The Construction and Validation of the Primary Scientific Reasoning Test



Volume I

Ng Yee Ping Diana

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy in Education

Michaelmas Term 2018

The Construction and Validation of the Primary Scientific Reasoning Test

Ng Yee Ping Diana

St Anne's College

Doctor of Philosophy in Education

Michaelmas Term 2018

Abstract

A major task of science education is to foster scientific reasoning. Despite extensive research, it is still unclear how best to assess this cognitive proficiency; a limited consensus of its nature has arguably constrained the development of classroom assessments of scientific reasoning. Therefore, this research aimed to develop a valid and reliable test of scientific reasoning – the Primary Scientific Reasoning Test *(*PSRT) – for primary-aged pupils. Developments in science education, as well as emerging and novel insights across substantive literature, formed the theoretical framework underlying the PSRT.

Mixed methodology was used to develop and validate the PSRT over four phases. The first phase developed the theoretical framework and the draft items. In the next phase, an 18-member expert panel from Singapore and England evaluated and endorsed the conceptual validity of the construct and the content validity of the draft items. The third phase field-tested and refined the draft instrument with 138 Year 7 pupils from four Singapore secondary schools. In the fourth and last phase, 431 Year 6 pupils from six Singapore primary schools attempted the finalised 100-item PSRT. Results from psychometric analysis using Rasch modelling supported inferences of a unidimensional construct underlying the test, and examination of the item difficulty estimates produced evidence that the order of difficulties generally accorded with the framework of scientific reasoning.

Given that the explanatory concepts in the framework generally explained the pupils' performance and drawing on the strength and scope of other evidence, the thesis concludes that the PSRT displayed reasonably strong construct validity. This conclusion supports the use of PSRT as a valid and reliable measure of scientific reasoning for primary-aged pupils. The theoretical basis of the PSRT and its construction procedures also bear lessons for test developers and researchers. Overall, the study contributes clarification and understanding to the nature and assessment of scientific reasoning, particularly for primary-aged pupils.

Dedication

I dedicate this thesis to all primary school pupils, in whose hands the future lies.

Acknowledgements

I am indebted to the following individuals:

Husband and children

Alex, Annabelle and *Ashleigh*

Supervisors

Professor Jo-Anne Baird and *Associate Professor Joshua McGrane*

Participating secondary and primary schools, and their pupils

Study sponsor

Singapore Examinations and Assessment Board

Advisors

Professor David Andrich and *Associate Professor Therese N. Hopfenbeck*

Friends and colleagues at

Oxford University Centre for Educational Assessment

Singapore Examinations and Assessment Board

Department of Education, University of Oxford

Ministry of Education, Singapore

Table of Contents

Abstract 2

Dedication 3

Acknowledgements 4

Table of Contents 5

List of Figures 10

List of Tables 14

List of Abbreviations 21

List of Appendices 23

Chapter 1 Why Study Scientific Reasoning in Primary Children? 24

1.1 Introduction 24

1.2 Background 25

1.3 Rationale 29

1.3.1 Current Scientific Reasoning Tests 29

1.3.2 Limitations of Current Scientific Reasoning Tests 29

1.4 Contributions of Thesis 33

1.5 Research Question and Research Aims 34

1.6 Organising Themes in the Thesis 35

Chapter 2 Literature Review: What is Scientific Reasoning? 36

2.1 Introduction 36

2.2 Contemporary Conceptions of Scientific Reasoning 37

2.2.1 Logic as Scientific Reasoning 37

2.2.2 Induction as Scientific Reasoning 39

2.3 Post-Logic and Post-Inductive Scientific Reasoning 44

2.3.1 Expert Reasoning as Scientific Reasoning 44

2.3.2 Procedural Reasoning as Scientific Reasoning 45

2.3.3 Epistemic Reasoning as Scientific Reasoning 46

2.4 Science as a Practice-based Activity 48

2.5 Construct of Scientific Reasoning in this Thesis 49
2.5.1 Theoretical Support for Construct 50
2.5.2 Specific Reasoning Associated with Scientific Activities 53
2.6 Conclusions 57
Chapter 3 Methodology: Validation Framework 59
3.1 Introduction 59
3.2 Validation Framework 60
3.2.1 Validity and Validation 60
3.2.2 Implications of the *Standards'* Recommendations for the Study 66
3.2.3 Validation Framework in Phases 74
3.3 Validation and Ethical Considerations 80
Chapter 4 Initial Test Design 83
4.1 Introduction 83
4.2 Validation in Initial Test Design 84
4.3 Development of Definitions and Behavioural Descriptors 86
4.3.1 Types of Knowledge 87
4.3.2 Science Practices 96
4.3.3 Sub-Practices and Attainment Levels 97
4.4 Development of Test and Items 108
4.4.1 Centrality of Learning Metaphors 109
4.4.2 Learning Theories and Assessment Implications 110
4.4.3 Impact of Cognitive Learning Theories on Item and Test Design 112
4.5 Test Framework 120
4.5.1 Characterisation of Items 120
4.5.2 Assessment Structure 122
4.5.3 Tables of Specifications 125
4.6 Conclusions 129
Chapter 5 Expert Panel Review 130
5.1 Introduction 130

5.2 Validation by Expert Panel Review 131
5.3 Expert Panel 132
5.4 Materials 135
5.5 Procedure 139
5.5.1 Before and During Review 139
5.5.2 After Review 140
5.6 Data Analysis 140
5.6.1 Rating Scales in the Proforma 140
5.6.2 Qualitative Feedback from Proforma and Closure Meetings 143
5.7 Results and Discussion 144
5.7.1 Content Validity Evidence of the Scientific Reasoning Construct 144
5.7.2 Content Validity Evidence of the Draft PSRT Items 150
5.7.3 Item Outcomes Based on Validity Evidence from the Expert Panel 156
5.8 Conclusions 158
Chapter 6 Pilot Study 160
6.1 Introduction 160
6.2 Validation in the Pilot Study 160
6.3 Schools and Participants 161
6.4 Materials 162
6.5 Procedure 162
6.6 Data Analysis 164
6.6.1 Classical and Modern Test Theories 164
6.6.2 Rasch Measurement Theory 166
6.6.3 Rasch Analysis 171
6.7 Results 177
6.7.1 Draft Primary Scientific Reasoning Test (Draft PSRT) 177
6.7.2 Item and Person Fit Statistics of the Draft PSRT 179
6.8 Discussion 183
6.8.1 Final Item Outcomes Based on Validity Evidence from Phases II and III 184

6.9 Conclusions 186
Chapter 7 Main Study – Dimensionality of Construct 187
7.1 Introduction 187
7.2 Validation in the Main Study 188
7.3 Validation Stage 1 190
7.3.1 Participants and Materials 191
7.3.2 Procedure and Data Analysis 191
7.3.3 Results and Discussion 191
7.4 Validation Stage 2 193
7.4.1 Schools and Participants 194
7.4.2 Materials 195
7.4.3 Procedure 199
7.5 Data Analysis Strategy 201
7.5.1 Data Analysis 202
7.6 Results 206
7.6.1 Overall Fit Statistics of the PSRT 207
7.6.2 Item and Person Fit Statistics of the PSRT 207
7.6.3 Review and Resolution Outcomes 211
7.6.4 Revised PSRT 219
7.6.5 Item- and Person-Fit Statistics of the Revised PSRT 222
7.6.6 Revised PSRT Booklets 224
7.7 Discussion 226
7.8 Conclusions 228
Chapter 8 Main Study – Identification of Construct 230
8.1 Introduction 230
8.2 Item Difficulty and Item Features 230
8.3 Data Analysis 233
8.4 Results 235
8.4.1 Item Difficulty Estimates in the Revised PSRT 235

8.4.2 Item Difficulty Estimates of Dichotomous Items 237
8.4.3 Threshold Estimates of Polytomous Items 251
8.4.4 Analysis of Pupils' Responses 260
8.5 Discussion 273
8.6 Conclusions 277
Chapter 9 Validity of the PSRT 278
9.1 Introduction 278
9.2 Construct Validity of the PSRT 278
9.2.1 Evidence from Theoretical and Methodological Frameworks 279
9.2.2 Evidence from Development, Review and Pilot of the PSRT 280
9.2.3 Evidence from the Validation of the PSRT 283
9.2.4 Construct Validity of the PSRT 286
9.3 Significance of the Study for Theory and Practice 287
9.3.1 Defining the Construct of Scientific Reasoning 288
9.3.2 Measuring Scientific Reasoning 289
9.3.3 Validating the PSRT 290
9.3.4 Implications for Curriculum and Instruction 291
9.3.5 Analysing Performance Data from the PSRT 293
9.4 Limitations of Study 296
9.4.1 Theoretical Limitations 296
9.4.2 Methodological Limitations 298
9.5 Future Research 299
9.6 Final Thoughts 300
Appendices ..... **Error! Bookmark not defined.**
References 302

List of Figures

Figure 1.1. Question made up of two items adapted from a Singapore primary school paper. 24
Figure 3.1. Items 1 and 2 are adaptations of items from the 2010 science key stage 2 and 2011 Year 6 science examinations. 59
Figure 4.1. An item in the Draft PSRT crafted using ideas gathered from a news event. . 84
Figure 4.2. A physical science learning outcome demonstrating the incorporation of content and skills. Adapted from Ministry of Education Singapore (2013, p. 50). 90
Figure 4.3. A life science learning outcome demonstrating the incorporation of content and skills. Adapted from Ministry of Education Singapore (2013, p. 45). 90
Figure 4.4. Four attainment stages to guide the development of level descriptors in each sub-practice. Adapted from Curriculum Council (2005). 98
Figure 4.5. A three-item question in the Draft PSRT. Items 1 and 2 have three-options SROS formats while item 3 has a CRSS format. 117
Figure 4.6. A grid model for characterising items along the axes of knowledge type, science practices, and levels of cognitive demand. Each of the three colours (blue, pink, and yellow) signifies a science practice. 120
Figure 5.1. An item in the Draft PSRT that had a reasoning flaw detected during the review. 130
Figure 5.2. Extent to which members of the panel agreed with the statements about the scientific reasoning construct. 145
Figure 6.1. The ICC of the dichotomous item – A1(b) from the anchor block. The item had a location estimate of 0.316. The black dots show the average response of the three class intervals and the curve shows the model expectations. Visual inspection suggested there were slight deviations between the observed means and the model. The item fit residual of -0.097, which was close to the expected value of 0, indicated acceptable overall deviation from model. 174
Figure 6.2. Uniform differences observed among subgroups (schools) reflect the presence of DIF in a dichotomous item – BIQ3(ii) from block I. This meant there was a consistent deviation in performance on the item among the four schools. 175
Figure 6.3. The category probability curve (CPC) of the polytomous item – BIIIQ3(b) from block II, demonstrating ordered thresholds. Persons with ability estimates ranging from approximately -2 to 0.5 logits on the scale were more likely to attain the middle response category of one mark compared to the adjacent categories of no mark or two marks. ... 176
Figure 6.4. The CPC of the polytomous item – BIIQ1 from block II, demonstrating disordered thresholds. The middle scoring category of one mark is not functioning as intended. 176

Figure 6.5. Person-item threshold distribution of the Draft PSRT. This graph superimposes the distribution of the items (in blue-patterned bars) and the persons (in red-patterned bars) on the same scale. 178
Figure 6.6. ICC of an item from question 2 of block IV showing under-discrimination relative to model expectation. 179
Figure 6.7. Person-item threshold distribution by gender of the Draft PSRT. Blue and red columns represent the location estimate distribution of girls and boys respectively. 181
Figure 7.1. A two-item question in the PSRT to illustrate how evidence from various methods of analysis gave insights about the functioning of the items..... 187
Figure 7.2. The “before” and “after” changes to an item based on the two-member review. 192
Figure 7.3. Flow-chart of the item diagnostic process comprising an iterative calibration and remediation loop. *Depending on the item, there could be multiple remediation attempts. 205
Figure 7.4. Person–item threshold distribution of the 100-item PSRT..... 208
Figure 7.5. An item requiring pupils to extrapolate from existing data. Credit was awarded for any numerical answer from 18 to 25 seconds. 208
Figure 7.6. ICC of item AQ3(e)(ii) showing the systematic DIF between genders. The plots in red and blue are for the performance of boys and girls respectively. 209
Figure 7.7. CPC of item BXQ2(a)(i) demonstrating the disordered thresholds. The scoring (response) category of one mark did not function as expected. 209
Figure 7.8. ICC of item BIQ2(b). The responses demonstrated an under-discriminatory fit to the model with a fit residual of 3.220 and a statistically significant chi-square probability. 210
Figure 7.9. A four-item question exemplar from the PSRT..... 213
Figure 7.10. Item characteristic curve of item BIIQ3(a) in run 1. 215
Figure 7.11. CPC with disordered thresholds obtained in run 2. 216
Figure 7.12A. CPC in run 3. 217
Figure 7.12B. ICC in run 3. 217
Figure 7.12C. CPC in run 4. 217
Figure 7.12D. ICC in run 4..... 217
Figure 7.13A. CPC in run 5. 218
Figure 7.13B. CPC in run 6. 218
Figure 7.14. Person–item threshold distribution of the Revised PSRT..... 219
Figure 7.15. Person–item threshold distribution by gender of the Revised PSRT. Blue and red columns represent the location estimate distribution of girls and boys respectively. 220

