rather, understanding develops over a spectrum of evolving proficiencies and conceptions. The Development Stage model is the core theoretical contribution of my original research proposal, and has remained mostly unchanged. However, it has been refined to include the boundary stages of EMERGENT and IMMINENT, which account for transitions between the primary stages of ABSENT, ACTIVE and STABLE. As learners start their learning in a micro-domain they become a novice and are consider to be in the EMERGENT stage, which quickly leads to active learning in the ACTIVE stage. As their knowledge matures, they get close to mastery in the IMMINENT stage. This model represents the fine-grained development stages of a learner in a micro-domain, and also has the potential to identify which diagnostic test items are best for learners in the different stages.

This model assists both with the interpretation of learner responses and the explanation of these responses as being one of the following:

- correct responses which are truly reflective of achieved proficiency;
- slips by otherwise proficient learners;
- the use of preconceptions or intermediate conceptions; which are both considered as misconceptions that are an integral element of conceptual development;
- random guessing, resulting in both correct and incorrect responses, from learners who have insufficient knowledge to address a problem.

The five stages of the model represent a continuity of learning in a micro-domain, from novice (ABSENT → EMERGENT), through active learner (ACTIVE → IMMINENT), to mastery (IMMINENT → STABLE).

The approach adopted for the data analysis in Chapter 6 was to position learners on the stages of the model and specifically to identify misconception usage in the ACTIVE stage. This information can be useful to teachers if it can be provided automatically and immediately—with little effort on the teacher's part—as part of daily instruction based on formative assessment practices. Such fine-grained information on

learner knowledge is not currently available through other means, since the amount of data required to be gathered and processed is too large for a teacher to handle. However, with the potential for widespread availability in the classroom of devices connected to the Internet, this has become a possibility.

My finding is that it is possible, using the approach developed in this study, to position learners into these stages for each micro-domain, and to accomplish this with a small number of items.

The application of this model to each micro-domain has highlighted some issues which point to further work, as follows

- How the micro-domain knowledge and proficiency is mapped to the stages for the micro-domain.
- The role and meaning of the ABSENT stage.
- Setting a reasonable cut-point in the Rasch measures which indicate when a learner is in the STABLE stage, given that the cut-points used are inconsistent between the micro-domains.
- Setting a reasonable cut-point for the ABSENT stage from the Rasch analysis for which learners are considered to know nothing about the topic.

One additional outcome is that the stages in this model can be used as a common vocabulary for teachers; since these stages are meaningful across micro-domains and they point to appropriate remediation for groups of learners who are in a particular stage. Thus the model may provide a benefit to learners similar to what would result from personalized instruction, as was contemplated by Bloom (1984).

It is my finding that this Development Stage model is valid and indicates real stages in the growth of proficiency; and consequently that it can provide teachers with fine-grained knowledge of their learners' misconceptions in a way that cannot be provided using traditional means. However, the data processing to compute these measures from the learners' responses requires computational support.

This finding on the Development Stages raises the question of whether this model should itself have been included as one of the research questions for this study, but I chose to adopt the position that this model was a means to an end, where the end is to better understand effectiveness and efficiency in diagnostic testing of misconceptions.

An additional finding is that learners who are in the ABSENT stage are those for whom the wording, notations, and nature of the problems are beyond the learners' understanding. In other words, these learners lack the capability to even read the question and to understand what is expected in a solution. These learners are failing on Polya's (1945/1990) first step in problem-solving; that they are unable to read the question. When the learners move beyond the ABSENT stage, they are commencing their conceptual development in learning the mathematical language of the problems in this micro-domain. The finding is that this ABSENT stage is not well understood within this model and that this warrants further research into the initial emergence of conceptions within new micro-domains.

Finding 2: Some Test Items are Better Suited for Diagnostic Assessment

My second finding concerns the suitability of test items for diagnostic purposes, and it is my conclusion that this suitability is measurable using a method analogous to how the Rasch method addresses the notion of item difficulty.

Rasch analysis was used to measure the propensity of learners to use specific misconceptions in responding to items. This satisfies RQ1 to determine why some items are better than others for diagnostic purposes; and it is shown that Rasch analysis is adequate for this purpose.

The Rasch analyses were conducted in parallel for the misconceptions being assessed, given that there were multiple combinations of conceptions and misconceptions that could account for learner responses to individual items. This approach is detailed in Appendix F which explains how the learners' responses were marked for input into the WinSteps program.

The results indicated which learners held which misconception and to what level, and also which test items were more likely to be selected by learners who were using a particular misconception. During this process items were removed when they did not fit the model and also when they did not correlate with calculated learner measures, since these items may have distorted or degraded the model (Linacre, 2002).

For each micro-domain, items were identified which are better suited for diagnostic purposes. Thus some items are shown to be effective indicators of misconceptions, while others are not; and the criteria for what constitutes a good diagnostic item was not evident in many cases from a visual inspection of the item.

However, there were some cases in which a visual inspection was useful and this may help to predict the diagnostic value of items. Thus, a test item may be counter-productive for diagnostic purposes unless the quality of the item is assessed using a method such as the parallel Rasch approach which has been used in Chapter 6.

Finding 3: Which Micro-Domains are Suitable for Diagnostic Assessment

The micro-domains used in this study were located within the domain of the rational numbers and were not independent. Some misconceptions were found across these micro-domains, many are based on prior whole number knowledge, and their effect across these micro-domains was similar in terms of the learner responses.

These micro-domains included problem types which are well-established and documented within the misconceptions literature and included problem types which have not been subjected to such intensive analysis. These latter types of problems are suited for the exploration and discovery of new misconceptions—which may be local to a single micro-domain.

When misconceptions are common, then any micro-domain can be used to isolate learner use of these misconceptions. Thus it is better to use a micro-domain which is easier to apply. For example, it is clear that the problem of identifying the relative magnitude of decimal numbers provides ripe ground for uncovering and documenting a range of fine-grained behaviours. However, it is also possible that other types of problems may elicit evidence of the same misconceptions, and may achieve this with less diagnostic effort.

The items in the Number-Line micro-domain have helped to uncover a misconception which has received little attention in prior research, and which I refer to as TICKSPARTS. This is evidenced by learners equating the number of ticks on the number line with the number of parts into which the number line is divided. This misconception was observed initially from a frequency analysis of learner responses. This leads to the possibility of using large-scale Web-based data on learner responses as a research tool to continually discover fine-grained ways of thinking; on the basis of observed and unexplained patterns of responses. This approach raises solvable ethical issues on access to information, and it may allow for the identification of misconceptions which are specific to a particular learner community, such as a single class or school.

Micro-domains are considered as ideal units of knowledge for diagnostic work. From the experience of the eight micro-domains used in this study the following characteristics are identified as suited to define the ideal micro-domain:

- Having a small scope, and with the availability of simple test items which are easy to administer quickly.
- Well understood in terms of the conceptions and misconceptions which learners use when addressing problems in the micro-domain.
- Represent a barrier to learning if not understood and may become persistent if not attended to.
- Applicable to classroom work, either on its own or as a pre-requisite to new work.

Finding 4: Self-Knowledge adds little value to the Diagnosis of Misconceptions

The Difficulty Index proposed in Chapter 4 was conceptually well-founded, and is based upon a motivated argument from earlier studies. The finding is that the additional effort in applying the Difficulty Index does not add incremental value to the diagnostic assessment process to warrant its regular usage in the classroom. Rather, learner self-knowledge should be used only where it enhances the conclusions, and when a finer level of positioning into the Development Stage model is needed. This can help to isolate new ways of thinking which are not currently identified, formalized, and being detected.

I conclude that the Difficulty Index should not be used as normal practice in classroom settings, given that it requires additional time and effort for collection and analysis, and that the same results are obtained from the Rasch analysis alone.

One exception is that this approach has potential benefit in classrooms where there is no access to the computing requirements for Rasch analysis.

Finding 5: No Common Rule for the Number of Items Needed

Classroom time is a limited resource, and deriving learner knowledge from the least assessment effort is always beneficial. The alternative to short, targeted tests is to use larger batteries of items which take time to administer and to evaluate; and the CAPS curriculum states that diagnostic assessments should be conducted as efficiently as possible (DBE, 2011a).

For some micro-domains, a single question will suffice to distinguish learners who know, and are in the STABLE stage; from those who are using the misconception, and are in the ACTIVE stage; and from those who are indeterminate, and who are placed into the ABSENT stage. On the basis of a single test item this is as much as can be inferred. Such a test item must satisfy, as far as possible, the properties of the "semi-dense" criteria of Bart et al. (1994), so that learners' responses are sufficient to distinguish between different ways of thinking.

However, for the micro-domain of Decimal Number Ordering using items with only two choices, it is not possible to distinguish between correct responses and rich distractors using a single item, and this limits the usefulness of these items for diagnostic purposes.

Finding 6: Diagnostic Items Should be MCQ and Need Sufficient Choices

Constructed-response items are not suited for diagnostic assessment due to the number of possibilities provided by learners, and also the challenge in interpreting these responses. These responses require a strong semantic model for interpretation; and consequently this is left to future models and technologies which may open up opportunities for such analysis. On the other hand, MCQ items are limited by forcing the learners to select only from the choices provided; and these choices may miss important misconceptions if they are not designed properly.

MCQ items with only two choices are suggested as being too limited for general use as diagnostic instruments in the classroom, due to the larger number of such items required to provide evidence of the various misconceptions. Rather, consideration should be given to meet the semi-dense criteria of Bart et al. (1994), which stipulate the relationship between the ways of thinking and the item choices; and this naturally requires more choices to reflect the various ways of thinking.

Finding 7: High- and Low-Proficiency Learners Need Different Methods of Measurement

Learners with high proficiency, being those in the STABLE stage, must be measured using a different approach than learners who are not proficient.

This finding is based on the theoretical model of schemas in which a learner's knowledge consists of a set of schemas which come into play as required for a particular

situation. Changes in schemas equates to the process of learning, and at every point in time these schemas will be fit for some purpose, but perhaps not for all of the purposes which the learner is presented with.

Measurements of learner ability is effective and meaningful only when the learner's schemas are sufficient to address the problems presented in a micro-domain. However, when the schemas are insufficient then the approach must change to focus on identifying which of many possible misconceptions the learner has used.

On this basis two forms of measurement are provide for the analysis. The first is the traditional approach to measure proficiency or ability; ideally using a Rasch analysis scored on the correct choices in the MCQ items. The second is conducted by running separate Rasch analyses in parallel, one for each known misconception, to determine the most likely explanation for the errors made by the low-proficiency learners.

My finding is that this approach—separating the measurement of high-proficiency and low-proficiency learners—has considerable value for diagnostic assessment while also clarifying the distinction between the measurement of ability and diagnostic assessment.

Finding 8: Very Low Values may Indicate Better Conceptual Development

Very low scores may result from a single learner's responses to a number of MCQ items; but can also result from the responses of a number of learners to a single MCQ item. Given an MCQ item with four choices, then learners with no conceptual basis to address the item can only resort to guessing; and they should achieve an expected score of around 25%. However, a learner may achieve a score which is far lower than 25%, such as 10%, and this result indicates a systematic pattern which can only result from the use of misconceptions in selecting the choices. Thus a score of 10% is an indication of some level of conceptual development, whereas a score of 25% is more likely to result from random guessing.

I argue that some level of conceptual development, as exhibited in learners who score 10%, will always trump the lack of conceptual development, as exhibited by learners achieving 25%. Thus these low scores are not correlated with conceptual development, and this applies no matter how these scores are computed, such as from a CTT sum of scores or from a Rasch analysis of the learners' responses. This argument

applies for all scores which are below 25% for cases of four choices per MCQ item. A similar argument holds for MCQ items with different numbers of choices.

In summary, for the low scoring learners, it is the case that a lower score represents more conceptual development than a higher score.

The outcome of this is that the low performing learners need to be measured using a different approach, as discussed in Finding 7 above, and this approach has been used throughout Chapter 6 as part of the process to detect misconceptions.

CHAPTER 8 : CONCLUSIONS AND FUTURES

8.1 Contributions of this Study

I now present the major contributions which emanate from this study, based on the detailed data analysis in Chapter 6 and the findings summarized in Chapter 7. Some of the arguments are repeated from previous chapters to support the contextualization of this study. These contributions are:

- Methods to assess proficiency are not suited to the measurement of low-proficiency learners.
- Diagnostic assessments can be conducted effectively and efficiently. This includes the claim that traditional scoring is inapplicable for low-proficiency learners; requiring an alternative approach which addresses the schemas of the conceptual models of the learners rather than their composite ability.
- Information on learner self-knowledge of their proficiency adds little or no cost-benefit over quantitative methods of measurement.
- Small micro-domains are suited for diagnostic assessments.
- The Development Stage theoretical model is suited to position learners based on their development stage; which aids the formative assessment process.

Methods to assess proficiency are not suited to the measurement of low-proficiency learners

My first contribution is the approach to differentiate the measurement processes for high-proficiency learners, who are in the STABLE and IMMINENT stages of development, from other learners who are in active learning stages.

This was not posed as a research question, but was applied throughout the data analysis of the online data. There was some benefit in the measurement processes for firstly identifying the proficient learners and separating them out. Given that the purpose of this study was to explore the diagnostic assessment of misconceptions, it was thus important to use that subset of the learners who were not proficient, rather than to include proficient learners for whom diagnostic assessment was not required.

This is a contribution which can be applied to all situations in which diagnosis is needed, and is applied as a two-phase approach, with the first phase identifying the

proficient learners, for whom there is no evidence of content-related challenges, and the second phase, which uses the alternative methods recommended in this study, for the specific identification of misconceptions which are the content-related challenges.

The economy of diagnosis

The second contribution is that diagnostic assessments can be conducted effectively (RQ1) and efficiently (RQ2) by asking the right items in the right quantity, and by processing the responses to produce valid measurements. Thus, this study can be viewed as a contribution to the economics of educational diagnostic practices.

This economical approach requires a new way to gather and process data about learning thinking. The approach explored in this study is a novel application of Rasch analysis, calculating measures in parallel for known misconceptions; then using this to identify the most likely conceptions which account for learner responses. This approach is contrasted against the measurement of learner ability, which is seen as a composite of schemas as the measurable construct.

This approach addresses the challenges of measuring low-ability learners using conventional scoring, and which is summarized in Finding 8 in Chapter 7.

This approach also addresses the critique on the unsuitability of IRT for diagnostic purposes (Stacey & Steinle, 2006) which uses learner "ability" as the construct being measured and in terms of this usage I agree with their critique; considering that this critique of IRT will apply similarly to Rasch analysis. This critique is addressed by shifting the target of assessment to fine-grained misconceptions which occur in the mathematical reality of the learners. This approach identifies the conceptions and misconceptions which are more likely to account for learner responses, including both correct and incorrect responses. This provides a fine-level window into learner thinking which is not readily accessible using any other means; and is achieved by treating diagnosis as the presence of misconceptions rather than as the absence of general ability.

This approach is contrasted against the consideration of expertise as a composite, but unknown, set of schemas which are sufficient to answer all problems in a micro-domain. Thus to measure expertise it is not necessary to know the specific conceptions which a learner has used, since expertise is inferred purely through consistently correct responses without any need to understand the specific conceptions used.

There is little or no cost-benefit to using learner self-knowledge for diagnosis

My third contribution concerns the usage of information on what learners think about the difficulty of the items which they are required to answer in assessment tests.

Finding 4, as presented in Chapter 7, concludes that whereas the information on learner self-knowledge can help to infer a learner's Development Stage, it does not add anything which is not already available by using the Rasch analysis process. In essence, the quantitative analysis of the data from learner responses to good diagnostic test items is sufficient to meet the goals of effectiveness and efficiency of diagnostic assessment.

Such a Rasch analysis requires computer support, since the inferences require a complex calculation which uses the learner's responses, and the item characteristics to infer the learner measure as the most likely measure which would produce these responses. The learner measures for usage of misconceptions are only valid as they are measured, since the learners' conceptions change fast in the work in micro-domain. Thus if there is no computing support then these results cannot be calculated. In this case, the suggestion is to use the learner self-knowledge, which can complement the scores to make inferences without needing a complex calculation. However, these cannot provide the finer details of learner misconceptions as are available within the Rasch analysis, but they can support the positioning of the learners into the development stages.

The Development Stage model can position learners in a micro-domain

A fourth contribution concerns the Development Stage model which I have used to support my study. There is a need to better understand learning thinking in order to inform instruction; and this need was the motivation for the Development Stage model which is a supporting theory throughout this study . Even though this model was included as a research question at the time this work commenced, I now see this as a valuable contribution to knowledge of learning thinking in micro-domains, and as a separate outcome from this study.

This model positions learners into one of five development stages and has been applied to the micro-domains I have used; specifically, to those micro-domains with sufficient data. Prior work in development models has addressed larger-scale, often multi-year, learning trajectories, such as the work of Confrey et al. (n.d.) on equipartitioning and other components of the Common Core State Standards for Mathematics (National

Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). The Development Stage model provides a quantitative, fine-grained scaling of the novice-expert spectrum, with the potential to position all learners, at every point in time, onto one of the stages, based on their personal development within a micro-domain. A learner is expected to move through these Development Stages rapidly during the learning of a new micro-domain, and I envision a future in which it will be possible to measure progress through these stages instantaneously as learners increase their proficiency through their engagement in classroom activity and with targeted assessments. This measurement information can be used to provide an "X-Ray" view of the mental structures of the learner and could be presented in a dashboard user interface to aid a teacher in understanding the current stage of development of each of their learners, to enable the teachers to engage in a kind of "brain surgery" to help the learners to fix the gaps in their conceptual model.

I propose that this Development Stage model is generally applicable to all micro-domains, and could be coupled with an automated process to position learners on the basis of the processes and algorithms described in Chapter 6.

Micro-domains are suited for diagnostic assessments

A final contribution arising from this study is an outcome from the nature of the diagnostic work conducted, rather than being an outcome from the research questions. This concerns the usage of "micro-domains" as the targets for diagnostic work, and is contrasted against the application of diagnostics to larger domains of knowledge. I argue that micro-domains are the ideal size of knowledge for the application of diagnostic assessments for identifying incomplete, incorrect, or inapplicable schemas which influence learners' responses; however, this size is not defined exactly but is expected to be small, being addressed within a maximum of 1-2 days of classroom instruction.

Criteria for micro-domains are proposed in Chapter 7, which may be used to demarcate a micro-domain to meet the needs of diagnostic work. These criteria are outlined in "Finding 3: Which Micro-Domains are Suitable for Diagnostic Assessment" on page 266.

The alternative to micro-domains is to attempt to diagnose larger domains of knowledge, such as the entire rational numbers at a unit; however larger domains are not suited for the diagnosis of misconceptions, since there are too many conceptions which

come into play in these larger domains, leading to a challenge in identifying which schemas are used. Such larger domains are better suited for measuring ability, which is by its nature a composite of conceptions.

8.2 Relating Findings to Theory and Literature

I return to the literature as reviewed in Chapter 2 to position my work with regard to the established theories and practices and how my findings address the gaps and opportunities which I had identified. The scope was wide due to the multi-disciplinary nature of this work and includes constructivism, misconceptions research assessment practices, educational diagnosis, Web-based assessment, Rasch analysis, Cognitive Diagnostic Assessment, and the rational numbers. A few core works and themes were selected which have informed both the motivation and the form of this study, and which now I reflect on in terms of my contributions.

Formative Assessment and Diagnostic Assessment

There has been an increasing focus on assessment for learning over the past 20 years. Wiliam (2011a) asserts that "assessment for learning" and formative assessment are almost identical in their usage, and that their goal is to provide evidence about learner thinking to teachers to inform instructional planning and activities. To be beneficial, this evidence is required to include diagnostic information on learner conceptual difficulty, and thus diagnosis is best positioned as an element of formative assessment practice (Black & Wiliam, 1998a; Black & Wiliam, 1998b; Wiliam, 2011a; Wiliam, 2011b; Stacey, 2013; Stacey, Price & Steinle, 2012).

Black and Wiliam (2009) have proposed a set of strategies for formative assessment, including that to be effective, classroom activities tasks should elicit evidence of student understanding. I have argued that this evidence must include misconceptions as an integral element of student understandings.

My work has taken this requirement for diagnostic information and has explored how diagnostic assessment can become both more effective, by selecting items which are better for diagnosis, and more efficient, in identifying the number of items needed. I have argued that traditional approaches to the assessment of ability are not suited for diagnostic purposes, and that an approach that targets specific misconceptions is required.

Assessment Practices

Pellegrino et al. (2001) suggest that assessment consists of three interrelated elements, structured into the "assessment triangle". This structure applies to all forms of assessment in which the understanding of learner thinking is the goal, and this includes diagnostic assessment.

The first element is the conceptual model of the students, which I have treated as consisting of a set of schemas, which the learners use to address mathematical problems. I assume that this is also the sum total of the knowledge of the learner; thus there is no other knowledge which exists outside of the schemas. The goal of diagnostic assessment is to discover the schemas which are not fit for the purpose of specific micro-domains, and which are thus considered as misconceptions.

The second element consists of the observations which are taken from learners, and which help to determine the learner's conceptual model. These observations are the learners' responses to "good" diagnostic questions.

The third element consists of the inferences which are drawn from the observations, and which convert raw observations into meaningful measurement of the learner's usage of misconceptions. The inferential approach used for this study consists of a parallel application of Rasch analysis, which uses the learner responses to identify the most likely schemas to account for these responses.

Rational Number Misconceptions

The domains of application for this study are specific micro-domains of the rational numbers. One of the important micro-domains concerns the problems of ordering decimal numbers, and inferences which can be drawn about learner misconceptions. I have based my work in this micro-domain on the detailed model of learners' ways of thinking on decimal numbers as identified and coded by Steinle (2004a). Four of Steinle's codes—L1, L2, L3 and S3—have been included into my study, and these have helped to identify specific misconceptions used by learners. For some learners there was significant evidence of the usage of a single one of these misconceptions whereas other appeared to be using many.

It was not my intention in this study to add to the knowledge base of misconceptions in the rational numbers, but rather to explore how the wealth of prior knowledge can be applied for diagnostic assessment practices.

Measurement

Educational measurement is concerned with the inference element in the assessment triangle of Pellegrino et al. (2001). The measurement process takes observations, consisting of learners' responses to items, and from these concludes something about the learners' conceptual models. It is important that educational measurement is consistent and standardized, so that the measures have an interpretation across different groups of learners, on different tests, and at different times. The notion of "fundamental measurement" has been a primary motivation for the development of better methods of measurement in the social sciences (Bond & Fox, 2012; Wright, 1997).

My purpose has been to find a more effective approach to diagnostic assessment by using the right questions to detect and isolate misconceptions used by learners, which is my RQ1. This is a measurement problem and is intended to result in consistent measures so that if a learner achieves a particular measure on these tests, then this will result in a corresponding inference on which schemas, and in particular which misconceptions, best account for the learner's responses.

Whereas there are a range of models which can support educational diagnosis, such as those outlined by Rupp et al. (2010), I have chosen an alternate approach by using simpler Rasch models in parallel and then identifying the most likely schema which a learner is using. These results are then used to help determine whether a learner is proficient or not, and to what extent they are actively learning, by positioning the learners into one of the Development Stages.

Web-based diagnostic assessment

There had been little work reported on Web-based diagnostic assessment of mathematics at the time I prepared my research proposal for this study; however, in my updated literature review, conducted towards the end of this study, Web-based diagnosis is emerging as a topic of interest for research and development. This has provided the opportunity to position my work in a broader body of work and to relate my findings and contributions with this evolving knowledge base.

SMART tests ("Specific Mathematics Assessments that Reveal Thinking") (Price et al., 2011) provide teachers with an automated, Web-based environment to conduct assessment for learning. The SMART test approach is close in concept and in application to my work and has similar goals. The SMART application is in use in the State of Victoria in Australia. My work differs from SMART in three ways. Firstly, the manner in which I have selected test items suited for diagnostic assessment, using Rasch analysis to identify good diagnostic items. Secondly, my goal is to gain maximum diagnostic information from the smallest number of test items, exploring an economic view of the efficiency of diagnosis. Finally, my generic model of Development Stages is distinct from the domain-specific learning hierarchy as provided within SMART.

For the SMART tests, Price et al. (2011) provide a learning hierarchy which they compare to the six-level OECD scale of proficiency used in the PISA study (OECD, 2010). My Development Stage model provides an alternative dimension for analysis which is distinct from the models of Price et al. and the OECD. My focus is on fine-grained conceptual development in micro-domains of knowledge, and my goal has been to understand where a learner is positioned in terms of their personal development from being a novice to exhibiting expertise in a micro-domain. Both Price et al. (2011) and the OECD (2010) present models that are learning trajectories and whereas my model is structured as a trajectory, it is based upon the shift in proficiency at a fine-grained level. I have scoped this study to address the problem of understanding learner proficiency by identifying their position in my Development Stage model. I have not attempted to assess or inform learning; which I view as changes in the conceptual model of a learner triggered by effective feedback and informed instruction.

8.3 Surprising Results

The most surprising result obtained was that in many of the micro-domains a single good diagnostic test item can be used as diagnostic indicator, even though more test items are needed to establish sufficient validity. It is possible to use these test items in non-technology classrooms and to provide the teachers with the some of the benefits as would be available in technology-linked classrooms.

I had the initial expectation that the self-knowledge of the learners would be universally useful to assisting the diagnostic assessment process. However, it became

apparent during the data analysis that these additional questions on whether the learners found the test items to be Easy, Just Right, or Difficult was unreliable, and was also unnecessary, given the power of the statistical inferences provided by the Rasch Analysis. In essence there was little or no added value in using this self-knowledge as an input for diagnostic purposes. However, there were a few cases in which this approach proved useful, and it was noted that this can also complement the use of diagnostic items when there is no computational support for calculating the Rasch measures.

A final, but highly significant, surprise result has arisen from my analysis of the use of MCQ items which include rich distractors. When responding to such test items, learners with a lack of schemas to address the test items will be randomly guessing their responses, and will achieve a higher score than those learners who have developed a few schemas and who are more likely to select the rich distractors. Thus for low-proficient learners, the raw scores obtained are the opposite of the learners' actual proficiency, since the lowest scores, which are well under the average expected from random guessing, can only result from systematic choices based upon some level of conceptual development. This is a general result which occurs in all cases where MCQ items have rich distractors which are based on misconceptions and when the scores are applied in low-performing learners. In conclusion, some conceptual development in a learner's conceptual model should always trump the lack of conceptual development, where the opposite is shown from the raw scores.