Figure 8.1. Schematic representation of the five key item features hypothesised to impact difficulty. SROS is selected-response objective scoring, CROS is constructed-response objective scoring, and CRSS is constructed-response subjective scoring. 231
Figure 8.2. Item map from RUMM2030 of the 86-item Revised PSRT. Every “X” represents two pupils. Item label indicates the block and order of items; e.g., A1(b) is the second item of the first question in the anchor block, BIIQ3(b) is the second item of question 3 from block II. Labels in blue indicate the eight polytomous items. The seven DIF-resolved items were left out of this item map. 238
Figure 8.3. Items of each demand level, as a percentage of the total items (indicated by the number below the bars) in each location estimate interval..... 240
Figure 8.4. Items of each format, as a percentage of the total items (indicated by the number below the bars) in each location estimate interval. SROS is selected-response objective scoring, CROS is constructed-response objective scoring, CRSS is constructed response subjective scoring..... 240
Figure 8.5. Items measuring each knowledge type, as a percentage of the total items (indicated by the number below the bars) in each location estimate interval..... 241
Figure 8.6. Items measuring each science practice, as a percentage of the total items (indicated by the number below the bars) within each location estimate interval. Practice 1 – giving scientific explanation, practice 2 – designing and evaluating investigation, practice 3 – interpreting and analysing data and evidence..... 241
Figure 8.7. The figure above is part of BVQ1. The item’s location estimate of 3.137 was the highest in the Revised PSRT..... 248
Figure 8.8. Items (i) and (ii), of BIIIQ2(b), had locations of -1.523 and 1.481, respectively. 249
Figure 8.9. An item that tested understanding of gravitational force with a location of 0.248. 251
Figure 8.10. An item which required processing of graphical information. 254
Figure 8.11. An item with a graphical stimulus composed of two linear relationship. 254
Figure 8.12. An item testing the ability to draw out scientific ideas from an abstract model. 255
Figure 8.13. A polytomous item of low cognitive demand with change recommendations. 257
Figure 8.14. A polytomous item of high cognitive demand with change recommendations. 257
Figure 8.15. ICC of BIIQ1. 258
Figure 8.16. CPC of BVIIQ2(e)(ii). 258
Figure 8.17. A polytomous item with change recommendations. 259

Figure 8.18. Two items testing competencies in the first sub-practice of practice 1 – giving scientific explanation. 261
Figure 8.19. Two items in the same question testing the recall and application of content knowledge in the changes of state of water. 262
Figure 8.20. An item demonstrating the assessment of conservation and seriation reasoning. 263
Figure 8.21. BIXQ1(c) and BIXQ1(d) assessed procedural and epistemic knowledge using competencies outlined in practices 3 and 2 respectively. The first two items are not shown. 264
Figure 8.22. Two single-mark items testing the control-of-variable (COV) strategy. 265
Figure 8.23. Two items that measured pupils' reasoning with experimental variables. .. 267
Figure 8.24A. ICC of BVIIIQ1(a)(i). 268
Figure 8.24B. ICC of BVIIIQ1(a)(ii). 268
Figure 8.24C. CPC of BVIIIQ1(a)(ii). 268
Figure 8.25. An item testing the ability to estimate the area of a leaf using square grids. 269
Figure 8.26. An item testing procedural understanding related to accuracy in measurement. 269
Figure 8.27. Two items that assessed pattern identification and extrapolation competencies. The key in (b) is that as the length gets longer, the time for 10 complete swings becomes longer. The key in (e)(i) is 10.0 seconds. 271
Figure 8.28. An item assessing the justification of an epistemic claim from evidence. ... 272
Figure 8.29. An item assessing the justification of a procedural claim from evidence. ... 273
Figure 9.1. Key evidence gathered under each of the three validity sources in current study. 286
Figure 9.2. An example to show how results from Rasch analysis informed adjustments to the scoring guide of an item for a fairer assessment. 294
Figure 9.3. A novel item designed based on statistical findings from flawed items. This item assesses pupils' proficiency in justifying a claim. 295
Figure 9.4. Question made up of two items adapted from a Singapore primary school paper. 300
Appendix 1, Figure 1.1. The three domains of learning in the primary science curriculum framework. From Ministry of Education Singapore (2013). . **Error! Bookmark not defined.**
Appendix 10, Figure 1. Modified SAS code. ..... **Error! Bookmark not defined.**

List of Tables

Table 1.1 .....
Validated scientific reasoning tests categorised by participants' age range and task 32
Table 2.1 .....
Gagné's basic and integrated process skills 40
Table 2.2 .....
Styles of reasoning in European science, and examples of the propositional construct ... 53
Table 3.1 .....
Four types of validity and the approach to validation 61
Table 3.2 .....
Summary of mixed-method design focus in research 72
Table 3.3 .....
Cross-over analytic strategies in research 73
Table 3.4 .....
Summary of mixed-method validation framework in the four-phased study 74
Table 3.5 .....
Test content validity and methods to collect validity evidence in phase I 76
Table 3.6 .....
Validation plan in phase II: expert panel review 77
Table 3.7 .....
Validation plan in phase III: pilot study 78
Table 3.8 .....
Validation plan in phase IV: main study 79
Table 4.1 .....
Relationship between content validity and nomological network analytical activities 85
Table 4.2 .....
Domains in the primary science curriculum of Singapore 88
Table 4.3 .....
Essential "takeaways" of themes in primary science curriculum of Singapore 89
Table 4.4 .....
Curated set of 10 basic procedural knowledge appropriate for primary-aged science pupils
..... 92
Table 4.5 .....
Descriptions of the seven outcomes in the domain of attitudes and ethics 93
Table 4.6 .....
Three basic epistemic knowledge appropriate for primary-aged science pupils 95

Table 4.7 .....
A comparison of the skills and uses in the Singapore's primary science curriculum and the science practices in this study 96
Table 4.8 .....
Four stages of learning attainments 99
Table 4.9 .....
Learning outcomes in the five levels of sub-practice 1.1: Identify, use and evaluate models and representations in explanations and hypotheses 100
Table 4.10 .....
Learning outcomes in the four levels of sub-practice 1.2: Recall and apply appropriate scientific knowledge 101
Table 4.11 .....
Learning outcomes in the five levels of sub-practice 2.1: Plan for effective scientific investigations or evaluate approaches 102
Table 4.12 .....
Learning outcomes in the five levels of sub-practice 2.2: Identify and manipulate variables within an investigation 103
Table 4.13 .....
Learning outcomes in the four levels of sub-practice 2.3: Collect precise and reliable data during the gathering of evidence 104
Table 4.14 .....
Learning outcomes in the levels of sub-practice 3.1: Process, transform, and analyse data 105
Table 4.15 .....
Learning outcomes in the five levels of sub-practice 3.2: Evaluate evidence to construct and critique conclusions, hypotheses, and arguments 106
Table 4.16 .....
The three practices and their sub-practices in the scientific reasoning construct 107
Table 4.17 .....
A comparison of cognitive demand levels across three categorisation schemes 115
Table 4.18 .....
Testing advantages of the three basic item formats..... 116
Table 4.19 .....
Adopted guidelines for item design and their rationale 119
Table 4.20 .....
Characterisation scheme of a three-item question shown in Figure 4.5 121
Table 4.21 .....
Arrangement of anchor block and ten question blocks across five Draft PSRT booklets 123

Table 4.22 .....
Number of items in anchor block and ten question blocks across five Draft PSRT booklets 124
Table 4.23 .....
Distribution of the 87-item Draft PSRT by level of cognitive demand and item format.... 125
Table 4.24 .....
Mark distribution in the Draft PSRT according to the disciplinary themes in the curriculum 126
Table 4.25 .....
Distribution of marks according to knowledge types and science practices 127
Table 4.26 .....
Mark distribution according to knowledge type, science practice, and cognitive demand 128
Table 5.1 .....
Nine review activities conducted by the expert panel..... 133
Table 5.2 .....
Professional profile of expert panel 135
Table 5.3 .....
Eight statements about construct and relationship to specific evidence of content validity 137
Table 5.4 .....
Thirteen statements about items and relationship to specific evidence of content validity 138
Table 5.5 .....
Statistics summarising the panel members' quantitative ratings in the proforma 142
Table 5.6 .....
Means, standard deviations, and 95% score confidence intervals for eight statements about the scientific reasoning construct..... 146
Table 5.7 .....
CVIs of panel's feedback on 13 statements about the first seven items in the Draft PSRT 151
Table 5.8 .....
Number of items by CVI obtained in statement 4..... 152
Table 5.9 .....
Number of items by CVI obtained in statement 5..... 152
Table 5.10 .....
Summary of item outcomes for the 87-item Draft PSRT from findings of phase II 158

Table 6.1 .....
Sources and methods used to collect validity evidence in the pilot study..... 161
Table 6.2 .....
Distribution of the 142 participating pupils according to their gender and school 161
Table 6.3 .....
Distribution of the five-booklet Draft PSRT by schools and gender..... 163
Table 6.4 .....
Test-level fit statistics and statistical tests examined in phase III..... 171
Table 6.5 .....
Item and person-level fit statistics and statistical tests examined in phase III 173
Table 6.6 .....
Average and range of reported duration to complete the Draft PSRT per school..... 178
Table 6.7 .....
Confidence intervals of pairwise differences in mean difficulty estimates between booklets
..... 181
Table 6.8 .....
Confidence intervals of pairwise differences in mean ability estimates between schools182
Table 6.9 .....
Summary of final outcomes for 64 items in the Draft PSRT from findings of phases II and III
..... 185
Table 7.1 .....
Summary of validity plan in the two stages of the main study..... 189
Table 7.2 .....
Evidence of content validity collected from review activity by the expert panel 190
Table 7.3 .....
Analytical methods to address two research aims and questions in validation stage 2.. 194
Table 7.4 .....
Distribution of pupil sample in the main study by school and gender..... 195
Table 7.5 .....
Arrangement of anchor block and ten question blocks across five test booklets, including
the number of items in each block..... 195
Table 7.6 .....
Distribution of the 100-item PSRT by item format..... 196
Table 7.7 .....
Mark distribution in the PSRT according to the disciplinary themes in the curriculum.... 197
Table 7.8 .....
Mark distribution according to knowledge type, scientific practice, and cognitive demand
..... 198

Table 7.9 .....
Summary of the historical development of the 100 items in the PSRT 199
Table 7.10 .....
Distribution of five-booklet PSRT by school and gender 200
Table 7.11 .....
Summary of the fit statistics examined in the PSRT 203
Table 7.12 .....
Summary of corrective measures for statistical anomalies in item-fit statistics or tests .. 204
Table 7.13 .....
Summary of item irregularities in the PSRT from the initial Rasch analysis 210
Table 7.14 .....
Formation of the Revised PSRT from the outcomes of review and corrective measures211
Table 7.15 .....
Item characterisation scheme of the third question of block II (BIIQ3) 212
Table 7.16 .....
Residual correlations in the initial analysis between items 214
Table 7.17 .....
Results of the four items from an initial Rasch analysis (run 1) on the PSRT 215
Table 7.18 .....
Remediation outcomes of three run strategies (runs 2–4) 216
Table 7.19 .....
Remediation outcomes of two run strategies (runs 5 and 6) 217
Table 7.20 .....
The 95% confidence intervals of pairwise differences in mean ability estimates for the different schools 221
Table 7.21 .....
The 95% confidence intervals of pairwise differences in mean difficulty estimates for the different booklets 222
Table 7.22 .....
Fit statistics of two items in Revised PSRT with anomalous fit residuals 222
Table 7.23 .....
Five item pairs with anomalous residual correlations 223
Table 7.24 .....
Reliability indices and fit statistics of booklets 1 to 5 224
Table 7.25 .....
Eigenvalues of the first three principal components from the PCA analysis of person-item residuals for the five booklets 225

Table 7.26 .....
Results of *t*-tests and binomial tests of person estimates from subsets of items identified from first principal component across all booklets 225
Table 8.1 .....
Distribution of 86 items in location estimate intervals, in descending order of difficulty .. 235
Table 8.2 .....
Mean location estimates and 95% confidence intervals of 86 items across four item features 236
Table 8.3 .....
Details of the only pair of items within a question in block VIII that had identical features 242
Table 8.4 .....
Details of the item pair within a question in block III that differed by cognitive demand.. 243
Table 8.5A.....
Details of the first of the four item pairs found in questions that differed by their item format 244
Table 8.5B.....
Details of the second and third item pairs found in questions that differed by their item format 245
Table 8.5C .....
Details of the last of four item pairs found in questions that differed by their item format246
Table 8.6 .....
Features of 20 items in the “easiest” and “hardest” groups in ascending location estimates 250
Table 8.7 .....
Development history of the eight two-mark items in ascending order of location estimates 252
Table 8.8 .....
Features and item estimates of three two-mark items that underwent re-scoring 253
Table 8.9 .....
Features and item estimates of three two-mark items with change recommendations... 256
Table 9.1 .....
Summary of organising themes in thesis and validity evidence collected across the phases 279
Appendix 22 Table 1 .....
Two items in Revised PSRT with anomalous fit residuals... **Error! Bookmark not defined.**
Appendix 22 Table 2 .....
Five item pairs with anomalous residual correlations..... **Error! Bookmark not defined.**

Appendix 22 Table 3A .....
Suggested Amendments to Item BVIQ2(a)..... **Error! Bookmark not defined.**
Appendix 22 Table 3B .....
Suggested Amendments to Item BVIQ2(a)..... **Error! Bookmark not defined.**
Appendix 22 Table 3C .....
Suggested Amendments to Item Pairs with Anomalous Residual Correlations..... **Error! Bookmark not defined.**

List of Abbreviations

AERA	American Educational Research Association
AN	Anchor block
APA	American Psychological Association
C.I.	Confidence interval
COV	Control-of-variable
CPC	Category probability curve
CROS	Constructed-response objective scoring
CRSS	Constructed-response subjective scoring
CUREC	Central University Research Ethical Committee
CVI	Content validity index
DIF	Differential item functioning
DOK	Depth of Knowledge
HADS	Hospital Anxiety and Depression Scale
ICC	Item characteristic curve
IRT	Item response theory
MHV	Mill Hill Vocabulary
NCME	National Council on Measurement in Education
NCMUE	National Council on Measurements Used in Education
NFER	National Foundation for Educational Research
NGSS	Next Generation Science Standards
NOS	Nature of science

OECD	Organisation for Economic Co-operation and Development
PCA	Principal component analysis
PGCE	Postgraduate Certificate in Education
PISA	Programme for International Student Assessment
PSI	Person Separation Index
PSRT	Primary Scientific Reasoning Test
QuASSR	Quantitative assessment of socio-scientific reasoning
RMT	Rasch measurement theory
SD	Standard deviation
SDDS	Scientific discovery as dual search
SPM+	Standard Progressive Matrices-Plus
SROS	Selected-response objective scoring
TIMSS	Trends in Mathematics and Science Study
TOS	Table of specifications
TOSLS	Test of scientific literacy skills

List of Appendices

Appendix 1. A Summary of the Primary Science Curriculum in Singapore ..... **Error! Bookmark not defined.**
Appendix 2. Scientific Thinking and Scientific Reasoning... **Error! Bookmark not defined.**
Appendix 3. A Summary of Key Stages 1 and 2 in the National Curriculum ..... **Error! Bookmark not defined.**
Appendix 4. Excerpt from CUREC Form Addressing Ethical Concerns **Error! Bookmark not defined.**
Appendix 5. Science Practices, Sub-Practices, and Levels **Error! Bookmark not defined.**
Appendix 6. Draft PSRT, Characterisation Scheme, Rasch Fit Statistics **Error! Bookmark not defined.**
Appendix 7. Item Format and Cognitive Demand in Draft PSRT Items **Error! Bookmark not defined.**
Appendix 8. Item-Curricular Theme Correspondence in the Draft PSRT **Error! Bookmark not defined.**
Appendix 9. Table of Specifications of Draft PSRT by Knowledge Types, Science Practices and Cognitive Demand..... **Error! Bookmark not defined.**
Appendix 10. Score Method ..... **Error! Bookmark not defined.**
Appendix 11. Content Validity Indices of Items in Anchor and 10 Question Blocks (Blocks I to X) of Draft PSRT ..... **Error! Bookmark not defined.**
Appendix 12. Summary of Key Issues and Item Outcomes from Expert Review about Draft PSRT in Phase II..... **Error! Bookmark not defined.**
Appendix 13. Original Validation Plan in Main Study ..... **Error! Bookmark not defined.**
Appendix 14. Fit Statistics of Items in the Draft PSRT ..... **Error! Bookmark not defined.**
Appendix 15. Summary of Final Item Outcomes in the Draft PSRT **Error! Bookmark not defined.**
Appendix 16. PSRT, Characterisation Scheme, and Rasch Fit Statistics **Error! Bookmark not defined.**
Appendix 17. Item Format of Items in the PSRT..... **Error! Bookmark not defined.**
Appendix 18. Item-Curricular Theme Correspondence in the PSRT **Error! Bookmark not defined.**
Appendix 19. Table of Specification of PSRT by Knowledge Types, Science Practices and Cognitive Demand..... **Error! Bookmark not defined.**
Appendix 20. Table of Revision History ..... **Error! Bookmark not defined.**
Appendix 21. Summary of PSRT Initial and Revised Rasch Fit Statistics **Error! Bookmark not defined.**

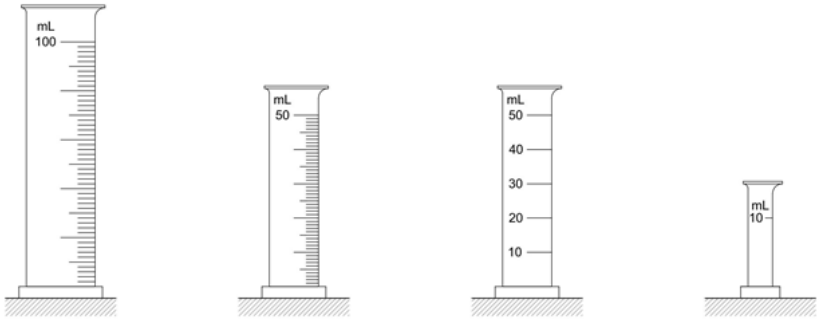
Appendix 22. Summary of PSRT Initial and Revised Rasch Fit Statistics**Error! Bookmark not defined.**
Appendix 23. Item Features of Revised PSRT Arranged in Order of Difficulty..... **Error! Bookmark not defined.**
Appendix 24. Features of 47 Items Without Changes in the Revised PSRT ..... **Error! Bookmark not defined.**

Chapter 1 Why Study Scientific Reasoning in Primary Children?

1.1 Introduction

How do primary school pupils reason scientifically? As an illustration, imagine an 11-to-12-year-old science pupil attempting the science question in Figure 1.1.

The diagram below shows four types of a piece of equipment.



A B C D

(a) Identify the equipment shown above. _____

(b) Which type of equipment will measure 15 cm^3 of water most accurately?

Circle your answer below.

A / B / C / D

Figure 1.1. Question made up of two items adapted from a Singapore primary school paper.

To answer the question correctly, the pupil will first need to identify and understand the function of the equipment. He/she must then reason that the smallest division on the scale of the equipment determines the precision of the reading. Using contextual cues (e.g., from the markings on walls), the pupil next determines which one of the four scales is most suitable for measuring the prescribed water volume. Despite the focus on quantitative reasoning processes, resolution requires the interactive engagement of these processes with the context and scientific content (e.g., functions of a measuring cylinder and concepts of volume and accuracy). While the underlying reasoning is widely applicable to other non-scientific areas of study, such as mathematics and financial literacy, the collection and

interpretation of readings from a scale necessitates an appreciation of the scientific context (Flavell, Miller, & Miller, 1993).[1] Therefore, though the science content assessed in the question is straightforward, what is less clear is how this imagined pupil reasons with scientific concepts to work out the answer.

Despite extensive research on reasoning processes, this chapter argues that limited understanding and consensus of the kinds of performances that exemplify scientific reasoning have affected the development and evaluation of test instruments. Well-developed substantive theories underlying psychological constructs are necessary for designing valid and reliable items and measurement methods (Borsboom, 2006). However, it is uncertain if and how established learning and theoretical measurement frameworks guide the design of current classroom scientific reasoning tests. Many of these tests are not validated and therefore their psychometric properties are unknown. Also, very few of these tests are designed solely for assessing primary school pupils. This study attempts to fill these gaps by developing and validating a scientific reasoning test for 11-to-12-year-old primary school pupils.

1.2 Background

Over the last decade, attention to the development and importance of reasoning in science has significantly intensified (Koerber, Mayer, Osterhaus, Schwippert, & Sodian, 2015; Osterhaus, Koerber, & Sodian, 2017; Zembal-Saul, McNeill, & Hershberger, 2012). This heightened interest in scientific reasoning, defined simply at this juncture as a set of higher-order mental abilities to critically "think with and about science knowledge" (Hogan & Fisherkeller, 2005, p. 95), is attributed to complex cognitive functions expected of 21st century citizens. Underlying such expectations is the premise that citizens with well-developed reasoning abilities engage more competently in science-related civic and social discourses (Rudolph & Horibe, 2016), and manage non-routine and ill-defined science problems more effectively (Hill, 2007). Moreover, there is empirical evidence that bolstering

[1] Quantitative reasoning will be discussed in the literature review (Chapter 2, subsection 2.5.2).

the scientific capacity of a nation's general workforce promotes social and material well-being as well as enhancing economic outcomes (Hanushek & Woessmann, 2010, 2012). For example, Jaffe et al. (2013) found that countries that invested heavily in scientific enterprises, such as research, generated more economic wealth (Jaffe et al., 2013). Thus, there are personal, social, and economic imperatives for helping individuals develop their scientific reasoning capacities.

One way to develop the desired scientific competencies is to nurture them early in school children (Harlen, 2001; National Research Council, 1996). In an influential report of a seminar series funded by the Nuffield Foundation in the UK, researchers and educators argued that early exposure is important because the process of acculturation to scientific ideas, practices, and attitudes is necessarily lengthy (Millar & Osborne, 1998). Moreover, research with young children, some of whom were as young as five, has found that they intrinsically develop scientific ideas about the world they experience (Harlen, 2007). That this development of ideas occurred even in the absence of instruction led Harlen (2008) to reiterate a commonly raised finding: the origins of scientific thinking began before formal education. The findings of some research have gone further and suggested that initial precursors of scientific reasoning emerged as early as infancy and progressed significantly after that (Schulz & Gopnik, 2004). Shaping and nurturing these rudimentary abilities appropriately into more advanced and wide-ranging competencies during children's formative years in primary schooling is a critical step for schools to undertake (Harlen, 2010).

Many countries have endorsed this view by frequently changing their primary science education programmes to foster reasoning and thinking skills in pupils (Abd-El-Khalick et al., 2004; Cleary, 2008). Concurrently, these countries' curriculum authorities engaged in scrutiny and international comparison of primary science curricula, assessment, and prescribed thinking and reasoning strategies (Achieve, 2010). For instance, the United States-based Next Generation Science Standards (NGSS) have generated both excitement and controversy outside of the country (Pruitt, 2014). Designed for pupils from as young as four (attending kindergarten) through to 12th grade (17 to 19 years old), the innovative

premise of the NGSS lies in their multi-dimensional knowledge and reasoning (Gross et al., 2013). Drawn from a wide range of research literatures, such as cognitive science, philosophy of science, and science education, the NGSS are a notable attempt to identify attributes of scientific thinking by explicitly stating the performance expected of pupils at various grade levels (NGSS Lead States, 2013).

Promoting thinking and reasoning in young children is not limited to science education programmes and national initiatives (e.g., NGSS). It is also heavily emphasised in international testing programmes, such as the Trends in Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). Questions generated by these large-scale international testing programmes have increasingly incorporated the assessment of complex cognition and reasoning in science over successive testing cycles (Jones, Wheeler, & Centurino, 2013; OECD, 2017). There is also robust evidence that the results of these testing programmes have influenced science educational reforms and calls for restructuring of the science curriculum in some countries (Duit, 2007; Kamens & Mcneely, 2009; Martens, Niemann, & Teltemann, 2016).

Despite these efforts, some studies have found that not all pupils demonstrated the expected scientific reasoning skills at the usual instructional junctures, or even applied the skills appropriately (Adey & Csapó, 2012; Zhai, Jocz, & Tan, 2014). For instance, a two-year research programme with 153 fourth-grade (Year 4) children found that they did not demonstrate hierarchical reasoning, even with classroom instruction (Chin, Dohmen, & Schwartz, 2013). However, after participating in an intervention programme utilising specialised computer software, there was improvement in their reasoning abilities with content involving scientific hierarchies, such as inferring the taxonomic grouping of animals by shared properties. In another study, researchers found that Chinese university physics students performed similarly to their North American counterparts on a test of logical reasoning: the Lawson's Classroom Test of Scientific Reasoning. However, the same Chinese students outperformed their peers on a physics test, presumably attributable to their more rigorous curriculum (Bao et al., 2009). These results suggest that the growth of

scientific reasoning abilities is not necessarily commensurate with instruction, greater content mastery, or developmental maturation.

Collectively, studies examining scientific reasoning abilities among school-going participants indicate that reasoning difficulties are affected by some key factors (Korom, Németh, Antal, & Csapó, 2012; Pellegrino, DiBello, & Goldman, 2016; Ruiz-Primo et al., 2012). These include the conceptual complexity of theories, the cognitive demand of the assessment, and pupils' naïve misconceptions (Haglund, Jeppsson, & Andersson, 2012). Misconceptions occur when children's notions about scientific phenomena are incongruent with established understanding (Osborne, 1985). Harlen (2008) reported that pupils often held on to erroneous ideas because these ideas agreed with their experiences and were reasoned to make more sense than the abstract notions taught to them in the classroom. Further, once these misconceptions become entrenched in pupils' mental schemas, they become resistant to instruction (Campanario, 2002).

It is thus vital to assess the reasoning abilities of pupils with science knowledge and skills early and accurately (Harlen, 2008). Diagnosing reasoning flaws forms part of the "instruction, support and practice" (Morris, Croker, Masnick, & Zimmerman, 2012, p. 62) repertoire for nurturing scientific reasoning in classrooms. Once identified, the appropriate instructional and pedagogical interventions can help educators address these reasoning gaps. For this purpose, it is crucial for an assessment instrument to be based on a sound theory of scientific reasoning, and to produce valid and reliable inferences (Osborne, 2013). Unfortunately, despite the existence of multiple frameworks and reasoning tests – some of which are close to a century old (see e.g., Burmester, 1953; Downing, 1928) – very few of the available classroom tests are suitable for primary school pupils, and the existing tests are problematic. The next section describes the limitations and inadequacy of current reasoning tests and discusses how this thesis can contribute by developing a scientific reasoning test for primary-aged pupils.

1.3 Rationale

1.3.1 Current Scientific Reasoning Tests

A variety of scientific reasoning tests exist, which may broadly be placed in two categories, depending on the scale of their administration. The first category includes large-scale assessments involving thousands of test-takers, such as national tests (e.g., National Assessment of Educational Progress Science Assessment) and international tests (e.g., TIMSS, PISA). The second category includes classroom-orientated tests administered to smaller groups on an ad-hoc basis. Unlike the large-scale tests, few classroom-orientated tests have been systematically examined to establish the validity of their reasoning frameworks and the reliability of the test responses (Pellegrino & Wilson, 2015).

1.3.2 Limitations of Current Scientific Reasoning Tests

Opitz, Heene, and Fischer (2017) reviewed studies of 38 large-scale and classroom-based scientific reasoning tests and noted the paucity of inquiries made about the validity of the scientific reasoning construct being assessed. Among the studies, only six used other reasoning tests as measures for establishing concurrent validity, and only four attempted to check divergent validity by concurrently assessing cognitive attributes such as intelligence. Opitz et al. (2017) also highlighted the inconsistent deployment of checks on the tests' other psychometric qualities such as reliability and content validity. Given the scarcity of evidence about the validity of the construct of scientific reasoning underlying many of these classroom-based tests, the authors questioned the results and claims about test-takers' reasoning abilities through their performance on the tests. Their conclusions – that current classroom reasoning instruments lack substantive and empirical evidence to support the interpretation and usage of the scores – accorded with the observations of other researchers (Kind, 2013; Osborne, 2013). Understanding the performance of pupils through a valid interpretation of their scores is important to support classroom teaching and learning (Pellegrino et al., 2016).

Osborne (2013) further attributed some of the uncertainties about the quality of current tests to weak consensus about the scientific reasoning construct. One contested area is the role of domain knowledge, or science disciplinary content, in eliciting scientific reasoning. He cautioned that very few contemporary studies of classroom assessments have investigated reasoning with tasks that require test-takers to draw on scientific knowledge. Minimising the influence of knowledge while assessing scientific reasoning is known as a "knowledge-lean" or "domain-general" approach (Perkins & Salomon, 1989; Zimmerman, 2000). This approach originated from the seminal research of Piaget and his collaborators, who postulated that reasoning abilities consisted of general cognitive heuristics that are knowledge-independent and transferable across contexts (Inhelder & Piaget, 1958). This influential perspective on scientific reasoning has been criticised for conflicting findings, which suggested that domain knowledge and contextual factors have significant impact on the quality and extent of reasoning (Kind, 2013; Pellegrino & Hilton, 2012; Zeineddin & Abd-El-Khalick, 2010).