8.4 The Future: Implementation and Research

I embarked on this study with two aims. Firstly, to advance knowledge in diagnostic assessment of mathematics misconceptions. Secondly, to inform practical intervention into classroom practice, using computers for learning. My motivation was clear from the start—that the problems of mathematical teaching and learning in South Africa are at a crisis point which will not be resolved in the short term using the current approaches to teacher development, improvements in classroom resources, or by curricula reform. This crisis is far larger than any of these approaches can address, either individually or in concert with each other. To address this crisis, we need to exploit the power of automation for the benefit of education. My own background is artificial intelligence and knowledge representation, which has provided me with a unique lens concerning potential theoretical

and practical opportunities for addressing this crisis. It has been a significant shift in my academic and professional focus to move into the rich world of educational research and practice. Throughout the past 40 years I have maintained a mathematics tutoring activity with a few selected students every year, and I have been driven to conduct this research through my goal to apply methods which I naturally use, as a tutor of mathematics, to a far wider audience.

I now raise the question of what can be done to turn the findings and theories of this study into practical application in support of my original aim—to achieve the potential of making a systemic difference in the quality of classroom mathematics of South Africa. In other words, this is a vision to create a widespread and sustainable increase in mathematics proficiency among the learners of the country.

The White Paper on e-Education (DOE, 2004), and the more recent Schooling 2025 Programme (DBE, 2011b) both highlight the need for computers and the Internet, but focus on the usage of computers for administrative support, and on access to Internet resources, and do not address the role of computers in support of learning. In particular, they completely ignore the role of computers for assessment. Since the introduction of the electronic calculator into the classroom in the 1980s, there has not been any major shift in the widespread use of technology in the classroom to support mathematics teaching and learning. The potential now exists, from the combination of inexpensive tablet computers and access to the Internet, to create the critical mass, the tipping point, at which the system may change forever in a positive way.

At the outset of my study, I found little evidence of work being conducted on the widespread usage of computers for assessment, occasionally referred to as "Computer-Aided Assessment" (CAA) and which specifically addressed the diagnostic assessment of rational number knowledge. This is the case even though computers had been available for educational usage for at least 30 years—since the advent of personal computers and laptops—and yet computers continue to remain largely unused for supporting education in the rational numbers in the classroom.

My work can move forward in two directions in parallel from this point. Firstly, as the practical application of this work into the classroom, initially on a pilot basis with a few schools while exploring the potential for widespread implementation. I recommend this as a combination of research and practice, where the focus is on the introduction of

these new practices into the classroom, and for which research should explore the readiness of the teachers, learners, schools, and classrooms to implement technology into the classroom. This also requires a study on how feedback may be used to close the teaching-learning loop, using the diagnostic information to provide fine-grained information about the learners' conceptualizations.

My second proposed direction for this study is to extend the research to include other micro-domains in mathematics and to seek common patterns between these micro-domains. Even within the existing micro-domains that I have studied here, there are many unexplained patterns of errors, which may be researched as new ways of thinking and then applied back into the diagnostic model.

The nature of this proposed research should also include the following core problems that I encountered in my work which are gaps that require a research approach in combination with a practical implementation:

- Redefining the role of the teacher in the technology-rich classroom—would teachers be better used as mediators of knowledge with support from trusted online knowledge bases than as educators working without technological support?
- Exploring the culture of the classroom, with regard to both the teacher and the learners, in terms of readiness to adopt assessment technology and formative assessment practices.
- As a side-effect of the introduction of technology into the classroom, investigating the acceptance or resistance of teachers to the removal of key responsibilities for teaching, and how this could affect the implementation process.
- Finer-grained analysis of micro-domains, and establishing the linkage between micro-domains, seen as small and incremental steps in learning, with the larger learning trajectories that map progress over many years of learning and which are linked into the curriculum standards and practices.
- Repeating this study in other micro-domains both within and beyond the rational numbers. Suggestions include percentages, ratio and proportion, algebra, factorization, and eventually the entire Senior Phase curriculum. The purpose would be to stimulate mathematical proficiency to enable more

learners to be well-prepared for pure Mathematics in Grades 10-12, rather than taking the "easier" option of Mathematical Literacy.

- Exploring the creation of a constructivist learning environment in the classroom using formative assessment practices driven by a Web-based tool.
- Finally, investigation into the handling of very large amounts of data, commonly called "big data", that is gathered at the systemic level from all classrooms concerning diagnostics and misconceptions. How can this be used to support data-driven research and theory evaluation? Also what are the ethical and privacy issues that arise in making such data available for secondary analysis?

8.5 The Prospects for Web-Based Diagnostic Assessment in the Classroom

My research hypothesis for this study is that Web-based online diagnostic assessment is an effective way to improve mathematics at the systemic level, in situations where there is access to computing facilities and broadband technologies within the schools. This is a technology-based extension of the position taken by Wiliam (2011a) on the importance of the teacher's knowledge of their learners' knowledge to support formative assessment practices. My goal has been to extend this position to address online diagnostic assessment in the classroom.

The introduction of computers in schools in South Africa has been slow, and also fraught with challenges and difficulties, including the lack of security to protect the computers, and the lack of capacity of the teachers and learners to use the computers effectively. However, the potential offered by connecting these computers to the Internet is large, and is a significant benefit for learners in schools who have this capacity. Access to the Web does not require the same level of computer literacy as needed to operate a computer, and thus accessing a Web browser poses fewer issues of learner and teacher capacity, and is thus more amenable to under-resourced and under-capacitated schools.

I commenced this study on the basis that I was required to use schools that already had the infrastructure and the capacity in terms of a good computer facility; and in which both learners and teachers were computer-proficient. It was essential for my work that

these computers were also connected to the Internet and that the learners knew how to use the Internet.

Given these pre-conditions, which I hope to be the norm in all schools within a few years, there is an emerging opportunity for how to use these computing facilities, and my research is focused on the usage of these facilities for classroom-based diagnostic assessment.

My study involved the creation of a few tests which were administered over the Web to four classes in two separate schools. Arising directly from the study were a number of important findings concerning the usage of the Web for administering the tests:

- Firstly, that the amount of data gathered is large, even for a small number of classes and tests, resulting in hundreds of thousands of data elements which required data processing. When implemented on a larger scale, this will move into the area of "big data" and will provide the basis for detailed analyses of such data in ways not possible at present.
- Secondly, that the feedback was provided immediately to the learners at the end of a test and, given the analyses performed in Chapter 6, this feedback could be automated to provide a comprehensive analysis to the teacher in real-time.
- Thirdly, that I, as the test administrator, was able to check in real-time what was happening for each individual learner, and to detect any issues.

I firmly believe, based upon this study, that diagnostic assessment can be incorporated successfully into daily classroom practice, but that this will require access to the Internet in every mathematics classroom. Whereas a few years ago, when I started this study, this was infeasible due to the costs of procuring the computers, there are now tablet computers that are lower in cost than textbooks and this can drive an effective paperless educational environment. There are currently initiatives underway to provide all learners with their own tablet, such as Gauteng's announcement to provide learners in Grades 4-9 with tablets, which commences in 2015 with an initial 21 schools[6]. It is expected that these tablets will be connected to the Internet, and this offers an opportunity

[6] http://www.itwebafrica.com/mobile/320-south-africa/233202-gautengs-plan-to-provide-learners-with-tablets

to also use these tablets for diagnostic assessment as part of formative assessment classroom practices.

On the basis of my study I consider that the following potential exists to extend and apply this work further:

- The data is collected at a fine-grained level from each individual learner in response to each test item, and this provides a massive bank of data which can be used as an ongoing research tool to support exploration for further misconceptions in a way not possible from existing approaches.
- The diagnostic assessments are analyzed 'on-the-fly' and provide the teacher with customized information about their individual learners' knowledge as soon as the tests are completed; allowing the teacher to adapt their instructional plans and practices. These instruments provide a quick insight into the learner thinking, and specifically on misconceptions that may be obstacles to their learning.
- Learners with persistent issues can be identified early and can be attended to with remedial work.
- Learners with similar misconceptions can be grouped to provide a benefit similar to one-to-one instruction.
- Further micro-domains in the rational numbers can be added and this approach may be extended to other topics in the mathematics curriculum that can be structured into conceptual bases to identify learning trajectories including the misconceptions that commonly occur and which are obstacles to learning.
- Data mining on the large banks of data can be used to identify further patterns of system usage, and can extend the knowledge of the learning trajectories and the Development Stages in the micro-domains.
- This approach to centralized data collection can provide an alternative to the expensive, and administratively challenged, Annual National Assessments, in providing better quality information, at lower-cost, and with almost zero administrative overhead.

8.6 Final Words

This is end of a long and difficult study in which I have had a number of personal challenges while conducting the research. I wish to note that I have not encountered anything that I would rather have done during the time I have spent on this study. I hope that my work may provide one small part of a solution for realizing a rapid and systemic improvement in mathematics knowledge so that, within a generation, South Africa can move up the international ranks, and that society as a whole can benefit from school-leavers who are far more mathematically proficient than is the case at present.

REFERENCES

Bart, W. M., Post, T., Behr, M. J., & Lesh, R. (1994). A Diagnostic Analysis of a Proportional Reasoning Test Item: An Introduction to the Properties of a Semi-Dense Item. *Focus on Learning Problems in Mathematics*, *16*(3), 1-11.

Behr, M., Lesh, R., Post, T., & Silver R. (1983). Rational Number Concepts. In R. Lesh & M. Landau (Eds.), *Acquisition of Mathematics Concepts and Processes* (pp. 91-125). New York: Academic Press.

Bell, A., Swan, N., Taylor, G. (1981). Choices of Operation in Verbal Problems with Decimal Numbers. *Educational Studies in Mathematics, 12*(4), 399-420.

Bejar, I. I. (1984). Educational Diagnostic Assessment. *Journal of Educational Measurement, 21*(2), 175-189.

Black, P., & Wiliam, D. (1998a). Assessment and Classroom Learning. *Assessment in Education: Principles, Policies & Practice*, 5(1), 7-74.

Black, P., & Wiliam, D. (1998b). Inside the Black Box. Raising Standard through Classroom Assessment. *Phi Delta Kappan, October 1998*, 139-148.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5-31. doi://10.1007/s11092-9068-5.

Bloom, B.S. (1984). The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Leadership*, May 1984, 4-17.

Bond, T., & Fox, C. (2012). *Applying the Rasch Model. Second Edition*. New York: Routledge.

Bright, G., Behr, M., Post, T., & Wachsmuth, I. (1988). Identifying Fractions on Number Lines. *Journal for Research in Mathematics Education, 19*(3), 215-232.

Brown, J.S., & Burton, R.R. (1977). *Diagnostic Models for Procedural Bugs in Basic Mathematical Skills*. Report No. 3699. ICAI Report No. 8. Bolt, Beranek and Newman Inc.

Brueckner, L. (1928a). Analysis of Errors in Fractions. *The Elementary School Journal*, *28*(10), 760-770.

Brueckner, L. (1928b). Analysis of Errors in Decimals. *The Elementary School Journal*, *29*(1), 32-41.

Buswell, G.T., & John, L. (1926). *Diagnostic Studies in Arithmetic. Supplementary Educational Monographs, No. 30*. Chicago: University of Chicago.

Ciofalo, J. F., & Wylie, C. E. (2006) *Using diagnostic classroom assessment: one question at a time.* Teachers College Record. Published: January 10, 2006: http://www.tcrecord.org ID Number: 12285.

Cobb P., Confrey, J., diSessa, A., Lehrer, R., Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher, 32*(1), 9-13.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education* (6th ed.). New York: Routledge.

Confrey, J. (1990). A Review of the Research on Student Conceptions in Mathematics, science and programming. *Review of Research in Education, 16*, 3-56.

Confrey, J. & Maloney, A. (2012). Next-Generation Digital Classroom Assessment Based on Learning Trajectories. In C. Dede, & J. Richards (Eds.), *Digital Teaching Platforms* (pp. 134-152). New York: Teachers College Press.

Confrey, J., Maloney, A., Nguyen, K., & Corley, D. (2012). *A Design Study of a Wireless Interactive Diagnostic Systems Based on a Mathematics Learning Trajectory*. AERA 2012 Conference, 13-17 April 2012. Vancouver: AERA.

Confrey, J., Rupp, A.A., Maloney, A.P., & Nguyen, K.H. (n.d.). *Developing a Learning Trajectory for Equipartitioning: Synthesis, Diagnostic Assessment Design, and Explanatory Item Response Modeling*. Provided by personal communication.

Cooke, D.H. (1931). Diagnostic and Remedial Treatment in Arithmetic. *Peabody Journal of Education, 9*(3), 143-151.

Cooke, D.H. (1932). Diagnostic and Remedial Treatment in Arithmetic II. *Peabody Journal of Education, 10*(3), 167-171.

Cromley, J. G., & Mislevy, R. J. (2004). *Task Templates Based on Misconception Research* (CSE Report No. 646). College Park, MD: University of Maryland.

DBE. (2011a). *National Curriculum Statement (Mathematics). Curriculum Assessment and Policy Statement (CAPS). Foundation Phase: Grades 1-3, Intermediate Phase: Grades 4-6, Senior Phase. Grades 7-9*. Department of Basic Education. Pretoria.

DBE. (2011b). *Action Plan to 2014. Towards the Realisation of Schooling 2025.* Department of Basic Education. Pretoria

DBE. (2012). *Education Statistics in South Africa 2010*. Department of Basic Education. Pretoria.

De Morgan, A. (1831/1898/2013). *On the Study and Difficulties of Mathematics*. Forgotten Books. Retrieved from http://www.forgottenbooks.com/books/On_the_Study_and_Difficulties_of_Mathematics_1000000807 on 29 December 2013.

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), Handbook of Statistics (Vol. 26, Psychometrics) (pp. 979–1027). Amsterdam: Elsevier.

DOE (2002). *Revised National Curriculum Statement Grades R-9 (Schools) : Mathematics*. Department of Education. Pretoria. (Government Gazette, Vol.443. No. 23406).

DOE. (2004). *White Paper on e-Education. Transforming Learning and Teaching through Information and Communication Technologies (ICTs)*. Department of Education. Pretoria. (Government Gazette Notice No. 1922)

DIAGNOSER. (2012). www.diagnoser.com. Last accessed 9 October 2012.

Dunne, T., Long, C., Craig, T., Venter, E. (2012). Meeting the requirements of both classroom-based and system assessment of mathematics proficiency: The potential for Rasch measurement theory. *Pythagoras, 33*(3), http://dx.doi.ord/10.4102/pythagoras.v33i3/19.

Empson, S. (2011). On the idea of Learning Trajectories: Promises and Pitfalls. *The Mathematics Enthusiast*, *8*(3), 571-596.

Feischbein, E., Deri, M., Nello, M. S., & Marino, M. S. (1985). The Role of Implicit Models in Solving Verbal Problems in Multiplication and Division. *Journal for Research in Mathematics Education, 16*(1), 3-17.

Gardner, J. (2012). Assessment and learning: Introduction. In J. Gardner (Ed.), *Assessment and Learning* (2nd ed., pp. 1-8). London: Sage Publications.

Gillings, R. (1982). *Mathematics in the Time of the Pharaohs*. New York: Dover.

Graf, E., (2008). *Approaches to the Design of Diagnostic Item Models*. Educational Testing Services. Report ETS-RR-08-07.

Griffin, P. (2009). Teacher's Use of Assessment Data. In C. Wyatt-Smith & J.J. Cummings (eds.). *Educational Assessment in the 21st Century* (pp. 183-208). Dordrecht: Springer.

Grossman, A. (1983). Decimal Notation: An Important Research Finding. *The Arithmetic Teacher*, *30*(9), 32-33.

Guiler, W. (1945). Difficulties in Fractions Encountered by Ninth-Grade Pupils. *The Elementary School Journal*, *45*(3), 146-156.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer.

Hergenhahn, B., & Olson, M. (2005). *An Introduction to Theories of Learning*. (7th ed.). New Jersey: Pearson Education.

Hiebert, J., & Wearne, D. (1985). A Model of Students' Decimal Computation Procedures. *Cognition and Instruction. 2*(3/4), 175-205.

Hiebert, J., & Tonnessen, L.H. (1978). Development of the Fraction Concept in Two Physical Contexts: An Exploratory Investigation. *Journal for Research in Mathematics Education*, *9*(5), 374-378.

Howell, D. C. (1999). *Fundamental Statistics for the Behavioural Sciences*. (4th ed.). London, England: Duxbury Press.

Huntley, B. (2008). *Comparing different assessment formats in undergraduate mathematics*. (Doctoral thesis). Retrieved from http://upetd.up.ac.za/thesis/available/etd-01202009-163129/.

IEA. (2007). *TIMSS2003 Mathematics Items. Released Set Fourth Grade. International Association for the Evaluation of Educational Achievement (IEA)*. Retrieved from http://timss.bc.edu/PDF/T03_RELEASED_M4.pdf.

IEA. (2013). *TIMSS 2011 User Guide for the International Database. Released Items. Mathematics – Eighth Grade. TIMSS & PIRLS International Study Centre*. Boston College. Chestnut Hill:MA.

Kehoe, J. (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation*, *4*(9),

Kieren, T. (1976) On the mathematical, cognitive, and instructional foundations of rational numbers. In R. Lesh (Ed.), *Number and measurement: Papers from a research workshop* (pp.101-104). Columbus, OH: ERIC/SMEAC.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.) (2001) *Adding it Up: Helping Children Learn Mathematics***.** Washington DC: National Academy Press.

Kluger, A., & DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. *Psychological Bulletin, 119*(2), 254-284.

Leahy, S., & Wiliam, D. (2012). From teachers to schools: Scaling up professional development for formative assessment. In J. Gardner (Ed.), *Assessment and Learning* (2nd ed., pp. 49-71). London: Sage Publications.

Leighton, J.P., & Gierl, M.J. (2007a). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. New York: Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (2007b). *Why Cognitive Diagnostic Assessment?* In J. P. Leighton & M. J. Gierl (Eds.), Cognitive Diagnostic Assessment for Education: Theory and Applications (pp. 3-18). New York: Cambridge University Press.

Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: a critical appraisal. *Learning and Instruction, 11*, 357-380.

Linacre, J.M. (2002). What do Infit and Outfit Mean-Square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J.M. (2013). Winsteps® Rasch Measurement Computer Program and User Guide. Beaverton, Oregon: Winsteps.com.

Long, C., Dunne, T., & Craig, T. (2010). Proficiency in the multiplicative conceptual field: using Rasch measurement to identify levels of competence. *African Journal of Research in MST Education, 14*(3), 79-91.

Mack, N.K. (1990). Learning Fractions with Understanding: Building on Informal Knowledge. *Journal for Research in Mathematics Education*, *21*(1), 16-32.

Mack, N.K. (1995). Confounding Whole-Number and Fraction Concepts When Building on Informal Knowledge. *Journal for Research in Mathematics Education*, *26*(5), 422-441.

Matters, G. (2009). A Problematic Leap in the Use of Test Data: From Performance to Inference. In C. Wyatt-Smith & J.J. Cummings (Eds.). *Educational Assessment in the 21st Century* (pp. 209-226). Dordrecht: Springer.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education.

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, *50*(9), 841-749.

Metaphysics Research Lab. (2014). *Simplicity*. Metaphysics Research Lab, CSLI, Stanford University. http://plato.stanford.edu/entries/simplicity/

Monroe, W. S. (1917). The Ability to Place the Decimal Point in Division. *The Elementary School Journal, 18*(4), 287-293.

Moss, P. (1994). Can There be Validity without Reliability? *Educational Researcher, 23*(3), 5-12.

Mouton, J. (2001). *How to succeed in your master's & doctoral studies*. Pretoria: Van Schaik.

Movshovitz, N., Zaslavsky, O., & Inbar, S. (1987). An Empirical Classification Model for Errors in High School Mathematics. *Journal for Research in Mathematics Education, 18*(1), 3-14.

Mullis, I.V.S. Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J. & Smith, T.A. (2000). *TIMSS 1999 International Mathematics Report. Findings from IEA's Repeat of the Third International Mathematics Science Study at the Eighth Grades*. International Study Center, Lynch School of Education, Boston College. Retrieved from http://timss.bc.edu/timss1999i/math_achievement_report.html.

Mullis, I., Martin, M., Foy, P., & Arora, A. (2011). TIMSS 2011 International Results in Mathematics. TIMSS & PIRLS International Study Center. Boston College. Boston:MA.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Chrostowski, S.J. (2004). *Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from http://timss.bc.edu/PDF/t03_download/T03INTLMATRPT.pdf.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington D.C: National Governors Association Center for Best Practices, Council of Chief State School Officers. http://www.corestandards.org/Math/.

National Research Council. (2005). *How Students Learn: Mathematics in the Classroom*. Committee on How People Learn, A Targeted Report for Teachers, M.S. Donovan and J.D. Bransford, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NCES (2015). Mathematics Concepts Mathematics Items. TIMSS Grade 8 Mathematics Released Item Results. Downloaded from http://nces.ed.gov/timss/pdf/TIMSS1999_G8_Math.pdf.

Neal, E.A., & Foster, I. (1928). An Experiment with Remedial Work in Common Fractions. *The Elementary School Journal*, *29*(4), 280-282.

Nesher, P. (1987). Towards an Instructional Theory: The Role of Student's Misconceptions. *For the Learning of Mathematics, 7*(3), 33-40.

Ni, Y. (2000). How Valid is it to Use Number Lines to Measure Children's Conceptual Knowledge about Rational Number? *Educational Psychology, 20*(2), 139-152. http://dx.doi.org/10.1080/713663716.

Novillis-Larson, C. (1980). Locating Proper Fractions on Number Lines: Effect of Length and Equivalence. *School Science and Mathematics, 80*, 423-428.

OECD (2010), *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science. (Volume I)*. Retrieved from http://dx.doi.org/10.1787/9789264091450-en.

Olander, H.T. (1933). The Need for Diagnosis. *The Elementary School Journa*l, *33*(10), 736-745.

Olivier, A. (1989). *Handling Pupils' Misconceptions*. Thirteenth National Convention on Mathematics, Physical Science and Biology Education, Pretoria, 3-7 July 1989. Retrieved from http://academic.sun.ac.za/mathed/MALATI/Misconceptions.htm.

Pearn, C., & Stephens, M. (2007). Whole Number Knowledge and Number Lines Help to Develop Fraction Concepts. In J. Watson & K. Beswick (Eds.), *Mathematics: Essential Research, Essential Practice, Proceedings of the 30th Annual Conference of Mathematics Education Group of Australia* (pp. 601-610). Adelaide: MERGA.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Committee on the Foundations of Assessment, Board on Testing and Assessment, Center for Education, National Research Council. Washington, DC: National Academy Press.

Piaget, J. (1964/2003) Development and Learning. *Journal of Research in Science Teaching*, *2*(3), 176-186 (Reprinted in V40, S8-18, 2003).

Piaget, J. (1970). *Genetic Epistemology*. New York: WW Norton & Company Inc.

Piaget, J. (1985). *Equilibration of Cognitive Structures*. University of Chicago Press. Chicago.

Polya, G. (1945/1990). *How to solve it*. Penguin Books, Princeton University Press. Harmondsworth, Middlesex, England.

Popper, K. (2002). *The Logic of Scientific Discovery*. Routledge, London. Originally published 1959 in English.

Posner, G., & Gertzog, W. (1982). The Clinical Interview and the Measurement of Conceptual Change. *Science Education, 66*(2), 195-209.

Price, B., Stacey, K., Steinle, V., Chick, H., & Gvozdenko, E. (2011). Getting SMART about Assessment for Learning in 2011. *Reflections*, *36*(3), 3-7.

Radatz, H. (1979). Error Analysis in Mathematics Education. *Journal for Research in Mathematics Education*, *10*(3), 163-172.

Rational Number Project. (2014). Retrieved from http://www.cehd.umn.edu/ci/rationalnumberproject/ on 30 December 2014

Resnick, L.B., Nesher, P., Leonard, F., Magone, M., Omansom, S. & Peled, I. (1989). Conceptual Bases of Arithmetic Errors : The Case of Decimal Fractions. *Journal for Research in Mathematics Education*, *20*(1), 8-27.

Robertson, M.S. (1924). A Diagnostics Fractions Test. *Peabody Journal of Education*, *1*(6), 301-310.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: The Guildford Press.

Sackur-Grisvard, C., & Leonard, F. (1985). Intermediate Cognitive Organisations in the Process of Learning a Mathematical Concept: The Order of Positive Decimal Numbers. *Cognition and Instruction, 2*(2), 157-174

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Siegler, R, Thompson, C., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*, 273-296.

Simatwa, E. (2010). Piaget's theory of intellectual development and its implication for instructional management at presecondary school level. *Educational Research and Reviews, 5*(7), 2010.

Simon, M. (1995). Reconstructing Mathematics Pedagogy from a Constructivist Perspective. *Journal for Research in Mathematics Education*, *20*(2), 114-145.

Smith, D.E. (1958). *History of Mathematics, Vol. II*. Dover: New York.

Smith, J.P. (1995). Competent Reasoning with Rational Numbers. *Cognition and Instruction*, *13*(1), 3-50.

Smith, J.P., diSessa, A.A., & Roschelle, J. (1993). Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition. *The Journal of the Learning Sciences*, *3*(2), 115-163.

South African Treasury (2010). Budget Review 2010/11. Retrieved from http://www.treasury.gov.za/documents/national%20budget/2010/review/Budget%20Review.pdf.

Sprague, J.B. (1939). Diagnostic Testing to Improve Mathematical Ability in Grade X. *The School Review*, *47*(6), 431-438.

Stacey, K., Helme, S., & Steinle, V. (2001). Confusions between decimals, fractions and negative numbers: A consequence of the mirror as a conceptual metaphor in three different ways. In M.v.d. Heuvel-Panhuizen (Ed.), *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 217-224). Utrecht: PME.

Stacey, K. (2005). Travelling the Road to Expertise: A longitudinal Study of Learning. In Chick, H.L. & Vincent J.L. (Eds.). *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Educations* (Vol. 1, pp 19-36). Melbourne: PME.