Similar arguments against the conception of scientific reasoning as composed of a set of general reasoning skills to be taught systematically are put forth on philosophical and pedagogical grounds (Finley, 1983). Moreover, some studies have called into question the validity of results from existing test instruments that adopt a "domain-general" approach (Hofstein & Mandler, 1985; Stefanich, Unruh, Perry, & Phillips, 1983). The literature review in the next chapter (subsections 2.2.1 and 2.2.2) comprehensively reviews these arguments.

In sum, this study adopts a domain-specific view and concurs with Osborne's assertion that "scientific reasoning requires reasoning with scientific knowledge" (Osborne, 2013, p. 269). As an indication of how this perspective on reasoning with knowledge is gaining ground, it is notable that large-scale assessments, such as TIMSS and PISA, assess pupils' reasoning with "knowledge-rich" items (Jones et al., 2013; OECD, 2017). Unfortunately, these large-scale assessments have not advanced understanding of how science content affects reasoning in science and vice versa (Liu, Lee, Hofstetter, & Linn, 2008). Instead, they focus on comparative levels of pupil achievement across countries (Hopfenbeck et al.,

2017). With few available insights into the relationship between scientific knowledge and scientific reasoning, there has been limited progress in the development and validation of suitable instruments to support classroom instruction (Schulz & Gopnik, 2004).

Besides the contested view about the role of science content knowledge in reasoning, there is little agreement among researchers about the other theoretical aspects of the scientific reasoning construct (Kind & Osborne, 2017). Contemporary scientific reasoning conceptualisations include scientific argumentation (Berland & McNeill, 2010), subsumed skills from the nature of science literature (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) and a broad-based set of skills such as problem identification and the ability to question and communicate results (see e.g., Fischer et al., 2014). Given the disparate conceptual frameworks underlying many classroom-based tests, it is unsurprising that Opitz et al.'s (2017) review of studies on similar tests reported a lack of common indicators used to measure scientific reasoning skills. They warned that this hampers the comparison of outcomes measured by the tests and the ability to predict future performance.

Another limitation of current classroom-based reasoning tests is the paucity of validated instruments suitable for children aged six to 13. Of the 34 classroom tests examined by Opitz et al. (2017), nine were intended for this age group. However, only one of these nine had undergone validation (see Gobert, Sao Pedro, Raziuddin, & Baker, 2013). In this single test, pupils' responses were computer-scored. Also, a search of selected databases for relevant tests published over the past ten years only surfaced two additional validated classroom-based scientific reasoning tests.[2] Table 1.1 compiles the two tests that were found with those identified by Opitz et al. (2017).

[2] These databases were ERIC and PsychINFO. The search strings consisted of all possible combinations of the terms "scientific reasoning" and "scientific thinking", together with the terms: "test", "assess*", or "scale".

Table 1.1

Validated scientific reasoning tests categorised by participants' age range and task

Approximate Age Range of Participants (Years)	Tests with Knowledge-Lean Tasks	Tests with Knowledge-Rich Tasks
6–13	Science-P (Koerber et al., 2015; Mayer, Sodian, Koerber, & Schwippert, 2014) Springs task (Linn & Rice, 1979) Test of enquiry skills (Fraser, 1980) Test of logical thinking (Tobin & Capie, 1981) (ages 12-14) Test of science process skills (Molitor & George, 1976)	Detector – Inquiry Intelligent Tutoring System (Gobert et al., 2013) (ages 12-14)
14–19	Lawson's Classroom Test of Scientific Reasoning (Han, 2013; Lawson, 1978; O'Donnell, 2011) Experimental problem-solving (Ross & Maynes, 1983) Test of science process skills (Rabacal, 2016) Test of integrated process skills I & II (TIPS) (Baird, Perry, & Simon, 1989; Burns, Okey, & Wise, 1985; Dillashaw & Okey, 1980) Scientific reasoning test, version 9 (Sundre, 2008)	Evidence-based reasoning assessment system (Brown, Furtak, Timms, Nagashima, & Wilson, 2010; Brown, Nagashima, Fu, Timms, & Wilson, 2010) Abilities in scientific inquiry (Nowak, Nehring, Tiemann, & Upmeier zu Belzen, 2013) Chemistry concept reasoning test (Cloonan & Hutchinson, 2011) Online portfolio assessment and diagnosis scheme (Su, Lin, Tseng, & Lu, 2011)

(continued)

Table 1.1 (continued)

Approximate Age Range of Participants (years)	Tests with Knowledge-Lean Tasks	Tests with Knowledge-Rich Tasks
	Test of scientific literacy skills (TOSLS) (Gormally, Brickman, & Lutz, 2012)	Processes of biological investigations test (Germann, 1989)
	Constructive inquiry science reasoning skills (Weld, Stier, & Mcnew-Birren, 2011)	Rubric (Timmerman, Strickland, Johnson, & Payne, 2011)
		Practical tests assessment inventory (Tamir, 1974)
		Science process skill test (Feyzíoglu, Demírdag, Akyildiz, & Altun, 2012)
		Quantitative assessment of socio-scientific reasoning (QuASSR) (Romine, Sadler, & Kinslow, 2017)
		Instrument of scientific literacy assessment on the cycle theme (Rusilowati, Kurniawati, Nugroho, & Widiyatmoko, 2016)

Note: Four of the 38 tests in the study by Opitz et al. (2017) were not published in English, and hence are excluded from Table 1.1.

1.4 Contributions of Thesis

In summary, despite the increasing emphasis on the development and assessment of reasoning skills in young children, there remains a lack of consensus about the scientific reasoning construct itself. A premise of this thesis is that the limited agreement has affected the construction and evaluation of test instruments that can inform the extent of learning and instructional effectiveness in science. Another gap in current research is understanding how the primary science curriculum impacts on scientific reasoning (Kim, Tan, & Talaue, 2013; Passmore & Stewart, 2002; Perkins & Salomon, 1989). This study attempts to address these issues by developing and validating a scientific reasoning test for 11-to-12-

year-old primary pupils in Singapore. The knowledge-rich items in the test – the Primary Scientific Reasoning Test (PSRT) – have been designed using a novel conceptualisation of scientific reasoning, arising from recent research and drawn from various bodies of literature. A variety of psychometric and qualitative methods advocated by contemporary notions of validity are used to evaluate the PSRT. If found appropriate, the test and the underlying construct of scientific reasoning are intended to inform and guide primary school science teachers, as well as test developers.[3] The following section states the overarching research question and the research aims.

1.5 Research Question and Research Aims

The overarching research question of this study is: "To what extent does the PSRT validly measure scientific reasoning?"

There are two research aims. These are to:

(1) produce evidence that the PSRT is measuring a unidimensional construct; and

(2) examine whether the underlying construct measured in the PSRT conforms to the hypothesised construct of scientific reasoning.

This study adopts the definitions and operationalisations of validity recommended by the *Standards for Educational and Psychological Testing,* hereafter referred to as the *Standards*. The *Standards* refers to validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). The sample in this study consisted of Year 6 pupils from Singapore. The reason for selecting these pupils was the ease of access to local schools.[4] A brief description of the Singapore Primary Science Curriculum is given in Appendix 1 of this thesis.

[3] An evaluation of the consequences of test use is beyond the scope of the thesis.
[4] The access came about as the author of this thesis is a science assessment specialist with the Singapore Examinations and Assessment Board (SEAB).

1.6 Organising Themes in the Thesis

There are four major themes in this thesis. These themes are drawn from Onwuegbuzie, Bustamante, and Nelson's (2010) instrument development and construct validation framework. The first presents the theoretical and methodological frameworks guiding the study and consists of three chapters: an introductory chapter (Chapter 1), the literature review (Chapter 2), and the research methodology (Chapter 3). This third chapter introduces the validation framework and explicates the mixed-methods approach, using qualitative and quantitative data to address the overarching research question and aims.

The second theme describes the design, review and piloting of the draft test instrument – the Draft PSRT. This theme runs across chapters 4 to 6, which cover three of the four research phases. Chapter 4 describes the activities in phase I, such as the conceptualisation of the scientific reasoning construct, and the initial design of the draft test instrument and its questions. The remaining phases (II and III) implemented an expert review and pilot study respectively. Chapters 5 and 6 discuss these phases in detail, providing information about the participants, instruments, and procedure, as well as the findings. The third theme describes the process and results of the validation in phase IV: the main study. In this final phase, both quantitative and qualitative data from Year 6 pupils were collected to address the two research aims. Accordingly, chapters 7 and 8 present the analysis and findings in relation to each aim. The final theme of the thesis is presented in the last chapter (Chapter 9). It offers concluding inferences about the construct validity of the PSRT through integration and analysis of the collective evidence gathered over the four phases. Chapter 9 also discusses the implications of these conclusions for theory and practice. The chapter ends by discussing the limitations of the study and proffers possibilities for future research.

In summary, this study arose out of concerns about the lack of a sound theoretical basis in many scientific reasoning tests. As will be elaborated in the next chapter, the first step to address this gap was to review existing conceptions of scientific reasoning.

Chapter 2 Literature Review: What is Scientific Reasoning?

2.1 Introduction

Science is an invention of humankind (Huff, 2003). Historical records trace the origins of science to early cultures of antiquity (Lindberg, 2007), but interest in the cognition of a scientific mind was only kindled when ancient Greeks endeavoured to understand their natural world (Feist, 2006). Greek philosophers pioneered structured reasoning, such as deduction and syllogism, as the basis for evaluating hypotheses and arguments (Cohen, 2010). The emergence of modern-day science during the "Scientific Revolution" in the 1600s and 1700s saw renewed interest in and efforts to elucidate the nature of the mental processes engaged in experimentation and inquiry (Hall, 1983; Henry, 2008). Francis Bacon (1561–1621), for instance, was the first to describe the significance of scientific inductivity (Urbach, 1987). From then until the start of the 20th century, there were multiple perspectives on scientific thinking (Crombie, 1994; Hacking, 1992; Ng, 2017).

Each of these perspectives corresponded to historical influences, which arose from significant scientific discoveries and favourable factors in the social environment (Crombie, 1994). For example, taxonomic and evolutionary reasoning, which involved the grouping of objects based on some organising principles (such as biological lineages), gained prominence in the 19th century due to a conducive social and scholarly environment in Europe (Huxley, 2003). However, systematic inquiries into the cognitive mechanisms underlying scientifically engaged minds only began in the early 1900s (Dunbar & Klahr, 2012; Tweney, Doherty, & Mynatt, 1981). Wertheimer's (1945) seminal investigation into Einstein's cognitive abilities, using Gestalt theories, initiated a series of studies seeking to understand how scientists rationalise and analyse data. Another influential study reported the dominant personalities of scientists (Roe, 1953).

From the 1950s onwards, contemporary insights into the scientific reasoning of children and adults came from cognitive developmental research (Klahr, Chen, & Toth, 2001; Klahr &

Simon, 1999). Not surprisingly, findings from this field had the greatest influence on conceptions of scientific reasoning in science education (Flavell et al., 1993). Two well-established and widely applied developmental perspectives on scientific reasoning are Piaget's (1958) logic-developmental model, and Gagné (1968) and Bruner's (1956) inductive-based ideologies (Finley, 1983; Kuhn, Amsel, & O'Loughlin, 1988; Sanderson & Kratochvil, 1971). The following sections discuss both perspectives and their contributions to the understanding of scientific reasoning. The discussion then moves on to explain how dissatisfaction with the logic and inductive orientations led the education and research communities to search for alternative or supplementary accounts.

Three broad post-logic and post-inductive accounts of scientific reasoning are next described, all of which share a consensus that scientific reasoning involves the interaction of multiple cognitions such as deduction and quantitative reasoning. Subsequently, a novel framework of scientific reasoning is introduced as a synthesis of the collective accounts. As further theoretical support for the proposed framework, the discussion follows on by describing emergent insights from philosophy, psychology, and cognitive history. Drawing on these understandings from psychological and cognitive research, the concluding section outlines the reasoning types associated with scientific activities that are commonly required of pupils learning primary science.

2.2 Contemporary Conceptions of Scientific Reasoning

2.2.1 Logic as Scientific Reasoning

Piaget's theory describes reasoning and mental capabilities during four periods of physical development (Gold, 1987; Inhelder & Piaget, 1958). In the first *sensorimotor* period, the baby gains sensory and motor functions to help him/her respond to their surroundings. As they develop from birth to approximately two years of age, their actions progress from spontaneous reflexes to intentioned sensory-motor reactions. The next two periods – *pre-operational* and *concrete-operational* – last approximately nine years. As the toddlers grow older from two to 11 years old, they acquire and demonstrate more complex intellectual

abilities. These capabilities help the children to develop material, spatial, and temporal understanding.

Of interest to this research is the next and final developmental period, from 11 to 15 years of age. Cognitive activities in this *formal-operational* period are "second-order operations" because the adolescent utilises competencies gained from earlier developmental periods to manage and operate complex information. Notably, the adolescent's mental schemas develop from concrete ideas and concepts to encompass plausible and theoretical notions. This capability means the adolescent can use deductive reasoning to understand hypothetical situations and outcomes (Ennis, 1982). He or she is further able to weigh the relative merits of evidence from various lines of investigation and draw reasoned conclusions. Jointly, these competencies form the rudimentary basis of scientific thinking (Flavell, 1963; Flavell, Miller, & Miller, 2002). Seeking to develop these basic skills further in pupils, the science education and teaching communities actively investigate the role of reasoning strategies, such as the control-of-variable technique, in the learning of science (Fischer et al., 2014; Kuhn, 2011).

Challenges to Piagetian theory came from two fronts – logical and empirical (Kuhn et al., 1988; Miller, 2010). On the first front, researchers highlighted serious irregularities in the propositional logic model proposed by Piaget to denote mental processes using four "group operators" and 16 "binary operations" (Braine & Rumain, 1983). Also, similar studies conducted later were unable to reproduce the results of Piaget and his collaborators (Keating, 1980). Other studies produced strong evidence and findings that showed that the pattern of cognitive growth outlined in Piaget's theory is not universally or predictably attained by children (Beilin, 1992; Flavell et al., 2002). For instance, research showed that personal and external factors, such as motivation and instruction, affect transferability of cognitions across analogous contexts (see, e.g., Linn, Clement, & Pulos, 1983). These findings are in direct contrast to the Piagetian view of "context-free" cognitive performance (Millar & Driver, 1987).