Stacey, K. (2013). Bringing Research on Students' Understanding into the Classroom through Formative Assessment. In V. Steinle, L. Ball & C. Bardini (Eds.), *Mathematics Education: Yesterday, today and tomorrow (Proceedings of the 36th annual conference of the Mathematics Education Research Group of Australia)* (pp. 13-21). Melbourne: MERGA

Stacey, K., Price B., & Steinle, V. (2012). Identifying Stages in a Learning Hierarchy for Use in Formative Assessment – the Example of Line Graphs. In J. Dindyal, Cheng L-P, Ng, S-F (Eds.), *Mathematics Education: Expanding Horizons, Proceedings of the 35th Annual Conference of Mathematics Education Group of Australia* (pp. 393-400). Adelaide: MERGA.

Stacey, K. & Steinle, V. (2006). A Case of the Inapplicability of the Rasch Model: Mapping Conceptual Learning. *Mathematics Education Research Journal*, *18*(2), 77-92.

Stafylidou, S. & Vosnaidou, S. (2004). The development of students' understanding of the numerical value of fractions. *Learning and Instruction, 14*, 503-518.

Steffe, L. (2000). *Radical Constructivism in Action: Building on the Pioneering Work of Ernst von Glaserfeld*. Routledge Falmer: New York.

Steinle, V. (2004a). *Changes with age in students' misconceptions of decimal numbers*. (Unpublished doctoral thesis, University of Melbourne). http://eprints.unimelb.edu.au/archive/00001531/

Steinle, V. (2004b). Detection and remediation of decimal misconceptions. In B. Tadich, S. Tobias, C. Brew, B. Beatty, & P. Sullivan (Eds.), *Towards excellence in mathematics* (pp. 460-478). Brunswick: The Mathematical Association of Victoria.

Steinle, V., & Stacey, K. (2005). Analysing Longitudinal Data on Students' Decimal Understanding using Relative Risks and Odds Ratios. In H. Chick, J. Vincent (Eds), *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 877-911). Melbourne: PME.

Steinle, V., & Stacey, K. (2012). *Teachers' Views of using an on-line, formative assessment system for Mathematics*. Paper presented at the 12th International Congress on Mathematical Education: Topic Study Group 33. July 8 to 15, 2012. pp. 6721–30. Seoul, Korea: COEX.

Tatsuoka, K.K. (1983). Rule Space : An Approach for Dealing with Misconceptions Based on Item Response Theory. *Journal of Educational Measurement, 20*(4), 345-354.

Tatsuoka, K.K., & Tatsuoka, M.M. (1997). Computerized Cognitive Diagnostic Adaptive Testing: Effect on Remedial Instruction as Empirical Evaluation. *Journal of Educational Measurement, 34*(1), 3-20.

Uhl, W.L. (1917). The Use of Standardised Materials in Arithmetic for Diagnosing Pupils' Method of Work. *The Elementary School Journal*, *18*(3), 215-218.

Usiskin, Z. (1979). The Future of Fractions. *The Arithmetic Teacher, 26*(5), 18-20.

Usiskin, Z. (2005). The Importance of the Transition Years, Grades 7-10, In School Mathematics. *UCSMP Newsletter 33*. University of Chicago School Mathematics Project.

Vinner, S., Hershkowitz, R., & Bruckheimer, M. (1981). Some Cognitive Factors as Causes of Mistakes in the Addition of Fractions. *Journal for Research in Mathematics Education*, *12*(1), 70-76.

Vergnaud, G. (1994). Multiplicative Conceptual Field: What and Why? In G. Harel, J.Confrey (Eds.), *The Development of Multiplicative Reasoning in the Learning of Mathematics* (Chapter 2, pp. 41-59). New York, NY: State University of New York.

Vergnaud, G. (2009). The Theory of Conceptual Fields. *Human Development, 52*, 83-94.

Vygotsky, L.S., (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University, Cambridge, MA.

Wearne, D., & Hiebert, J. (1988). A Cognitive Approach to Meaningful Mathematics Instruction: Testing a Local Theory Using Decimal Numbers. *Journal for Research in Mathematics Education*, *19*(5), 371-384.

Webb, J. M., Stock, W. A., & McCarthy, M. T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Contemporary Educational Psychology*, *19*, 251-265.

Wiliam, D. (2011a). What is assessment for learning? *Studies in Educational Evaluation*, *37*, 3-14.

Wiliam, D. (2011b). *Embedded Formative Assessment*. Bloomington: Solution Tree Press.

Wilson, P.H., Mojica, G.F., & Confrey, J. (2013). *Learning Trajectories in Teacher Education: Supporting Teachers' Understandings of Students' Mathematical Thinking*. The Journal of Mathematical Behaviour, *32*, 103-121.

Wright, B. (1997). A History of Social Science Measurement. *Educational Measurement – Issues and Practice, Winter 1997*, 33-45.

Wright, B. ([2005]). Rasch Dichotomous Model vs. One-Parameter Logistic Model (1PL 1-PL). Downloaded from www.rasch.org/rmt/rmt193h.htm.

Wylie, C., & Wiliam, D. (2006, April). *Diagnostic questions: is there value in just one?* Paper presented at the Annual Meeting of the National Council on Measurement in Education. April 6 to 12, 2006. San Francisco, CA.

World Wide Web Consortium. (2008). *Extensible Markup Language (XML) 1.0* (Fifth Edition). World Wide Web Consortium. http://www.w3.org/TR/2008/REC-xml-20081126/

World Wide Web Consortium. (2010). *Mathematical Markup Language (MathML)* Version 3.0. http://www.w3.org/TR/2010/REC-MathML3-20101021/. World Wide Web Consortium.

World Wide Web Consortium. (2011). *Scalable Vector Graphics (SVG) 1.1* (second Edition). http://www.w3.org/TR/2011/REC-SVG11-20110816/. World Wide Web Consortium.

APPENDIX A : RESULTS OF DATA GATHERING

A large amount of data was collected in this study, when measured by the number of database records in the database. This database is divided into three sections:

- Firstly, the base data about the schools and learners (provided in codes rather than in names), with the lesson structures used by the online-system.
- Secondly, the data captured from the pretest, which provides the complete set of learner responses to each test item, and also contains the interpretation of the learner responses to the constructed-format questions.
- Thirdly, the data captured automatically on the server from the online-tests, which consists of 13,178 records each of which is a single response of a learner to a test item presented.

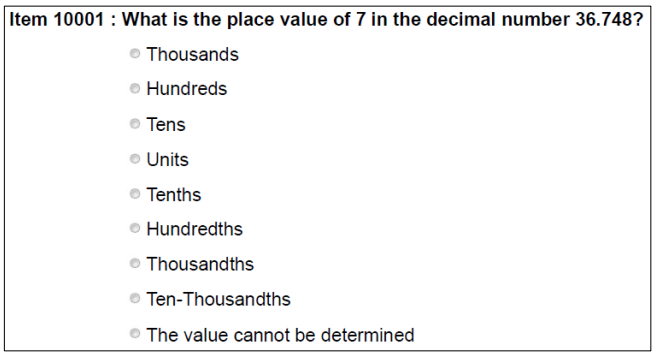
Each record from the database from the online tests consists of the following fields:

Table 42. Database fields for online test responses

Field	Type	Usage Notes	Sample Data
SCHOOL	Code	The school from which this record was drawn, being either A or B.	A
CLASS	Code	The class within the school, which was either 7 or 8 indicating the class Grade. This is unused for most of the study.	7A
LESSON	Number	The lesson number for the weekly online lessons, being 1-4.	1
TESTSEQ	Number	The test number within the lesson, with there being a number of tests deployed within each of the lessons. These are not sequential since other activities are also used, such as the display of supporting information.	3
ITEMSEQ	Number	The sequence of the test item within the test within the lesson.	1
USERCODE	Text	The user code that was created for this analysis, that removes all references to school and the individual learners.	A01
DATEOPEN	Date/Time	The date/time when the test item was initially presented to the learner, and displayed on the Web page.	2009/05/28 01:39:33 PM

Field	Type	Usage Notes	Sample Data
DATECLOSED	Date/Time	The date/time when the learner completed the test item, including all required fields such as the difficulty fields. From the DATEOPEN and DATECLOSED the duration of this response can be calculated in seconds.	2009/05/28 01:40:01 PM
ANSWER	Number	The response given by the learner to this test item.	2
ISEASY	True/False	True if the learners marked this as Easy.	0
ISJUSTRIGHT	True/False	True if the learner marked this test item as Just Right.	-1
ISDIFFICULT	True/False	True if the learner marked this test item as Difficult.	0
ISNOTUNDERSTOOD	True/False	True if the learner marked this test item as not being understood.	0
TESTITEMTYPE	Number	The type of test item, from a predefined set of types, which are explained below.	1
NUMANSWERS	Number	How many choices there are for this test item. This is copied in from the ItemBank table.	9
CORRECTANS	Number	The correct choice, for this item, being the one which is deemed to be correct, even if others are also partially correct.	5
ITEMBANKNUM	Number	The reference to the Item Bank number appears as Item# on the various diagrams in this thesis, as well as in the dump of the items given in the Appendix. These item bank numbers start from 10001.	10001
XML	Long Text	The full text of the question itself, including all of the HTML, SVG and MathML as outlined in the appendix.	<Problem><Question>What is the place-value of 7 in…
ERRORNOTE	Text	Any issues that were noted in the item description, form, or usage that needed to be considered when this was being analysed.	

For each of the test items, identified by the ITEMBANKNUM, there is an additional data table which marks every choice in the MCQ test items as either being the CORRECT choice, or being one of many possible MISCONCEPTIONS that would give rise to this response. The choices are numbered in terms of the sequence of the choice, counted from 1 for the first choice, rather than the text of the choice presented to the learners.



For example, for Item 10001, choice 1 is “Thousands”, choice 2 is “Hundreds” etc… and in this case choices 2 (“Hundreds”) and 6 (“Hundredths”) are marked as the WHOLE misconception, while choice 5 (“Tenths”) is marked as CORRECT.

I extracted the data for the analyses in Chapter 6 by selecting the items that I needed for a particular type, using the TESTITEMTYPE field, where a value of 1 (as in the above sample data) is the PV1 type, which is the first type of the place-value test items. I also matched the learner choice to the particular conception which I was examining, such as the WHOLE misconception. As required within the various steps of processing I removed certain Items or User Codes if these were misfitting to the Rasch model.

APPENDIX B : PRETEST

This appendix provides the details of the pretest and a short explanation of the source of the individual questions used. The pretest was conducted for each of the two schools in this study, and was administered prior to the online assessments.

The purpose of the pretest was to help to identify test items to use in the online assessments. The pretest consisted of a number of test items, some of which had been used in prior work on misconceptions in the rational numbers. All of the test items used in the pretest are short and simple, and were selected for their diagnostic value rather than to establish learner proficiency. The formalization of this "diagnostic value" was one of the primary goals of this study, following from the work of others (Bart et al., 1994; Huntley, 2008) who have investigated the nature of "good" mathematical questions.

Each of the questions used in the pretest also asked the learners whether they found the test item Easy, Just Right, or Difficult.

I confine my explanation to the motivation for inclusion as well as the provenance of each type of question included. In some cases, I refer to the TIMSS test items, and refer to TIMSS 1999 (Mullis et al., 2000; NCES, 2015) and TIMSS 2003 (Mullis et al., 2004; IEA, 2007).

Introductory Notes.

The following notes were provided on the question paper prior to the individual questions.

USER CODE: _____ **DATE: _____**

Please enter ONLY the User Code you have been assigned by your teacher and do NOT write your name on this test paper.

This test is part of a research project being undertaken by Roger Layton as part of PhD in Mathematics Education at the University of the Witwatersrand.

This test concerns your knowledge of mathematics and includes questions about the Rational Numbers.

This test does not play any part in your year mark, since the results will be used for research purposes only.

You have 20 minutes to complete the test.

INSTRUCTIONS:

1. Do not turn over the page until your teacher tells you to start.
2. Please do all of your workings on this paper.
3. Write the answer in the place provided for each question unless the answer is required to be placed elsewhere.
4. Indicate whether you found the question Easy, Difficult, or Just Right for your own level of knowledge of rational numbers by placing an X into the block next to this statement…
   For example **[X] Easy**.
5. If you do not know the answer, then write D/K (Don't Know) in the ANWER block.
6. If you run out of time, leave the rest of the answer blocks empty.

Question 1: Ordering of Common Fractions

The first question in the pretest was at a lower level than the age of the learners in the study, and it concerned the selection of the larger of two small simple common fractions. From the National Curriculum Statement this is a Grade 4 competence:

> Recognizes and represents the following numbers in order to describe and compare them: …. common fractions with different denominators including halves, thirds, quarters, fifths, sixths, sevenths, and eighths; … (DOE, 2002, p.40).

Question 1 : Which is larger: $\frac{1}{4}$ **or** $\frac{1}{6}$ **?**

ANSWER:_____ [ ] Easy, [ ] Difficult, [ ] Just Right

I used this first question to explain to the learners how they should answer the question, by writing their answer in right place, and to let me know whether they found this question Easy, Difficult or Just Right, by placing a cross in the block on the right. I asked the learners to do any workings on the flip side of the question paper.

The original intention of asking the learners about their perception on the difficulty of the question was to determine whether this can be used to isolate misconceptions, on the basis that if a question is indicated as Easy or Just Right and this is answered incorrectly, then this may provide additional evidence of a systematic error and a possible misconception, including situations in which this is not within the set of

misconceptions known in advance. This information also might provide useful information from the learners who identified a question as Difficult and who selected the correct answer, which might mean that they were guessing.

Question 2: Estimation of Value of Common Fractions

This question is based upon an example mentioned by Kilpatrick et al. (2001), asking the learner to select the closest to a common fraction sum which is too difficult to be worked out by hand.

For the pretest the following question was posed

Question 2 : Which is the closest to the sum $\frac{7}{8}+\frac{12}{13}$

a. 1
b. 2
c. 19
d. 21
e. 40

This question requires a conceptual understanding of common fractions, including their notation and magnitude. The two common fractions in this sum are both close to 1, and thus the sum is close to 2 (answer b).

Question 3: Place-Value

Place-value is at the core of all work in decimal numbers, and this particular question was included to quickly identify the level of knowledge of the learners. This question is within the curriculum expectations for Grades 7 and 8.

Question 3 : What is the place value of the digit 7 in the number 0.06758

a. Tens
b. Units
c. Tenths
d. Hundredths
e. Thousandths
f. Ten-thousandths

Question 4: Decimal number ordering

Which of the following is the smallest?
A. 0.25
B. 0.125
C. 0.5
D. 0.675
E. 0.375

This type of question concerning comparison of decimal numbers has been widely used in rational number research (Sackur-Grisvard & Leonard, 1985; Resnick et al., 1989; Steinle, 2004b), and is also referred to in Kilpatrick et al. (2001). This particular example is cited in various places in this thesis, and it has proven value to help expose a number of different misconceptions.

Question 5: Decimal notation

Question 5 : Which is the decimal representation of the number "two hundred and six and nine tenths"?
A. 206.90
B. 206.910
C. 206.09
D. 206+9/10
E. 206.9
F. 2006 9/10
G. 20069.10

This question was introduced to determine the ability of the learners to convert between a word description of a decimal fraction and the decimal representation. A similar type of question was included into the TIMSS 2003 public question sets (IEA, 2007).

Question 6: Fractions and percentages

TIMSS 2003 (IEA 2007) included questions concerning the relationship between fractions and percentages, and this proficiency is included within the assessment criteria in the National Curriculum Statement (DBE, 2011a) concerning conversions between the various types of rational numbers. Question 6 was introduced to detect particular misconceptions in the meaning of the fractions and percentages.

Question 6 : What percentage is equivalent to the fraction $\frac{3}{4}$?

a. 3%
b. 4%
c. 34%
d. 50%
e. 75%
f. 100%
g. 133.33%

I am particularly interested in the number of learners who selected choice c, considering that 34% is the percentage equivalent of the common fraction $\frac{3}{4}$.

Question 7: Equivalent fractions

This question was introduced to identify whether the learner knows how to find equivalent fractions. This was intended to detect various possible errors, such as the selection of 15 (choice f) since 4-3 = 1 and 16-15 = 1.

Question 7 : What is the value of x if $\frac{3}{4} = \frac{x}{16}$

a. 3
b. 4
c. 7
d. 12
e. 13
f. 15

Question 8, 9: Decimal addition and subtraction

A number of early studies in learner errors explored how learners would react to questions where there is no whole number prior to the decimal fraction. Whereas these are valid fractions they are not in common usage, with modern representation using the zero prior to the fraction such as 0.25 rather than merely .25

This question was posed in two ways, with question 8 being a multiple choice format, and question 9 being constructed-response.

Question 8 : What is the result of $.25 + .4$

a. .29
b. .65
c. 4.25
d. 25.4
e. 6.5

Question 9 : What is the result of 7 - .4

Question 10: Place-value estimation

This question was derived from early work in identifying learner errors in which the learners were asked to position a decimal point in a given answer to complete the answer. For example, the first of the two sub-questions asks the learners to multiple 657 by .7 and the answer is given as 46004. The learner is required to put the decimal point into this answer, and the correct answer is 460.04. I made an error in the wording for this question, since it should indicate "the decimal *point* is missing", rather than the decimal place.

Question 10 : The following arithmetic calculations have been worked out, but the decimal place is missing in the answer. Place the decimal point correctly into the following two answers:
(a) 657 x .7 = 46004
(b) 16.2 ÷ 3 = 54

Question 11: Decimal estimation

This question has also been derived from the public set made available from the TIMSS 2003 study (IEA, 2007), and it provides a useful measure of understanding of the decimal numbers.

This question requires the learner to inspect the question and to select the most appropriate and closest answer. The expectation was that some learners would perform the addition in full and then compare the result as the value of the sums in the individual choices. By doing this they would find that there was more than one alternative for a correct choice. The sum is 10.96 which is close to 11, and there are three choices which evaluate to 11 being choices b, c, and e. My intention was that a competent learner would select choice c as the closest, given that the individual numbers that comprise the addition sum are also close to the corresponding numbers in the choice, with 6.91 being close to 7.00 and 4.05 being close to 4.00.

Question 11 : Which of the following is closest to the sum 6.91 + 4.05
a. 6.00 + 4.00
b. 6.00 + 5.00
c. 7.00 + 4.00
d. 7.00 + 5.00
e. 8.00 + 3.00

Questions 12, 13, 14: Common fraction density

These three questions explore the learner's understanding of density of the common fractions, where density is the property that there are always fractions that exist between two other fractions. For example, between $\frac{3}{5}$ and $\frac{4}{5}$ there is a fraction $\frac{7}{10}$, even though there is no whole number between 3 and 4, since the whole numbers do not have the property of density.

Question 12 asks the learner to find an equivalent fraction to the one provided. Question 13 and 14 ask for new fractions between the given fraction and 1, however, for Question 14 there is no further fraction available with the same denominator and it is necessary to change the denominator.

Question 12 : Write down a fraction that is equivalent to $\frac{3}{8}$

Question 13 : Write down a fraction that is larger than $\frac{2}{7}$ and less than 1

Question 14 : Write down a fraction that is larger than $\frac{3}{4}$ and less than 1

Question 15: Decimal ordering

The final question in the pretest concerns the decimal number ordering for the case of selecting the smallest of pairs of numbers

Question 15: Place a tick against the smallest number in each of the following pairs.

A.	0.45	[ ]	0.39	[ ]
B.	0.4	[ ]	0.39	[ ]
C.	0.45	[ ]	0.3	[ ]
D.	5.45	[ ]	5.39	[ ]
E.	0.453	[ ]	0.3	[ ]
F.	0.398	[ ]	0.3	[ ]
G.	8.4	[ ]	8.3	[ ]
H.	7.45	[ ]	7.9	[ ]
I.	3.45	[ ]	3.33	[ ]
J.	0.45	[ ]	0.08	[ ]

APPENDIX C :
WEB-BASED IMPLEMENTATION

This appendix provides details of how the Web-based assessment was implemented, using examples to illustrate how this was designed, implemented technically, and presented to the learners.

Multiple-Choice Question (MCQ) Structure

I made a decision early in the research process to use only MCQs: firstly, due to their ease of gathering data; secondly, due to the extensive prior research on this form of question, thirdly, their suitability for additional analysis through Rasch Analysis; and finally, because of their widespread usage in computerized assessment.

In my implementation of MCQ each test item was composed of four distinct elements:

- a question which is posed;
- a set of possible answers;
- a place to enter the answer, and
- a question on how difficult the learner found this test item.

These test items were required to be presented using the Web, and this had two major challenges resulting from the nature of typical mathematical questions: firstly, the handling of mathematical notation on the Web, and secondly the usage of graphical elements as part of a question, such as for the number line.

Representing Mathematics on the Web using MathML

One of the key challenges in presenting mathematical notation on the Web has been the range of standards for mathematical representation and notation. Within the domain of the rational numbers, it is common that both a questions and its possible answers will include a number of notations and conventions, many of which are specific to mathematics, such as the structure of the common fraction. For example, the fraction for "three-quarters" is represented best as $\frac{3}{4}$. This is also commonly represented by 3/4 or ¾ but these are less adequate for mathematical notation. Figure 43. Example of Web mathematics requirement presents this challenge of displaying fractions on the Web.

To accommodate this requirement to provide true mathematical notation in a Web browser I needed to make a decision on whether I should create image files (such as using the JPG or PNG formats) for each of the mathematical expressions I was using, using an equation editor such as that provided with Microsoft Word. This approach of using such image files was commonly used in most Web-based mathematics at the time I commenced this study. It is a time-consuming task to create these image files, and they are not scalable as the page is resized. They are also not easily changed if there is a need to modify the Web page.

At the time of preparing the test items for the Web there was an emerging standard for mathematical markup called MathML, the Mathematics Markup Language (World Wide Web Consortium, 2010). MathML holds the promise of enabling mathematical expressions to be included into standard HTML Web pages and presented to the users in a familiar mathematical language. I made the decision to use MathML as the basis for all mathematical expressions in my test items, while I was simultaneously aware of the limitations of MathML support within many Web Browsers. For instance, Internet Explorer, the most popular Web browser at the time of this study, did not support MathML implicitly, and the MathML add-ins available for Internet Explorer Web browser were not easy to implement. However, the FireFox Web browser had a long history of support for MathML, and thus I decided on using MathML but imposed an additional constraint on the computer laboratories in the study schools that every computer being used must have the FireFox Web browser installed.

For the first school in the study the FireFox browser was already in use as the default, with the computers running on the Linux / Ubuntu operating system. However, for the second school this was a challenge, with Internet Explorer being the predominant Web browser, and a special setup was required to ensure that FireFox was available prior to the start of the online assessments.

Representing Graphics on the Web using SVG

A second technical consideration concerned the display of graphical items that commonly appear in questions on fractions, such as those within the public item banks of TIMSS 2003 (IEA, 2007). I planned to use this type of question in the tests and needed to incorporate both graphical images and mathematical notations. I illustrate this requirement using a problem from my item bank.

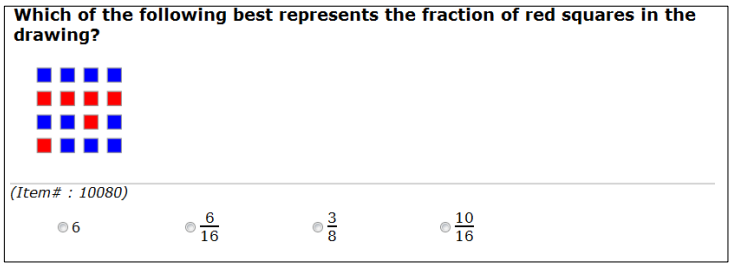


Figure 43. Example of Web mathematics requirement

Similar to the challenges with the presentation of mathematical notation, there were corresponding challenges in the handling of graphical structures. A similar decision was required for whether to generate image files in JPG or PNG for these graphical questions, or to use modern standards for representing graphic directly within a Web browser.

All Web browsers support the incorporation of images within Web pages, but this requires that all of the images must be pre-built using standard image formats. This can be done using software programs such as Adobe Photoshop or Microsoft Paint. I wanted this to be the option of last resort due to the effort required in setting up each of the questions, and the lack of flexibility if there was a need to change these images during the course of the study.

As an alternative I considered the SVG language (Scalar Vector Graphics) (World Wide Web Consortium, 2011), which has been defined as a World Wide Web Consortium Recommendation[7]. SVG is a language to define graphical elements and objects directly within HTML. The benefit of SVG is that it can be generated by other computer programs or by hand, and provides smaller sizes of files for transfer than the relatively large size of image files.

[7] A "World Wide Web Consortium Recommendation" is the highest level of standards provided by the World Wide Web Consortium in terms of how the Web is structured. However, since the World Wide Web Consortium is not an official standard-setting body these documents cannot be referred to as "standards".

Once again, I discovered an inherent limitation in the range of Web browsers that support SVG, and I discovered that the most popular browser, Internet Explorer, also does not provide support for SVG, whereas FireFox supports SVG implicitly. With the previous decision to use the FireFox browser exclusively for this study this was already taken care of in the set up for the study.

Access to the Web-Based Assessment Program

Each learner was given a logon code and a private password at the start of the research programme. These were handed out at the time of the pretest and these codes were then entered onto the paper pretests as one element of my implementation of privacy. This process of handing out the user codes was conducted by the school personnel, and I was not provided with the lists of which codes corresponded to which learner. The learners were requested not to share their password with others. The user codes were constructed by combining a prefix representing the school followed by a unique sequential number. However, it was only after the school codes were selected that it became evident that the school name could possibly be inferred from the short code used. As a result, the school codes have been changed within this thesis to reflect the requirement for anonymity and confidentiality that I had agreed with the school principals.

The Web site required the users to log on at the start of each of the weekly assessment sessions. The Web site also preserved the current status of the testing, so that if a learner logged off and then logged on again within a single session they would continue from where they left off.

Preparing the Test Items for the Web Assessment

All of the test items have been coded using XML (Extensible Markup language; World Wide Web Consortium, 2008) as follows:

```
<Problem>
        <Question>What is the place value of 0 in the decimal number 3.408?</Question>
        <Answers>
                <Answer Seq="1">Thousands</Answer>
                <Answer Seq="2">Hundreds</Answer>
                <Answer Seq="3">Tens</Answer>
                <Answer Seq="4">Units</Answer>
                <Answer Seq="5">Tenths</Answer>
                <Answer Seq="6">Hundredths</Answer>
                <Answer Seq="7">Thousandths</Answer>
```

```
                <Answer Seq="8">Ten-Thousandths</Answer>
                <Answer Seq="9">The value cannot be determined</Answer>
        </Answers>
</Problem>
```

Figure 44. Example of markup of a test item for the Assessment Markup Language

This XML test item is stored in a relational database which holds the bank of the test items. Each of the test items in this item bank has its own encoding, as shown in Figure 44, and its database entry contains detailed information about this question such as which is the correct choice, and which choices correspond to which known misconceptions, noting that there are situations in which correct choices may also be the same choices as a known misconception.