Overall, criticisms of Piagetian concepts arose from conflicting findings evidencing uneven and context-dependent cognitive growth in children (Flavell et al., 1993; Kuhn & Franklin, 2006). Regardless, some developmental psychologists argued that Piaget's insights into cognitive development were broadly accurate (Kuhn, 2008). Cognitive differences suggested between young and older children have found some support in the literature. These differences include the use of judgement (outward perception vis-à-vis inference to underlying reality), the ability to focus attention (centring as opposed to decentring), and the reversibility of thinking patterns. These qualitative differences have implications for the assessment of scientific reasoning (Adey & Csapó, 2012). For instance, the design of tasks (e.g., duration) needs to be age-appropriate.

In closing this subsection, it is worth mentioning that Keating (1980) noted Piaget's assertion that "there is more to thinking than logic" (Inhelder & Piaget, 1958, p. 335). We turn now to an alternative conception based on inductive reasoning. This reasoning involves the uncovering of rules that govern a series of events, which are then generalised to similar events or other unknown but observed, or yet to be observed occurrences (Dunbar & Klahr, 2012). Embodying this induction-as-scientific reasoning conception are the theories of two of Piaget's contemporaries – Gagné and Bruner. As will be discussed in the following subsection, their theories were greatly influential in shaping the conception of reasoning in the teaching and learning of science.

2.2.2 Induction as Scientific Reasoning

Gagné's "Process" Approach: Gagné (1963) argued that the development of inquiry and intellectual skills in pupils required enduring and context-transferable cognitions (Hricko, 2008). Accordingly, he devised a series of science activities to foster "process skills", or what Sanderson and Kratochvil (1971) termed, "intellectual tools of science" (p. 6). His approach served as the basis for problem-solving and scientific inquiry in research, instruction, and curriculum development for three decades – well into the 1980s (Fields,

2000). His theory also served to bridge the then predominant behavioural tradition and the emerging cognitive paradigm (Case & Bereiter, 1984; Sanderson & Kratochvil, 1971).

A well-documented curriculum initiative that applied Gagné's theory occurred in the mid-1960s (Fields, 2000). In this programme, science educators taught primary and secondary pupils process skills, with the view that their improved general cognition would enhance learning of scientific knowledge (Lawson, Karplus, & Adi, 1978). As shown in Table 2.1, two categories of process skills were taught. Unlike the basic processes taught to the younger pupils, the older children were exposed to instructional activities that fostered higher-order cognitions in hypothesis-setting and investigations (Finley, 1983).

Table 2.1

Gagné's basic and integrated process skills

Process	
Basic	Integrated
Observing	Operational definitions (posing questions)
Classifying	Formulating hypotheses
Describing	Controlling variables
Communicating	Interpreting data
Using numbers	Experimenting
Measuring	
Using space/time relationships	
Predicting	
Concluding	

Note. Adapted from Sanderson and Kratochvil (1971).

The process approach lost its influence in education as criticisms grew about its inductive basis of learning (Finley, 1983). Lipton (2004) identified two well-established philosophical arguments against inductive methods. In the first, he highlighted the difficulty of identifying the principles, rules, and methodology associated with this form of reasoning. Without knowledge of how inductivity works, it is not possible to explain the functioning of scientific processes, such as hypothesis generation. Millar and Driver (1987) concurred and asserted

that Gagné's processes have no historical, sociological, or philosophical support. Rather, they argued that these are, "aspects of our general cognitive functioning from infancy and that science can lay no special claim to them" (Millar & Driver, 1987, p. 36). The second argument against inductivity is its "underdetermined" nature, which means that its inferences can never be proven to always hold true (Lipton, 2004). This is because the premise (or evidence) for the inference (or claim) comes from a series of observations, which must be assumed to operate under all circumstances.

Also criticised are the empiricist roots of the process approach, which claim knowledge to originate from what can be perceived, experimentally verified, and measured (Finley, 1983). Instead, modern philosophers and cognitive scientists argued that experiences affect and shape the acquisition of conceptual knowledge (Hempel, 1965; Resnick, 1983). For instance, deciding the saliency of observations requires criteria drawn down from prior knowledge (Driver, 1983); hence, the perceptions of an expert will differ from those of a novice. From the pedagogical angle, there are doubts about the applicability of the process approach (Millar & Driver, 1987). It is unclear, for instance, how the teaching and assessment of discrete process skills can occur, as performance depends on the task. Moreover, no substantive theories about progression in these skills exist.

Regardless of criticisms of the process approach, inductivity is nevertheless important for knowledge acquisition and application (Adey & Csapó, 2012). Historically, too, some scientific discoveries have attributed their cognitive basis to inductive processes (Dunbar & Klahr, 2012). In education, Bruner's (1956) work illuminated the importance of the inductive mechanism that underlies concept attainment and knowledge representation in children.

Bruner's Knowledge Representation and Concept Attainment: Bruner et al. (1956) suggested that the formation of concepts involves decision-making and categorisation processes (Lutkehaus & Greenfield, 2003). There is a "search for and testing of attributes" (Bruner et al., 1956, p. 233) to group initial notions and objects. Inferences about shared associations during categorisation subsequently shape understanding of conceptual notions. In efforts to characterise the mental processes of decision-making, Bruner and his

collaborators studied many examples, and identified cognitive strategies used by individuals during concept formation. These strategies share three broad features: (1) frequent exposure to initial ideas or a variety of relevant instances; (2) identification and grouping of attributes among ideas or instances; and (3) deciding whether the grouping is correct, incorrect, or indeterminate.

Overall, these strategies imply that, for successful and valid categorisation during concept formation, learners must experience both positive and negative cues or examples. This method of gaining knowledge, through the construction of a general rule for an explanatory model from specific instances, is akin to inductive reasoning or concept induction (Joyce & Weil, 1972; Joyce, Weil, & Calhoun, 2011; Klahr & Dunbar, 1988). Significantly, as individuals call on their store of historical cues, it is possible for them to reason and understand incorrectly. Besides, numerous classifications and decisions are involved, affecting the outcome (Bruner et al., 1956). Consequently, understandings of the same phenomena will differ across individuals. In effect, attainment of concepts also relies on the personal and historical experience of the individual.

Bruner also contributed to an understanding of knowledge representation in children (Lutkehaus & Greenfield, 2003), who he suggested view knowledge in three modes: enactive, iconic, and symbolic (Bruner, 1964). These modes are forms of cognitive processing and emerge in a specific developmental sequence. The enactive mode appears first in an infant, whose interactions with his or her environment (e.g., manipulating objects) involve the activation of motor and neuronal functions. Iconic representation develops next; this describes the engagement of mental imagery unaccompanied by physical actions or the object of the imaging. In this mode, the growing child is capable of visualising objects and their properties. The most cognitively advanced mode is symbolic representation. In this condition, the individual can use symbol systems, such as mathematical notation and language, to encode information. Knowledge represented by symbolic systems has remote and arbitrary features (Schunk, 2014). Developmentally, this mode manifests last, and is the preferred mode for communicating understanding of abstract notions (Bruner, 1964). In

the long run, individuals retain the ability to represent knowledge among the modes and switch accordingly.

Brunerian ideas about the role of induction in the "continual deepening … to progressively more complex form[s]" (Bruner, 1960, p. 13) of reasoning, which occur during concept and knowledge development, were influential in science education (Schunk, 2014). Applications of these ideas, some of which are known as the "spiral curriculum" and "discovery learning", have appeared regularly as part of curricular and pedagogical reforms over the past 60 years (Gibbs, 2015). However, the role of induction in teaching and learning is still not well understood, despite an increase in relevant classroom research (see, e.g., Klauer, Willmes, & Phye, 2002; Sanz de Acedo Lizarraga, Sanz de Acedo Baquedano, & Oliver, 2010). For instance, researchers are unsure of how a pupil's background and the nature of assessment tasks impact on the elicitation of inductive reasoning (Csapó, 1997). Issues have also surfaced regarding the practicality and usefulness of applying inductive-based pedagogy (e.g., hands-on activities) in the classroom (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Gibbs, 2015).

In sum, logic and inductive perspectives presented theoretical and empirical challenges to the models of scientific reasoning (Kuhn et al., 1988). Critics charged that both forms of reasoning, while useful, are inadequate to improve learning outcomes (Hamers, de Koning, & Sijtsma, 1998; Lawson, 1985; Millar & Driver, 1987). In particular, accumulating evidence conflicted with the "context-free" and "transferable" notions of thinking skills characteristic of the Piagetian and inductive approaches (Kuhn et al., 1988; Perkins & Salomon, 1989). Findings from cognitive science research also suggest that scientific reasoning can involve multiple types of cognition (e.g., extrapolation), depending on reasoning contexts (Kuhn et al., 1988; Sodian & Bullock, 2008; Zimmerman, 2007). Later research, in the late 1980s, moved away from focusing on the reasoning mechanism to investigate the roles of content knowledge, science methods, and social factors in learning contexts (Erduran & Dagher, 2014c). These investigations resulted in revised accounts of scientific reasoning (Koslowski,

1996; Millar, Lubben, Gott, & Duggan, 1994; Sandoval & Reiser, 2004). The next section discusses these accounts.

2.3 Post-Logic and Post-Inductive Scientific Reasoning

2.3.1 Expert Reasoning as Scientific Reasoning

In response to criticisms of the prevailing Piagetian and inductive forms of reasoning in science education, a body of research arose to investigate the relationship between cognitive performance and knowledge by focusing on how children acquire and modify scientific concepts (Carey, 1985; Driver, 1983; Driver & Easley, 1978; McCloskey, 1983).

Among the early research in this area, the work of Posner, Strike, Hewson, and Gertzog (1982) was especially influential (Vosniadou, 2013). These scholars extended Piagetian ideas to suggest that conceptual change occurs through “accommodation”, which is the initiation and process of replacing misconceptions with complete and accurate scientific ideas. The key mechanism driving change is the cognitive “conflict” that surfaces when the acceptable understanding is compared with the flawed alternative (Hewson, 1981). Later scholars challenged these assumptions, countering that pupils’ existing knowledge can be productive for developing mature and appropriate understanding (Smith III, diSessa, & Roschelle, 1994). Evidence that contradicted the postulated process of change surfaced, which suggested instead that the process is necessarily slow and deliberate (Caravita & Halldén, 1994). There is also a lack of clarity about the roles that motivation and contextual factors play during concept change (Pintrich, Marx, & Boyle, 1993).

Instead of dismissing the naïve scientific knowledge of pupils, further developments in conceptual change research attempted to identify the cognitive strategies and reasoning differences between novices and experts, with the aim of systematically reconciling these differences (Clement, 1998). Recent efforts have elucidated the development of cognitive mechanisms in tandem with conceptual changes in content-specific areas (see, e.g., Kapon & diSessa, 2012). However, there is no conclusive evidence about the effectiveness of such

efforts to reduce misconceptions (Vosniadou, 2014). To sum up, characteristics of scientific reasoning posited in this body of research, such as the systematic revision of conceptual understanding, are likened to the reasoning of authorities, such as experts.

2.3.2 Procedural Reasoning as Scientific Reasoning

Research on "procedural knowledge" arose in the early 1990s as a critical reaction to the process approach, which was then predominant in science education (Millar, 1989, 1991). Millar et al. (1994) argued that the process approach lacked the concepts, reasoning, and strategies associated with inquiry science. Besides manipulative skills, it is equally important to understand the aims and purposes of investigative tasks, as well as the criteria for judging the quality of empirical evidence (Gott & Duggan, 1996).

In effect, procedural knowledge, unlike "process" and "skills", underscores the relevance of content knowledge and represents the "thinking behind the doing of science" (Phipps & Duggan, 1995, p. 24). Science is reconceived as composed of conceptual and procedural understanding, with the notion of "skills" broadened to encompass task design, data management, and quantification methods (Gott & Mashiter, 1991). For instance, instead of demonstrating discrete skills, pupils could be tasked with determining the "fairness" of a test by analysing relevant evidence, such as identifying its key variables, analysing its results, or simply making a prediction based on their current understanding. In this revised conceptualisation of science, reasoning with procedural knowledge involves the critical evaluation of evidence using multiple methods.

Notwithstanding the important contributions of research about procedural knowledge to science education, transfer across to the general classroom is limited (Erduran & Dagher, 2014a). For instance, there are persistent reports that pupils' conceptions of a "scientific method" are at best superficial and at worst incorrect (Halwes, 2000; Woodcock, 2014). Another issue is a lack of clarity about the progression of specific forms of procedural understanding (Kuhn & Franklin, 2006). There is, for example, little information about how primary school children's ability to treat anomalous data differs from that of those in

secondary education (Chinn & Brewer, 1993). Another significant limitation is the lack of explanatory frameworks in claim-making from evidence using procedural reasoning (Goswami, 2004; Kuhn, 2011). The next subsection discusses how research on pupils' abilities to reason and evaluate evidence, using epistemic science knowledge, has addressed some of these issues.

2.3.3 Epistemic Reasoning as Scientific Reasoning

Around the turn of the millennium, reports surfaced suggesting that science education goes beyond content mastery and inquiry-based procedures (Driver, Learch, Millar, & Scott, 1996; Millar & Osborne, 1998). In this literature, calls were made to help pupils appreciate the nature and spirit of the scientific enterprise, or what Driver et al. (1996) referred to as, "*knowledge about science* as opposed to *scientific knowledge*" (p. 3, emphasis in the original). Researchers argued that direct instruction about these epistemic requirements or criteria builds pupils' capacities to justify scientific propositions and claims about scientific knowledge (Duschl, 2000; Grandy, 1997). Concurrent scholarship from diverse disciplines such as cognitive psychology, anthropology, philosophy, and linguistics also provided fresh insights into pupils' reasoning when drawing on relevant evidence to make claims (Garcia-Mila & Anderson, 2008) For instance, some studies showed young learners to be capable of reasoning with evidence when provided with sufficient support (Koslowski, 1996; Sandoval & Reiser, 2004; Zimmerman, 2000). However, substantial findings from research, and in particular from the extensive nature of science (NOS) literature, suggest that pupils find epistemic notions challenging and obscure (see, e.g., Sandoval & Reiser, 2004).

Other researchers were more specific and proposed that science education situate cognitive development in epistemic understanding alongside inquiry practices and norms valued by the scientific community (Duschl, 2008; Duschl & Osborne, 2002; Garcia-Mila & Anderson, 2008). Noting that scientific outcomes are often the result of critical and competing social discourses such as debates and dialogues, these researchers recommended exposing pupils to similar dialectic processes to hone their understanding

and sharpen their reasoning (Newton, Driver, & Osborne, 1999). This perspective of a close relationship between higher-order cognitions and dialogic abilities used in the defence, critique, and counter-proposal of claims, has led some researchers to draw parallels with Vygotsky's views about mediated cognition (see, e.g., Kim & Roth, 2014).

Vygotsky's Theory of Mediated Cognition: In his sociocultural theory, Vygotsky (1994) argued that *all* higher-ordered intellectual abilities originate from social interaction and events. The formation of conceptual thoughts and the act of reasoning with concepts are made possible by "verbal thinking" (Vygotsky, 1994, p. 213). Accordingly, he viewed semiotic and signification mechanisms, the most prominent of which are language and its symbolic forms (written, verbal, and graphical), as psychological "tools" for mediating and transmitting thinking in social settings, and thus important for learning (John-Steiner & Souberman, 1978). Qualitative improvement in thinking occurs because of the contact with these tools (Wertsch, 1998). These improvements, which include cognitive processes important in science learning such as problem-solving and experimentation, are mediated by language in social activities and settings (Rowe & Wertsch, 2004).

Thus, a direct application of Vygotsky's view to science education involves increasing instructional opportunities for pupils to witness and participate in group dialogue. Learning to assert an opinion and respond publicly are the precursors to claim-making and other, more advanced forms of scientific discourses such as argumentation (Berland & McNeill, 2010; Siegel, 1995). Argumentation is widely promoted as an effective approach to the social construction of scientific explanations, hypotheses, and arguments (Erduran, Ozdem, & Park, 2015; Kuhn, 2010). It is a specialised discourse for advancing both claims and counterclaims, each backed by their networks of evidence (Erduran, Simon, & Osborne, 2004; Kuhn & Udell, 2003). Given the interest in this approach as a pedagogical tool, many recent studies and curricular initiatives have investigated the impact of argumentation-based tasks and assessments on pupils' understanding of epistemic ideas in science (Clark & Sampson, 2008; Sampson & Clark, 2008). A fuller discussion about how the findings of these studies impact on this research, and the identification of basic epistemic ideas that

are appropriate for primary-aged pupils from relevant NOS literature appear in Chapter 4 (subsection 4.3.1). To sum up, recent scholarship in science education has called attention to the importance of claim-based reasoning using evidence. This third form of scientific reasoning involves the abilities to make and justify scientific claims and propositions.

2.4 Science as a Practice-based Activity

As recounted in the preceding discussion, there have been three broad post-logic and post-process developments in science education since the late 1970s. In brief, these developments foregrounded the importance of reasoning with disciplinary knowledge – conceptual, procedural, and epistemic. The three forms of knowledge address the ontological (what is the material world like?), causal (how can we explain what we observe?), and epistemic (how can we know or how can we be certain?) features of science (National Research Council, 2012; Osborne, 2011, 2013). To investigate these features, Osborne (2011) outlined a model of three overarching science practices – *developing explanations and solutions*, *investigation*, and *evaluating* – which scientists engage in to produce the three forms of knowledge. Taken together, the practices constitute the typical work scientists do to reason with disciplinary knowledge in any scientific research or investigation (Erduran & Dagher, 2014c; Kind, 2013).

Osborne (2014) attributed the notion of science as a collection of specific practices to Thomas Kuhn, whose 1962 seminal work *The Structure of Scientific Revolutions* drew attention to the social aspects of scientific endeavours. Science is perceived as a community-centric activity with consensually agreed rules, value system, and norms (Kuhn, 1962/2012). Members engage in common practices to develop theories, produce empirical data, and evaluate evidence using theoretical understanding to further the goals of science. Kuhn (2011) explained how these practices fit into a four-phase iterative “cycle of scientific investigation” (p. 503). In the first phase (inquiry), research goals, hypotheses, or questions are formulated based on theoretical gaps or interest. Consequently, the research objectives drive the selection of methodological approaches in data collection and analysis. Therefore,

the first phase of Deanna Kuhn's model incorporated two of the three general science practices – theory generation and data gathering – identified by Osborne (2011). In the remaining three phases (data analysis, inference, and argument), justification of claims occurs using supporting or disconfirming evidence. Collectively, these phases correspond to Osborne's (2011) third scientific practice – the critical evaluation of evidence.

2.5 Construct of Scientific Reasoning in this Thesis

The discussion in earlier sections presents an emerging notion of scientific reasoning. In this conceptualisation, **scientific reasoning[5] is the coordination between claims and evidence using disciplinary knowledge while engaged in science practices** (Kind, 2013; Kind & Osborne, 2017; Kuhn & Franklin, 2006). In the context of this thesis, a claim is potentially falsifiable using evidence such as empirical data (Kuhn & Pearsall, 2000). Coordination takes place during reasoning as an individual evaluates the claim by weighing the evidence using scientific knowledge and applying science practices (Kind, 2013; Kuhn, 2011). In this light, reasoning is a deliberate search for understanding and "something people *do* [emphasis in original], not something they *have* [emphasis in original]" (Kuhn, 2011, p. 372). Coordination is contingent on the context (e.g., task) and involves the interaction of multiple types of reasoning such as analogical and probabilistic reasoning (Korom, Németh, et al., 2012). Only when there is successful coordination ("do") will there be scientific understanding ("have"). This conceptualisation aligns with the broader notion of scientific reasoning as a set of cognitive, meta-cognitive, and meta-strategic skills (Schraw, 2011).

Regardless of the nature or complexity of reasoning (e.g., inductive versus probabilistic reasoning), this conceptualisation of scientific reasoning is jointly defined by the scientific nature of evidence and knowledge *and* the methods for gaining understanding. As explained in the previous sections, the subject of reasoning (evidence and knowledge) in

[5] In this research, "scientific reasoning" is equivalent to "scientific thinking". See Appendix 2 for a brief explanation of this stance.

this thesis is drawn from the three distinct forms of disciplinary knowledge identified from developments in science education. The thesis also draws on Osborne's model of the three science practices as a succinct description of all activities encompassed in a scientific inquiry cycle. Chapter 4 (subsections 4.3.1 and 4.3.2) further explicates the nature of the disciplinary knowledge and science practices expected of primary-aged pupils in this thesis. This explication includes definitions and behavioural descriptors. The following subsections discuss other theoretical support for the construct and conclude with a brief review of the specific reasoning that primary science pupils are expected to develop or demonstrate according to this thesis.

2.5.1 Theoretical Support for Construct

Theoretical support for the construct of scientific reasoning has come from three sources. The first two are recent psychological and philosophical models of thinking and learning in science (Kind, 2013; Osborne, 2013, 2014). For example, as an indication of the importance of the three science practices, psychological studies of scientific reasoning are frequently organised around hypothesis generation, experimentation, and inference (Klahr, 2000; Klahr & Carver, 1995; Klahr, Zimmerman, & Jirout, 2011; Kuhn et al., 1988). According to Klahr & Dunbar's (1988) influential scientific discovery as dual search (SDDS) psychological model, scientific discovery consists of a coordinated search through two problem spaces – a space of hypotheses and a space of experiments.

Experimental results guide the search in both spaces. Prior knowledge and the type of hypothesis further impact on the searches in the hypotheses and experiment spaces respectively. Evidence evaluation assesses the fit between theory and evidence produced in each space. Therefore, the SDDS model comprises three components: *search hypothesis space*, *test hypothesis*, and *evaluate evidence* (Klahr & Fay, 1993). These components jointly control the process of scientific discovery (Klahr, 2000).

Although Klahr and Dunbar (1988) investigated domain-general (not requiring disciplinary knowledge) strategies in their work with SDDS, they acknowledged that in a realistic

scientific reasoning context, substantive knowledge and the form of investigative strategy are mutually influential (Kind, 2013). Prior content knowledge will influence the initial hypotheses, the type of experiments conducted, and the analysis of results. Based on this acknowledgement, Kind (2013) argued that the three forms of knowledge – conceptual, procedural, and epistemic – are important for shaping reasoning.

As noted by Osborne (2013) and Kind (2013), philosophers of science arrived at a similar three-practice investigative structure of scientific reasoning from their analysis of authentic scientific episodes. To illustrate specifically, Giere, Bickle, and Maudlin (2006) recounted how the discoverers of the DNA double helix constructed DNA models with wire and fabricated metal plates as a visualisation methodology. During the constuction process, Watson and Crick incorporated conceptual information from text sources and made endless modifications to their model using data and evidence from many sources. Giere et al. (2006) explained that the scientists set up an initial hypothesis by modelling using artefacts. While not all scientists build scale models like Watson and Crick did, Giere et al. (2006) argued that in an abstract sense, most scientists can be said to be engaged in constructing mental models of some aspects of the world.

The account also showed that the two scientists engaged in reasoning that drew on conceptual, procedural, and epistemic knowledge as they performed numerous rounds of revisions from constant testing and evaluation of evidence against established understanding (Kind, 2013). Osborne (2013) suggested that the similarities in the scientific reasoning models proposed independently by philosophers and psychologists reflect the reality of scientific endeavours. According to him, the philosophical account stressed, “the reasons and the criteria that scientists advance for their judgements”, while the psychological perspective focused on, “the cognitive act or reasoning itself” (Osborne, 2013, p. 270).

Another source of theoretical support for the model of scientific reasoning hypothesised in this study came from an emerging branch of interdisciplinary scholarship – cognitive history

(Kind & Osborne, 2017). Cognitive historians strive to understand the mental processes underlying science practices by drawing insights from written historical records and artefacts such as published books, research notes and scientific instruments (Nersessian, 1995). They focused their attention on providing retrospective, naturalistic explanations of reasoning and thinking, and argued that these cognitive abilities operate and evolve from the assembly of unique, contingent factors during the pursuit of scientific knowledge (Netz, 1999). Unlike cognitive scientists who investigate the mental mechanisms characterising scientific reasoning, cognitive historians analyse the culture of science – its ideas, customs and social norms – to understand why cognitive behaviours manifest in the way they do.

Crombie (1994) provided one of the most extensive accounts of scientific thinking from a survey of 2500 years of European science (Kwa, 2014). His analysis showed that, regardless of the historical periods, scientific endeavours always produce original and unique ontological (scientific knowledge), epistemological (rules and principles to justify knowledge), and methodological (methods to collect knowledge) propositions and constructs (Hall, 1983; Kind & Osborne, 2017). Crombie also demonstrated that scientists in the past were actively constructing explanations and theories, investigating phenomena, and evaluating data and evidence, using six distinctive patterns of reasoning that he termed “styles of reasoning” (Hacking, 1992; Kwa, 2014). See Table 2.2 for examples of the scientific content, procedural approach, and epistemic requirements associated with Crombie’s six reasoning styles.

In summary, the preceding discussion has presented recent philosophical and psychological perspectives on science, as well as insights from cognitive history, to support the postulated model of scientific reasoning in this study. According to Kind and Osborne (2017), Crombie’s six reasoning styles offered a “coherent, domain-specific vision of both the entities [ontic, procedural, and epistemic] required and the different forms of reasoning that the sciences have developed” (p. 10). In contrast, they opined that each of the perspectives offered by psychologists and philosophers focus on only selected aspects of scientific reasoning. For instance, as explained in the preceding subsection, the

psychological account refers to scientific reasoning as a set of domain-general abilities, while the philosophical emphasis is on the norms and epistemic criteria of science.

Table 2.2

Styles of reasoning in European science, and examples of the propositional construct

Historical Period in Europe	Style of Reasoning	Example of Propositional Construct		
		Ontology	Methodology	Epistemology
Ancient Greece (2,500 years ago)	Deductive reasoning	Pythagoras's theorems	Symbolic notation	Deductive arguments
16th–17th century	Experimental-abstraction	Newton's laws of motion	Calculus	Role of mathematical prediction
17th–18th century	Hypothetical-modelling	Descartes's hydraulic model	Modelling	Limitations of modelling
19th century	Taxonomic	Linnaeus's system	Classification	Speciation rules and principles
Late 19th century	Historical-evolutionary	Darwinian theory	Systematic observation	Importance of substantiating claims
20th century	Probabilistic	Mendelian genetics	Statistical testing	Limitations of probabilistic claims

Note. Adapted from Crombie (1994); Kind and Osborne (2017).

2.5.2 Specific Reasoning Associated with Scientific Activities

This final subsection of the literature review defines and briefly describes 12 specific reasoning types commonly expected of pupils in primary education (Adey & Csapó, 2012; Korom, Nagy, et al., 2012). Also known as "thinking skills", "schemata", or "patterns", these reasoning types correspond closely to normal cognitive development, and draw from a rich background in philosophy, psychology, and logic (Adey & Csapó, 2012; Holyoak & Morrison, 2012). The emphasis during reasoning is the development and making of inferences using some initial information (Holyoak, 2012). As highlighted in section 2.5, the theoretical model of scientific reasoning posits that multiple reasoning types may be elicited depending on the reasoning context. Chapter 4 (subsection 4.3.3) will explain the importance of these reasoning types to the descriptor development of competencies

underlying the science practices. The following discussion begins by reviewing deductive and inductive reasoning, which were mentioned briefly earlier in subsections 2.2.1 and 2.2.2.

Deductive reasoning is a goal-orientated search that processes information (i.e., the premises) in logically valid ways to produce conclusions (Goswami, 2010). Conclusions drawn are true so long as the premises on which the basis of truth is claimed are true. In this light, deductive reasoning cannot go beyond the given information (Johnson-Laird, 2012). Leighton (2004) noted that for successful execution, users must possess the germane contextual, background, or content knowledge. Therefore, domain-specific knowledge has a critical impact on the way this reasoning is deployed. A well-studied application of deductive reasoning in science is the hypothetico-deductive approach used in the evaluation of evidence to support or disconfirm theories or hypotheses (Mahootian & Eastman, 2012). Some researchers contend that this approach is the basis of most scientific developments and should be emphasised during instruction to pupils (Lawson, 2000).

Inductive reasoning involves the uncovering and generalising of rules from a series of events to other, similar events (Dunbar & Klahr, 2012). Lipton (2004) described induction as a process of weighing available evidence and making a judgement call. The concluding inference may be disproven given new information. This important cognitive skill manifests early and is documented in many studies with primary-aged children (Goswami, 2010). These studies (see, e.g., Gelman & Markman, 1986) indicate that inductive reasoning is facilitated when the properties of the premise (e.g., car A has two doors and car B has four doors) and the expected inference are similar (e.g., a yet-to-be-observed car will likely have doors) and typical (e.g., the yet-to-be observed car will likely have two or four doors).