The reasons I used XML for the test items is firstly to separate the content of the test items from the format in which they are presented to the user; and secondly that this can accommodate the specifics of the mathematical notations on the Web, including mathematical notation, and graphical items and diagrams. The mathematical notation is structured using MathML, the mathematical markup language, and the graphics were encoded using SVG, Scalable Vector Graphics.

APPENDIX D :
THE ITEM BANK FOR ONLINE TESTS

This appendix provides the full list of the test items used in online tests and these are presented in the same form as the users would see on the Web diagnostic system.

A total of 153 test items are included here, covering the range of the test item types as identified in Table 43. These are numbered 10001 to 10073, with a gap from 10032 to 10051 which were duplicates of other items and which were merged prior to the analysis.

Table 43. Summary of Item Bank by Item Type

Item#	Item Types	Note
10000-10020	PV1, PV2	Place-value in decimal numbers.
10021-10050	DO	Ordering of decimal numbers with two or five choice
10051-10055	PV1, PV2	Place-value identification
10056-10059		General problems which are not used for analysis in this study
10060-10069	CR	Representation of common fractions
10070-10079	NL	Common fractions on the Number line
10080-10091	CG	Common fractions in graphical structures
10092-10111	CO	Common fraction ordering – two choices
10112-10131	CE	Closest number to common fraction
10152-10171	CA	Addition of common fractions
10172-10173		General problems which are not used for analysis in this study

APPENDIX E : ONLINE LESSON STRUCTURE

Within this Appendix I provide the details of how the online lessons and the tests were conducted. This supplements the overview provided in Chapter 4.

This appendix explains the three elements used for the online lessons: the Information Pages, the Tests consisting of individual Test Items, and the Results Pages. It then outlines the structure of each of the four lessons used in this study to gather the information.

Information Pages

The information pages were structured into a set of pages that the users were required to open before they could move on to the next element in the lesson. The welcome pages introduce the online assessment process to the learners, and include details from previous lessons. Other information pages include remedial information to help the learners to understand the tests, and concluding comments which were provided at the end of each lesson.

The user was restricted from jumping ahead but could move back to any previous information page visited. This was to allow the learners to review the information and to

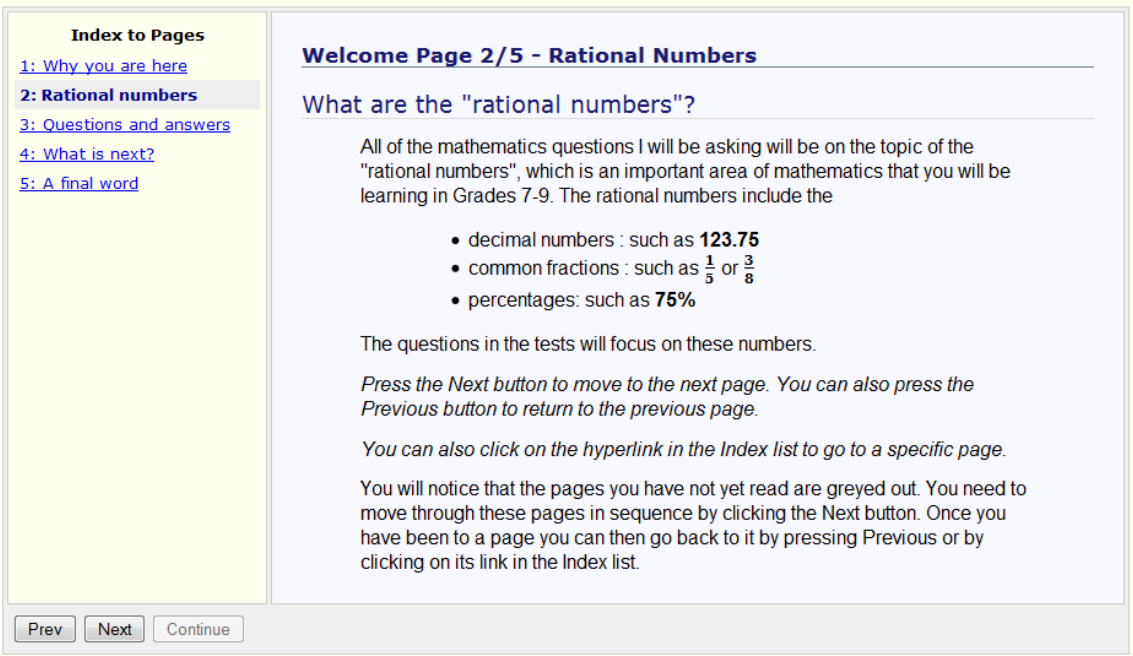


Figure 45. Example of the Information Pages during Web-based assessment

require them to at least view each page before moving on. I could not test whether the learners read the pages or simply skipped through these.

Figure 45 shows a sample of the information pages presented to the learners. This page is part of the introductory materials before the first test is completed. The user interface is intuitive for anyone with experience using the Web, with clear information on the hyperlinks, and the navigation buttons.

Tests

The tests were structured into a set of questions, called test items. These were provided to the learner one test item at a time. The learners must respond and cannot move on to the next test item without providing a suitable response. The learners respond by selecting one of the multiple choice responses, or select a check box indicating that they did not understand the question. The learners were also required to indicate whether they found the question Easy, Just Right or Difficult.

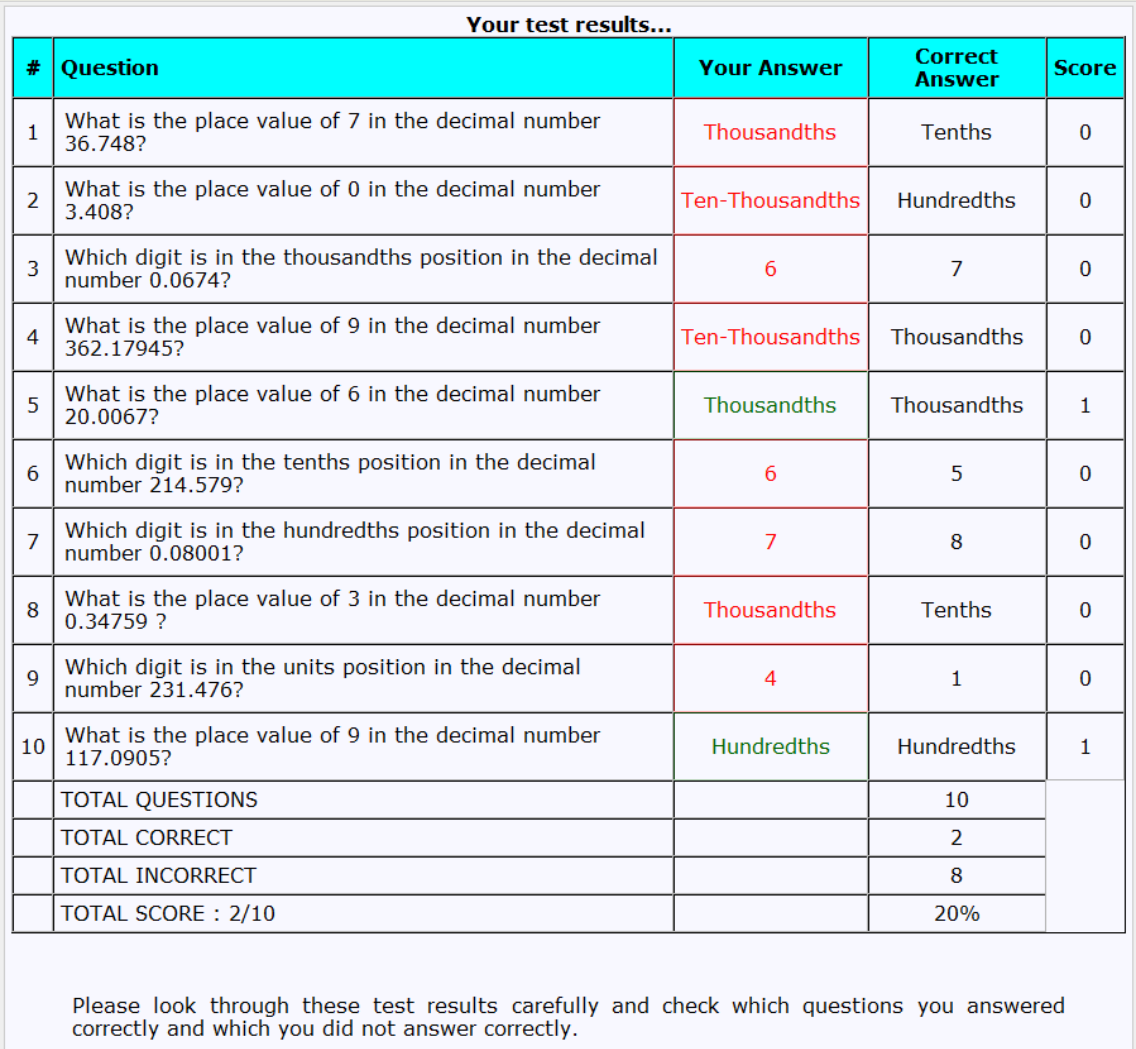
The learners were not able to return to a previous item to change a previous response, and for every item they were informed how many items in the test they were currently completing. When the test was completed, a results page was displayed automatically.

Results Pages

The results page showed the test items in the test just completed, and included the results in tabular format, with a total score at the bottom of the page. Incorrect responses were displayed in red, while correct items were displayed in black.

This results page also showed each entire test item as presented to the learners, including all mathematical notations and graphical elements used as part of both the questions and the various choices.

I intended that the learners would spend some time reflecting on the results and analyzing their results before moving on to the next part of the lesson.



Your test results...

#	Question	Your Answer	Correct Answer	Score
1	What is the place value of 7 in the decimal number 36.748?	Thousandths	Tenths	0
2	What is the place value of 0 in the decimal number 3.408?	Ten-Thousandths	Hundredths	0
3	Which digit is in the thousandths position in the decimal number 0.0674?	6	7	0
4	What is the place value of 9 in the decimal number 362.17945?	Ten-Thousandths	Thousandths	0
5	What is the place value of 6 in the decimal number 20.0067?	Thousandths	Thousandths	1
6	Which digit is in the tenths position in the decimal number 214.579?	6	5	0
7	Which digit is in the hundredths position in the decimal number 0.08001?	7	8	0
8	What is the place value of 3 in the decimal number 0.34759 ?	Thousandths	Tenths	0
9	Which digit is in the units position in the decimal number 231.476?	4	1	0
10	What is the place value of 9 in the decimal number 117.0905?	Hundredths	Hundredths	1
	TOTAL QUESTIONS		10	
	TOTAL CORRECT		2	
	TOTAL INCORRECT		8	
	TOTAL SCORE : 2/10		20%	

Please look through these test results carefully and check which questions you answered correctly and which you did not answer correctly.

Figure 46. Sample Results Page in the Web-based assessment

I now outline the details of the structure of each of the lessons as they were given to the first study school, and then repeated for the second study school. The individual lessons were planned based upon an initial analysis of the results, determining whether some types of items should be repeated or other types should be incorporated. Another decision concerned how many items to include per test and per lesson as a whole, as well as how much informational materials should be presented to the learners.

Each Test, in which the learners entered their responses to a set of test items presented to them, was immediately followed by a display of the Results Page for this test and these are not indicated in the lessons structure for each Test.

Lesson 1

The first lesson was structured into a small number of elements to ensure that the learners became familiar with the online test format. The learners quickly became familiar with the online lessons and there was time remaining at the end of the lesson.

The following elements were included into this first lesson.

Element	Type	Description	Num Questions
	INFORMATION	The welcome pages introduce the online assessment site, as well as the nature of the research project, in particular why learners make the same types of mistakes in mathematics. There is a short introduction to the rational numbers including decimal fractions, common fractions and percentages. The MCQ format is explained and also the need for the learner to answer whether the learners found this Easy, Just Right, or Difficult, as well as when they should use the Don't Know checkbox. The test structure is outlined to the learner, indicating that each test consists of around 5-10 questions, followed by results, and then followed by some feedback. There is also information given on how to use the online assessment, clicking on the Continue button to move onto the next page.	
1	INFORMATION	The introduction to the place-value types of question, and a small example is given using the decimal fraction 2.375.	
2	INFORMATION	The specific question types from the next test are explained using a sample question with its answer.	
3	TEST 1	Decimal place names – selecting the digit at a particular place name, and selecting the place name of a particular digit.	10
4/5	INFORMATION	There are 5 pages which explain the process. Detailed examples of questions are analyzed. These information pages also show how the names TENTHS, HUNDREDTHS etc. are created and how to recognize them in decimal numbers.	
6	TEST 2	Decimal place names – selecting the digit at a particular place name, and selecting the place name of a particular digit.	5
7	INFORMATION	Five detailed examples of the questions are given, and explained in detail.	
8	TEST 3	Decimal place-values – last 5 questions repeated from the first test in this lesson.	10
9	INFORMATION	A short note on what will be covered in the following week	

Lesson 2

Following from the first lesson, I adapted the content for the second lesson to meet the level of the learner's knowledge, and to place more emphasis on a few types of questions rather than to include too many individual types of questions.