Closely related to inductive reasoning is *analogical reasoning* (Holyoak, 2012). Most accounts of this reasoning agree that the underlying mechanism is inductive in nature and consists of two components – the target and the source (Holyoak, 2012). The target is the concept or problem for resolution. The source is another known piece of knowledge or fact that is used to understand the target or to explain the target to others. To make the analogy, features of the target are corresponded to the source by uncovering similar or novel features,

or even reimagined features, in the target. One direct application of analogical reasoning is the construction of models, which are simplified and idealised representations of complex phenomena (Harrison & Treagust, 2000). Scientific models are used for the purposes of "summarising data, making predictions, justifying outcomes, and facilitating communication in science" (Erduran & Duschl, 2004, p. 112). At primary level, model-based reasoning is used extensively to help pupils "visualise" unseen entities (e.g., structure of a cell) or draw relationships among components of macro-level scientific processes, such as a water cycle (Schwarz et al., 2009).

Extrapolation has close associations with inductive and analogical reasoning. This reasoning refers to the inductive application of rules and patterns from events or processes to estimate the possibility of attaining similar outcomes in unknown or not yet investigated contexts (Adey & Csapó, 2012). Demonstration of this reasoning commonly occurs when primary pupils are required to extend graphical relationships beyond the measured data or predict the behaviour of unknown data based on existing patterns in given data (Gott & Duggan, 2003).

The development of *classification, seriation,* and *conservation reasoning* is linked to normal cognitive growth (Flavell et al., 2002). First described by Piaget, the acquisition of these types of reasoning is postulated to occur between the pre-operational and concrete-operational stages, when children are between seven and nine years of age (Inhelder & Piaget, 1958). At its simplest level, classification is the ability to group elements into two groups (with or without a common characteristic). More complicated classification schemes involve more characteristics or more values of a characteristic (e.g., with, without, partial). Seriation is the mental ability to arrange elements using an organisational principle, such as those related to temporal sequences, spatial position, or magnitude of quantity. According to Korom, Nagy, et al. (2012), understanding the meaning of words describing the principle (e.g., before, placed in front of, larger than) improves seriation reasoning.

Conservation refers to the understanding that physical quantities such as mass (e.g., of a round clay ball) or number (e.g., five clay balls arranged in a row) do not change despite

modifications to shape (e.g., flattened) or position (e.g., balls rearranged into two rows). The ability to conserve is required for understanding basic scientific phenomena (Adey & Csapó, 2012). For instance, when pupils realise that a flattened clay ball does not change its mass, they become cognisant of changes in physical states of matter and the possible impact such changes have on the properties of objects. Studies investigating conservation abilities indicated that young children progress developmentally from a "qualitative, 'guestimate'-minded approach" …(to) a quantitative, measurement-minded one" (Flavell et al., 1993, p. 136). Therefore, *qualitative reasoning*, which focuses on descriptive and non-numerical understanding of objects and their properties, develops early in children. Quantitative thinking follows after, and indicates that the child is ready for or is already operating in the Piagetian stage of formal-operation (Lawson, Karplus, & Adi, 1978).

Quantitative reasoning is the logical application of numerical understanding and procedures to draw conclusions or solve problems related to physical quantities (Adey & Csapó, 2012). Two forms of propositional logic related to quantitative reasoning have been well-studied in science education (Adey & Harlen, 1986; Lawson et al., 1978; Shayer, Adey, & Wylam, 1981). These are *proportional and probabilistic reasoning.* Proportional reasoning is knowing how changes to one variable relate to the changes in another variable (Boyer, Levine, & Huttenlocher, 2008). Scientific concepts such as speed and rates of change, which are usually introduced in the primary science curriculum, require the application of proportional understanding (Howe, Nunes, & Bryant, 2011; Lawson et al., 1978).

Probabilistic reasoning is the ability to predict a future outcome by weighing evidence from historical precedents, theory, or prior knowledge (Girotto & Gonzalez, 2008). Using simple experiments, Piaget and his collaborators were among the first researchers to investigate children's natural facility with chance and probability (Piaget & Inhelder, 1975). Their results suggested that children under 11- or 12 years of age are just beginning to appreciate simple notions of probability and are yet to be capable of probabilistic reasoning (Green, 1978; Piaget & Inhelder, 1975). Korom, Nagy, et al. (2012) recommend that primary-aged pupils

be made aware of the probabilistic nature of certain natural phenomena that is contingent on the presence or absence of conditions, such as seed germination, and changes of state.

Control-of-variable (COV) reasoning is the recognition and manipulation of key experimental variables to draw unambiguous outcomes or inferences (Chen & Klahr, 1999). This reasoning is critical for experimental set-up and the subsequent interpretation of results (Adey & Csapó, 2012). First investigated with school children in the 1950s (see, Inhelder & Piaget, 1958), the findings of follow-up studies have shown that COV reasoning develops with age (see, e.g., Bullock, Sodian, & Koerber, 2009; Lawson, 1978; Zimmerman, 2000). Other studies suggest that COV reasoning can be improved with instruction and practice (see, e.g., Shayer, 1999).

2.6 Conclusions

As described in this chapter, two contemporary and prominent accounts of scientific reasoning after the 1960s were the logical reasoning model of Piaget, and the inductive frameworks of Gagné's "process" and Bruner's "concept attainment" approaches. However, dissatisfaction with these early conceptions led to the formulation of three alternative and supplementary accounts of scientific reasoning. Unlike the Piagetian and inductive accounts, which focused on the "object" or reasoning mechanism, these newer accounts emphasised the "subject" of reasoning – expert content knowledge, knowledge of data-collection methods, and knowledge used to justify claims and conclusions. These three forms of knowledge are gathered from the practices of theorising, investigating, and evidence-evaluating. Together, the practices outline the activities of an inquiry cycle that scientists engage in while conducting research.

A synthesis of both the content of reasoning and the practices involved during reasoning led to a novel articulation of scientific reasoning. This model posits that a scientifically reasoning mind engages with disciplinary knowledge during claim-making and evidence evaluation using science practices. Evidence to support the three knowledge types, and the three science practices in the postulated model came from recent insights in the

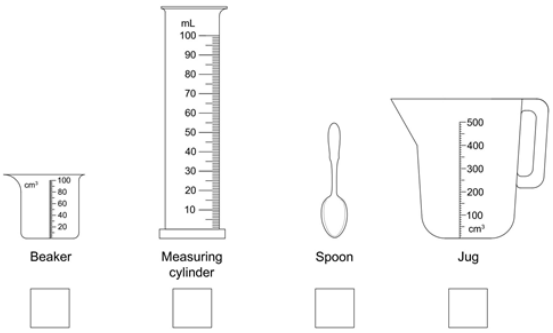
psychological, philosophical, and cognitive history literatures. This chapter has concluded with a discussion of the 12 distinctive reasoning types commonly expected of primary pupils in the learning and assessment of science. The next chapter will describe the validation design and approach taken in the construction and evaluation of a test based on this model of scientific reasoning.

Chapter 3 Methodology: Validation Framework

3.1 Introduction

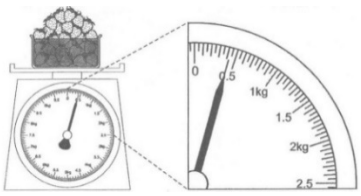
The two items shown in Figure 3.1 test the same scientific concept, which is identifying the equipment with the most appropriate scale.

Item 1: Which piece of equipment is most accurate for measuring 75 cm^3 of water?

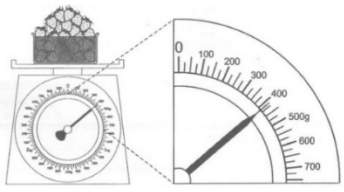


Item 2: A fruit seller used three instruments to find the mass of a box of fruit. Which one of the instruments gives the most accurate results?

(1)



(2)



(3)

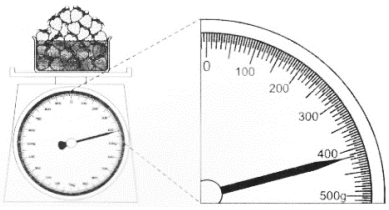


Figure 3.1. Items 1 and 2 are adaptations of items from the 2010 science key stage 2 and 2011 Year 6 science examinations.

Although both items have the same testing objective – to elicit evidence that our imagined 11- to-12-year-old pupils can *demonstrate* the concept, not merely *comprehend* it – the context of each item is different (Pollitt & Ahmed, 2001). However, are the two items measuring the scientific reasoning construct investigated in this study? More generally, can it be empirically determined whether items measure the same scientific concept? Part of this chapter addresses these questions. This chapter also discusses the validation framework that directs the research plan and design. The bulk of this framework focuses on how the different aspects of the research contribute to the validation of the PSRT.

3.2 Validation Framework

The validation framework encompasses the four phases (I to IV) of the research. Sequentially, these phases are: (phase I) an initial test design phase; (phase II) an expert panel review; (phase III) the pilot study; and (phase IV) the main study. The research plan in each of these phases identifies the methods for gathering selected sources of evidence.

This section begins by discussing the rationale for adopting the *Standards'* guidelines in this research, including the historical background which shapes modern validity notions and practices. Subsequently, the way in which these guidelines influenced the design of this research is delineated, including the decision to adopt a mixed-methods methodology to validate the inferences made from pupils' performances on the scientific reasoning instrument. The remaining sections of this chapter present the phase-specific aspects of the framework. These include the validation goals, the pertinent sources of validity evidence, specific research questions, descriptions of the methods for collecting validity evidence, and the techniques to be used for data analysis.

3.2.1 Validity and Validation

Modern validity theory posits that "all validity is construct validity" (Newton & Shaw, 2014, p. 110). This view of validity has its origins in Messick's (1989) well-known argument that construct validity subsumed all other traditional aspects of validity, such as criterion and

content validity. Current versions of the *Standards* advocate this contemporary notion of validity (Newton & Shaw, 2014).[6] The *Standards* is a document that specifies the criteria for evaluating tests, testing practices, and the effects of test use. Its guidelines have come to represent the consensus of the majority of practitioners involved in the fields of educational and psychological measurement (Zumbo, 2014). In particular, its recommendations about how tests and other forms of assessments should be developed, administered, and interpreted are closely adhered to and widely cited (Wiliam, 2014). This subsection discusses the shaping of contemporary notions of validity, as reflected in successive editions of the *Standards,* and the implications for test development in this research.

Historical Notions of Validity: When the systematisation of validity notions first occurred in the late 1940s, there were four conceptions of validity, of which construct validity was one (AERA & NCMUE, 1955). The remaining three types of validity were content, predictive, and concurrent. Each validity type had a different validation approach (see Table 3.1).

Table 3.1

Four types of validity and the approach to validation

Type of Validity	Approach to Validation (e.g., Vocabulary Test Scores)
Content	Demonstrate how well the content of the test samples the criterion; e.g., words in a vocabulary list
Predictive	Demonstrate how well the test scores correspond to a (future) criterion; e.g., academic success in university
Concurrent	Demonstrate how well the test scores correspond to another (present) criterion; e.g., use scores to compare with general abilities scores
Construct	Demonstrate the extent to which the test measures its intended psychological quality; e.g., reading ability

Note. Adapted from Newton and Shaw (2014).

Unlike the other types of validity, construct validity was created to evaluate the extent to which tests measure psychological constructs, such as personality and aptitude (Cronbach

[6] The latest edition of the *Standards,* published in 2014, is the fifth revision of this document.

& Meehl, 1955). The developers referred to a “construct” as a way of categorising “situations, persons, or responses”, and emphasised that it was a deliberate attempt on their part to group “experiences into general law-like statements” (Cronbach, 1971, p. 462). In this initial model of construct validity, nomological networks, or networks of related theories, accounted for behavioural manifestations of theoretical constructs. Therefore, both the test instrument and the theoretical aspects of the construct predicting consistent behaviours must be validated. In accordance with the prevailing philosophical view of science, the recommended methods of validation were positivistic, and modelled after scientific experiments, with proposed hypotheses and collection of confirmatory results (Cronbach, 1989).

Notably, three general techniques for establishing validation through test performance, based largely on correlational and factor analytics, were recommended (Cronbach & Meehl, 1955). These were: (1) differences in the performance of different groups on the same test; (2) intra-test (item analysis and reliability over time) and inter-test (with other tests) statistical relationships; and (3) consideration of test processes that might impact performance (e.g., scoring rubrics, test design). Citing the first edition of the *Standards*, which was published in 1954, the developers highlighted the need for comprehensive reporting of all data with inferences that support and account for the test performance (Cronbach & Meehl, 1955).

As was evident in two further editions of the *Standard*s, published in 1966 and 1974, the merging of two of the original four validity types – concurrent and predictive validity – into criterion validity shaped validity theory further (Newton, 2012b). The main reason for this merger was the acknowledgement that there was very little difference between the two, other than temporal (present and future) application of the criterion. The published guidelines in these two editions mirrored current thinking: that there were three types of validity, each with its own validation approaches. Interestingly, Newton and Shaw (2014) noted that the notion of construct validation in these two editions was increasingly used to organise guidelines on related measurement concepts and issues, such as sampling processes, speed effects on test performance, and measures of internal consistency.

During this period key writings emerged that laid the foundation for a subsequent shift from a “fragmented, trinitarian conception of validity … to a synoptic unitarian one” (Newton & Shaw, 2014, p. 111). For instance, Lennon (1956) clarified that content validation must consider the responses of test-takers to the questions, not just the content of the questions (Sireci, 2015).

Messick’s Unified Theory of Validity: As attested to in the fourth (1985) and fifth (1999) editions of the *Standards,* the shift to Messick’s influential unified view of validity became established in the last two decades of the 20th century (Newton, 2012b). Messick’s (1980, 1989) assertion that construct validity subsumed both content and criterion-related validities echoed Loevinger’s (1957) position more than 20 years earlier. Like Loevinger, Messick highlighted that the traditional understanding of content validity, which predominantly centred on deploying expert panels to judge the extent to which the testing content sampled the entire content in the achievement or proficiency spectrum, was inadequate as a yardstick to judge the validity of the tests (Messick, 1975). Drawing on Loevinger’s arguments – that it was fallacious to equate rigorous sampling in tests (high content validity) with equally rigorous measurement of the underlying attribute in the tests – Messick instead argued for a conceptualisation of validity that focused on inferences from performances in the tests.