For this second lesson, I continued the testing of the place-value, and introduced decimal ordering.

In School B, there had been a problem with setting up the computers with the FireFox browser for Lesson 1, and this resulted in some learners not completing the second test, so this was repeated as the first step in this second lesson.

Element	Type	Description	Num Questions
	INFORMATION	Introductory notes about what happened in the previous lesson.	
1	INFORMATION	Introducing the place-value again as a refresh of the types of examples given in Lesson 1.	
2	TEST 1	Decimal Place-Value – being the repeat of the last test from the first lesson.	10
3	INFORMATION	Summarizing the place-value tests.	
4	TEST 2	Decimal Ordering by selecting the larger or smaller of two decimal numbers	20
5	INFORMATION	A number of worked examples are provided, with detailed explanations for why some numbers are larger than others.	
6	TEST 3	Decimal Ordering by selecting the larger or smaller of two decimal numbers	10
7	INFORMATION	A short explanation of place-value and worked examples, prior to commencing the next test.	
8	TEST 4	Decimal Place-Value additional questions.	5
9	INFORMATION	Closing notes for this lesson	
10	TEST 5	Extra Tests for those who complete the other tests on time	4

Lesson 3

In Lesson 3 I decided to extend the range of the tests to include additional test types to help identify the various misconceptions within the areas of common fraction estimation, common fraction values on the number line and common fraction graphics. There was also an additional set of test tests on the common fraction ordering which were introduced in Lesson 2.

Element	Type	Description	Num Questions
	INFORMATION	This is a short summary of the previous lesson and introducing this lesson	
1	INFORMATION	This summarizes the work for this lesson and reminds the learners about how to use the online assessment system.	
2	INFORMATION	Explaining the questions of the common fractions in the next test.	
3	TEST 1	Common Fraction Estimation – which is the closest to a given word description of a common fraction.	10
4	INFORMATION	Short explanation of the estimation	
5	INFORMATION	Number Line - introduction to the types of questions in the following test	
6	TEST 2	Common Fraction Number Line – what is the value of the red arrow (which is pointing to a place on the number line).	10
7	INFORMATION	An explanation of the number line, following the results of the tests.	
8	INFORMATION	Fraction Graphics – introduction to the graphical questions	
9	TEST 3	Common Fraction Graphics – a range of questions in which various graphical objects are structured in terms of questions, such as what is the fraction which represents the blue squares in a set of blue and red squares.	12
10	INFORMATION	A short explanation of the graphical questions as a round up	
11	INFORMATION	Common Fraction Ordering – an explanation of the questions to follow.	
12	TEST 4	Common Fraction Ordering 2 numbers	10
13	INFORMATION	A detailed explanation of the types of questions and how to resolve them. This includes analysis of different situations of fraction ordering.	
14	TEST 5	Common Fraction Ordering	10
15	INFORMATION	Final Notes	

TEST 1 concerned the estimation of the value of common fractions. All of the questions were of the form: “Which of the following most closely represents the fraction ‘five and six-ninths’?” where four choices were presented to the learners in fraction notation.

TEST 2 used the number line, and asked the learners to select one of the given fractions as being the best representation for the number identified at this point. Most of these had the form: “What is the value of the red arrow on the number line as a common fraction?”, where four choices were presented to the learner for selection.

TEST 3 used graphical elements for solving particular fraction problems, such as "which fraction best represents the diagram?", with a corresponding diagram, and four possible fractional choices.

TEST 4 was based upon a comparison of fractions, asking which is the smaller of two common fractions. This was then followed by TEST 5 in which the question asked the learner which is the larger of two common fractions.

Lesson 4

This final assessment lesson introduced a few more types of test items, and returned to others.

Element	Type	Description	Num Questions
1	INFORMATION	Summarizing the structure of this lesson and what will be covered	
2	INFORMATION	Example of common fractions and an example of a question	
3	TEST 1	Common Fraction Estimation	5
4	INFORMATION	Explanation of the previous problem with a few worked examples.	
5	TEST 2	Common Fraction Estimation	5
6	INFORMATION	Explaining the different types of common fractions including proper, improper, and mixed.	
7	TEST 3	Common Fraction Ordering	5
8	INFORMATION	This includes five pages that explain the various cases of common fraction ordering with examples.	
9	TEST 4	Common Fraction Orders	5
10	INFORMATION	A simple example of common fraction addition as a prelude to the test.	
11	TEST 5	Common Fraction Addition	5
12	INFORMATION	Worked example of common fraction addition.	
13	TEST 6	Common Fraction Addition	5
14	INFORMATION	Final Notes	

There were six tests included in this final lesson, all of which were shorter than the previous tests in past lessons.

TESTS 1 and 2 introduced a new type of test item, that required finding the closest decimal number to a given fraction. A more complex version of this type of test, originally derived from the TIMSS 2003 study (IEA, 2007), was included in to the pretest. These

tests were structured in the form “Which is the closest number to the fraction N/M”, with four options provided, including both whole numbers and decimal numbers.

TESTS 3 and 4 returned to the common fraction comparison, to gather more data for this type of test.

TESTS 5 and 6 address the addition of common fractions.

This marks the end of the lesson designs, and a number of the original questions types used in the pretest, and which had been considered for inclusion, were not eventually included.

The reasons for the exclusion of some item types were that there was a limitation on what could be included into four lessons, and I decided it was better to focus on some test items in detail to demonstrate the principles of diagnostic assessment rather than to cover too broad a field, thus depth was prioritized over breadth while also providing sufficient breadth to enable many micro-domains to be included.

Structure of the Individual Online Tests

Within this section I provide further details on each of the individual tests conducted within each of the four lessons, in terms of the specific test items used from the item bank as described in Appendix D :
The Item Bank for Online Tests. Some of the test items were repeated from one lesson to the next.

Table 44. Test Items used by Test

Lesson	Test	Test Items
1	1	10001-10010
1	2	10011-10015
1	3	10016-10020, 10006-10010
2	1	10016-10020, 10006-10010
2	2	10021-10040
2	3	10041-10050
2	4	10051-10055
2	5	10056-10059
3	1	10061-10069

3	2	10070-10079
3	3	10080-10091
3	4	10092-10101
3	5	10102-10111
4	3	10112-10121
4	5	10122-10131
4	7	10092-10101
4	9	10102-10111
4	11	10152-10161
4	13	10162-10171
4	15	10172-10173

APPENDIX F :
RASCH ANALYSIS USING WINSTEPS

Much of my analysis in Chapter 6 uses the Rasch method, and I use the WinSteps program (Linacre, 2013) which provides a wide range of outputs and calculated measures for the analysis of test item responses and learner proficiencies.

The inputs to WinSteps are:

- the set of test items used, each identified by a code;
- the set of the learners, also each identified by a code, and
- the set of learner responses to each test item.

The responses are coded as 0 (for an incorrect response) or 1 (for a correct response). Missing values are accommodated and are coded using a period (.), and these are automatically omitted from the analysis, with the Rasch analysis providing a valid result for learner ability and item difficulty while accommodating missing values.

The primary outputs from the Rasch analysis are the grading of the test items on a scale of difficulty, centered on zero, and the grading of the learners using a calculated ability scale. The scales of item difficulty and learner ability are the same, which is a unique feature of the Rasch analysis technique.

Other outputs from the Rasch analysis are the fit of the response data to the Rasch model. The Rasch method finds the best-fitting model to the input data as provided, and there will be data which fits this model and data which does not fit, which are outliers. These are called “misfitting” data in terms of the model. The Rasch method structures the test items on a scale of “difficulty” and the learners onto a scale of “ability”. In the Rasch method the more proficient learners answer more of the difficult items correctly and the easy items are answered correctly by all of the learners. Misfitting data will distort the model and it is recommended by Linacre (2013) that these be removed prior to running the Rasch model again. Misfit test items are those for which the results are inconsistent, and these are removed to reduce the distortion in the model. Misfit learners are also removed, for cases where the learner responses do not fit the data results, and such misfits in learner data may be an indication of guessing, where the relationship between learner capability and item difficulty cannot be determined due to inconsistency in the responses from these misfit learners.

I apply Rasch analysis in two different ways within this study. Firstly, in determining the proficiency of learners on the basis of the test items presented, to identify and remove proficient learners from further analysis of the incorrect responses. If a learner has demonstrated proficiency, which is the STABLE stage of my model, then they will not have made enough mistakes to warrant the analysis of their misconceptions and any such mistakes are more likely to be slips. This is the traditional approach to Rasch analysis, in which the trait being measured is the learner's "ability", and this traditional approach is generally not suitable for the measurement of misconceptions (Stacey & Steinle, 2006).

The second way I apply Rasch analysis is by addressing particular misconceptions that would cause the rich distractors to be selected in the MCQs, and I analyze these responses to determine which test items are good indicators of each specific misconception, and also which learners have evidence of the usage of the misconception. Thus the trait I am measuring here is not "ability", in its traditional meaning, but rather the extent to which the learners' responses can be accounted for by their usage of a particular way of thinking.

The WinSteps program requires the creation of a control file which instructs WinSteps on the input data provided, the calculations to perform, and the reports to produce. Building this control file for each Rasch analysis is a relatively complex and time-consuming process, and I have semi-automated this process in my combined results database, using the WinSteps parameters that are of interest. WinSteps has a large number of controllable parameters, each of which influences the input, processing and output of the data and results.

I now outline how I have used the WinSteps approach to Rasch analysis throughout the detailed analyses in Chapter 6.

The control variables for WinSteps are provided at the top of the control file and inform WinSteps of the input, processing and output for each analysis run.

As an example, Table 45 shows part of the control file used for the first analysis for the micro-domain of place-value knowledge. Of interest here is that there are 25 items (NI variable on line 11), and the item codes are provided as 10001 to 10055 (lines 25-49). The CODES variable (line 18) indicates that the responses for this analysis are restricted to the values of 0 and 1, and that others will be ignored, and the CLFILE variable (lines

19-22) provides labels for the 0 and 1 responses. Other parameters in this example are used to determine the types of output, such as the CSV=1 on line 6, indicating that the results will also be produced as a CSV (comma-separated values) file for input into other systems, and the file names to be used to hold this CSV data, on lines 9 and 10. Files in the CSV format can be easily imported into other applications such as Excel and database systems.

Table 45. Example of Control File for Rasch analysis

```
1.  ; ------------------------------------------------------------------------------
2.  ; PV - SCHOOLS A+B - CORRECT
3.  ; Created on 2013-Nov-17 Sunday
4.  ; ------------------------------------------------------------------------------
5.  TITLE=  PV - SCHOOLS A+B - CORRECT
6.  CSV=    Y
7.  HLINES= N
8.  QUOTED= N
9.  PFILE=  PV-AB-LRNR.CSV
10. IFILE=  PV-AB-ITEM.CSV
11. NI=      25
12. ITEM1=  1
13. NAME1=  41
14. PERSON= LEARNER
15. ITEM=   TESTITEM
16. IMAP=   $3W3
17. PMAP=   $1W3
18. CODES=  01
19. CLFILE= *
20. 0 Not detected
21. 1 Detected
22. *
23. &END
24. ; The individual test items in this analysis
25. 10001
26. 10002
27. 10003
28. 10004
29. 10005
30. 10006
31. 10007
32. 10008
33. 10009
34. 10010
35. 10011
36. 10012
37. 10013
38. 10014
39. 10015
40. 10016
41. 10017
42. 10018
43. 10019
44. 10020
45. 10051
46. 10052
47. 10053
48. 10054
49. 10055
50. END NAMES      ; End of the test items, DATA follows
51. 000000000010011.....01111               A01
52. 000000000000010.....10101               A02
53. 111111111111111.....11111               A03
54. …
```

The data lines of the responses are from line 51 onwards, and only three lines of responses have been displayed in this example.

The outputs from WinSteps include diagnostic tables, result tables, and plots, and each of these have a specific WinSteps "table number" which is shown in the first line of the data tables used in the analyses.

I use the following WinSteps output tables frequently in the analysis of the response data:

- TABLE 26.1 ITEM POLARITY: a diagnostic table to help identify the correlations of the items with the learner measures, used to help to filter out items which do not fit the learners' responses. The test items are presented in point-measure correlation sequence.
- TABLE 2.6: ITEM-CATEGORY MEASURES: which provides evidence of misfit test items, identified by the mis-ordering of the responses against the expected values.
- TABLE 10.1: ITEM MISFIT: which shows the test items in the sequence of their misfit to the Rasch model. A good fit means that the test item difficulty is aligned with the learner ability, and if not then this TABLE 10.1 will show two special cases of misfit beyond the pure correlation provided in TABLE 26.1. INFIT is higher when the items with measures close to a learner's measure do not show an increased level of difficulty and is a weighted statistic. OUTFIT is an unweighted which is more sensitive to unexpected results at the extremes, such as caused by slips or guessing.
- TABLE 17.1: LEARNER MEASURE ORDER: showing the results by learner in the sequence of their estimated ability on the trait being measured. For my purposes this trait is the proficiency in the micro-domain for the first part of my analysis, and is then the level to which they are using a specific misconception in the second part of my analysis. This analysis is repeated if there is more than one misconception to account for the responses in the multiple-choice options.

ENDS