Messick (1989) also criticised the conventional understanding of criterion validity promulgated in earlier versions of the *Standards*, which was “demonstrated by comparing the test scores with one or more external variables considered to provide a direct measure of the characteristic or behaviour in question” (APA, AERA, & NCME, 1966, p.13). Noting that this understanding translated into examinations of statistical associations between test scores (predictor) and the scores of the same test-takers on selected external variables (criterion), he questioned the basis of the conclusions drawn from such empirical associations. He argued that the onus was on the test developers to provide a rational basis for linking the construct of the predictor to the construct of the criterion, which justifies the use of the test results to predict the criterion scores. Invoking nomological networks as an

approach for conceptualising the logical relationships between related constructs, Messick argued that criterion validity is a special variant of construct validation.

In sum, Messick found the traditional definitions of content validity and criterion validity lacking (Newton & Shaw, 2014). In his view, construct validity is a superordinate validity concept that organises all other traditional forms of validity. Validation is informed by theory, conducted scientifically, and is an ongoing process of justifying, collecting, and evaluating validity evidence from many sources, including analysing plausible rival hypotheses to account for performances on the tests. Developing the idea that there is no singular source of evidence sufficient to support a validity claim, Messick (1989, 1993) identified six criteria, or sources of evidence, for ascertaining construct validity. These are: the test content; the theoretical basis of the test, which directs the expected performance; the structural relationships among parts of the test; the relationship between test performance and performance on some stipulated criterion; generalisability of interpretations across sample populations and contexts; and evaluation of the intended or unintended consequences.

Significantly, the methods used to collect validity evidence from these sources extended from those traditionally implemented for each form of validity during the trinitarian era (Messick, 1995a, 1995b). Therefore, besides the statistical techniques formalised in the 1950s to evaluate construct validity and the usual deployment of expert panels to judge content validity, Messick recommended that other methods be used concurrently to collect validity evidence germane to the investigation. For instance, test developers should consider delineating the conceptual rationale and content domain of testing for new tests clearly. They should also produce evidence to determine that the new tests do not bias the performance of subgroups (e.g., gender). Taken together, the inferences from the evidence provide accounts of the strengths or weaknesses of the tests for score interpretation and score use.

Messick's view – that validity is a judgement made after considering evidence from relevant sources, using multiple methods to collect this evidence – gained acceptance during the

1990s and heavily influenced current versions of the *Standards* (Kane, 2013). The latest edition of the *Standards* continues to espouse the unitary view of validity (AERA, APA, & NCME, 2014). The sections that follow discuss the *Standards*' validity definition and validation guidelines and highlight implications for the current study.

Validity and Validation in the Standards: The *Standards* defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). This definition advocates that both theory and empirical evidence are necessary to validate inferences about the performance of test-takers. This stance is consistent with Messick's unitary view of validity. The *Standards* advocates five complementary and distinct sources of validity evidence, derived from Messick's six criteria[7] (Newton & Shaw, 2014): evidence based on test content; response processes; internal structure; relations to other variables; and consequences of testing. The collection and analysis of validity evidence from these sources involve multiple methods.

Overall, there have been significant disagreements over the interpretations of the *Standards*' definition of validity and some validation guidelines. Critics have charged that its definition and guidelines lack "precision, consistency, and clarity" (Newton, 2012a, p. 1) and provide little guidance on key issues, such as whether validity refers to test score interpretation, test use, or both (Sireci, 2015). Some commentators have noted that these issues largely centre on the lexical and theoretical meanings of validity (Newton & Baird, 2016). The lexical debates concern the semantic analysis of validity and its associated concepts. As an illustration, Markus (2016) noted that various validity definitions have expressed similar arguments about validity, and advised that future work contributes to move the validity field and its practices forwards instead of, "defending specific definitions or vocabularies" (p. 265). Conceptual differences form the second source of disagreement in the literature noted

[7] Although Messick's evidence related to generalisability of inferences is not highlighted as a distinct evidence source in the *Standards'* formulation of sources, evidence of test invariance across sub-populations is required in the source of internal structure.

by Newton and Baird. Like many such issues (e.g., whether ethical considerations form part of the understanding about validity), there is not yet any resolution.

Notwithstanding the ongoing debates, some proponents have commended the *Standard*s' guidelines and recommendations, arguing that these are contemporary to and complement existing practices (Chan, 2014; Diaz-bilello, Patelis, Marion, Hall, & Betebenner, 2014). For instance, Chan (2014) reviewed seven standards and guidelines [8] relevant to the development and evaluation of instruments and noted that only three – including the *Standards* – promoted modern validity notions. Others have added that there is enough agreement to move assessment practices forwards. For instance, Wiliam (2014) stated that there is "a reasonable consensus that validity is thought of as a property of inferences based on test outcomes rather than as a property of tests" (p. 29). Also, Bennett (2012) noted that the unease over the generality of the *Standards'* recommendations may be misplaced, as it "is desirable for a given formulation to have applicability to a wide range of situations, including future ones that cannot be anticipated" (p. 30). Wiliam (2014) reiterated this stance, noting that, while there is no explicit guidance for classroom assessment validation practices in the *Standards*, nevertheless the "contexts of application are so diverse that it would be difficult to imagine anything useful being said" (p. 29). In conclusion, the preceding discussion suggests that the validity definition and recommendations in the *Standards* are useful to guide instrument development and validation in educational research.

3.2.2 Implications of the *Standards'* Recommendations for the Study

The validation framework in this study identifies three of the five *Standards'* sources of evidence to collect validity evidence for evaluating the properties of the PSRT. The reasons for not considering the remaining two sources are explained below. On the whole, the evidence presented by the three selected sources provides a composite argument for the proposed inferences of assessment results. This subsection discusses the five *Standards*-

[8] These standards and guidelines apply to major professional associations and organisations in education, psychology, business, and health.

recommended sources with respect to the construction and validation of a classroom-based scientific reasoning test. It concludes that a mixed-methods research methodology is the most suitable design to organise the multiple methods for the collection and analysis of data.

Sources of Validity Evidence

(1) Evidence based on test content

Test content evidence pertains to the association between the content of the test and the underlying construct the test is designed to measure (AERA, APA, & NCME, 2014). Judging the extent to which the assessments tasks are representative samples of the larger domain of performance, from the perspectives of the test designer (content representativeness) and the users of the test results (content relevance), is a common method to collect validity evidence (Nitko & Brookhart, 2007). This study has adopted a more explicit and comprehensive definition of content validity for classroom assessments: “the degree to which a test measures the content domain it purports to measure” (Sireci, 1998b, p. 299). More helpfully, Sireci (1998a) ascribed content validity to four features of test quality. The next section provides a fuller discussion about these features and their role in the validation framework used in the current research.

(2) Evidence based on response processes

The *Standards* states that evidence based on the response process concerns “the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA, APA, & NCME, 2014, p. 12). Elsewhere, Zumbo and Hubley (2017) defined response processes as the “mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task” (p. 2). In this thesis, qualitative analyses of the items and psychometric investigations of the test instrument were the methods used to obtain evidence of response processes. Psychometric analysis involves an examination of the correspondence between the actual scores or performances (estimated difficulties) and the intended cognitive or physical competencies in test construct (theoretical difficulties).

Besides qualitative and psychometric analytical approaches, data from cognitive interviews was also collected as part of the evidence for response processes. However, the analysis of the interview data was not completed in time. Hence, this method was left out of the validation framework. This limitation will be discussed in Chapter 9, section 9.4.

(3) Evidence based on internal structure

Internal structure evidence provides empirical and qualitative information about the relationship among the items or parts of the test such as its dimensionality and reliability (AERA, APA, & NCME, 2014; Rios & Wells, 2014). Checking that these data corroborate the substantive or conceptual framework of the test's underlying construct provides important validation evidence. This study drew on psychometric methods from Rasch modelling to collect the necessary evidence.

(4) Evidence based on relations to other variables

The association between external variables and the hypothesised construct of the test provides a valuable source of validity evidence (AERA, APA, & NCME, 2014). For example, evidence to support inferences that the construct measured in the PSRT is not general intelligence could provide discriminant evidence of validity. Suitable validated instruments for this purpose include the well-established Raven's Standard Progressive Matrices-Plus (SPM+) and the Mill Hill Vocabulary (MHV) Test (Raven, 2008). Evaluating whether the data from these tests correlate moderately with performance on the PSRT will support the distinction between scientific reasoning and general reasoning. However, though evidence from these instruments was collected, the data had not been analysed fully. Therefore, the validation framework of this thesis excluded this source. The implications of excluding evidence from the two psychological tests will be discussed in Chapter 9, sections 9.4 and 9.5. These sections elaborate the limitations and possibilities for post-thesis research.

(5) Evidence based on consequences of testing

The *Standards* recommends the collection of information about the impact of test results on learning, instruction, and social equity. For instance, checks on how the test results influenced the pedagogical practices of teachers or implementation of policies for sub-populations of under-performing pupils are some ways to collect consequential evidence. The science education literature acknowledges both pernicious and constructive impacts of tests on learning and teaching practices (Roberts & Gott, 2004). Documented effects include elicitation of "teaching to the test" behaviours (Darling-Hammond & Rustique-Forrester, 2005), narrowing of the curriculum (Au, 2007), enhanced motivation, and clearer, more focused instructional targets (Alderson & Wall, 1993; Madaus, 1988). The effects of testing, also known as "washback", are so entrenched that some researchers have instead called for preventive efforts to be redirected to the study of how test design promotes positive washback (Davies, 1985). However, given the research goals and design of this validation study, studying the effects of the test is beyond the scope of the thesis.

In sum, validating the inferences from the PSRT scores involved the application of a variety of methods to collect the required quantitative and qualitative validity data. Accordingly, this study implemented a mixed-methods methodology[9] to organise the methods for gathering validity evidence to address the research question and research aims. The validation framework, to be presented in the next subsection of this chapter, details the collection of evidence from three of the five *Standards*-advocated sources across the research phases.

Mixed-Methods Methodology: According to Johnson, Onwuegbuzie, and Turner (2007), adopting a mixed-methods research methodology is for the "broad purposes of breadth and depth of understanding and corroboration" (p. 123). Their rationale highlights the "fundamental principle of mixed research" (Johnson & Onwuegbuzie, 2004, p. 14), which is to draw on the advantages of data collection, analysis, and inference methods conferred by

[9] This study adopted Nudzor's (2009) distinction between "methodologies" and "methods". The former refers to "strategies, plans of action, processes or designs lying behind the choice and use of particular methods" (p. 115). In contrast, methods are "the techniques or procedures used to gather and analyse data" (p. 115).

quantitative and qualitative approaches. The collective advantages ensure there is maximal extraction of information for stronger inferences. Notably, the introduction of *multiple* methods in the assessment validation literature in the mid-1950s incepted the idea of merging a few techniques for superior overall outcomes in social science research (Johnson et al., 2007). References to “multiple operationalism” and “operational delineation” in publications by Campbell and collaborators are widely credited for establishing the use of various methods to distinguish between the desired trait variance and the method variance (Campbell, 1960; Campbell & Fiske, 1959; Webb, Campbell, Schwartz, & Sechrest, 2000).

Later, Denzin (1970) advanced the concept of “triangulation”, first proposed by Webb, Campbell, Schwartz, and Sechrest (1966), by explicating how triangulation can occur in four ways. The first is data triangulation, which refers to triangulating data from a few data sources. The second and third are investigator triangulation (using multiple observers in the same research) and theoretical triangulation (approaching data from multiple theoretical perspectives for interpretation and analysis). The fourth and final way is methodological triangulation, including both within-methods triangulation and between-methods triangulation. The former involves multiple methods from only one tradition, while the latter refers to the use of multiple methods from both traditions. Denzin (1970) recommended the adoption of between-methods triangulation to overcome methodological threats, such as biases peculiar to any one method, by accruing strengths from each.

From the utility point of view, Teddlie and Tashakkori (2003) highlighted that mixed-methods research can draw on the advantages afforded by both numerical and narrative research methods for stronger inferences. For instance, mixed-methods research can assess the statistical associations between variables of interest and provide contextualised, humanistic accounts of the underlying causal mechanisms for the associations. Moreover, conflicting findings and possible theoretical or methodological issues can be uncovered. In effect, mixed-methods methodologies accommodate a wider variety of alternative perspectives than the possibilities afforded by a solely qualitative or quantitative approach.

Criticisms about mixed-methods approaches have centred on the lack of consensus about three fundamental issues: philosophical and epistemological stances; terminology related to research designs and processes; and the appropriate balance and use of qualitative and quantitative approaches (Teddlie & Tashakkori, 2012). Objections on grounds of dualistic and mutually incompatible epistemological and philosophical doctrines between the two research traditions (e.g., objectivity versus interpretivism) have formed a major argument against the use of mixed-methods research (Howe, 1988). Also, the emergent nature of the field has resulted in a wide and confusing plethora of terms, definitions, and practices, leading to uncertainties about the effectiveness of the methodology (Bryman, 2007). Finally, there are concerns and disagreements about how different methods from each of the two traditional methodologies should be combined (Bryman, 2006).

Addressing the potential criticisms of mixed-methods approaches, Collins, Onwuegbuzie, and Sutton (2006) recommended four thematic rationales for combining qualitative and quantitative approaches in research. Instrument fidelity is one of the rationales, which means ensuring that inferences made are valid, reliable, and appropriate for use. Collecting a variety of data sources and analyses from insiders' views and researchers' theoretical perspectives is particularly important when the developed instrument is to be used for the measurement of complex, multi-faceted psychological constructs (Onwuegbuzie, Bustamante, & Nelson, 2010).

Methodological Design in Study: Consequently, this research collected and analysed a wide range of data from both methodological orientations, such as past research findings, survey ratings, test performance results and written feedback from an expert panel. Methods to collect these data included administering survey forms and test instruments, surveying current literature, and conducting psychometric analyses. The criterion for method selection and, if required, combining evidence from across methods, was the potential of the gathered data to address the validation goals effectively and efficiently (Johnson & Onwuegbuzie, 2004). In effect, the goals determined the sequence, the

methodological orientation of the method, and the weight of the evidence from the method for making inferences (Morse, 2010).

To illustrate, as shown in Table 3.2, the methodological focus of phase II, denoted by the shortened and capitalised notation "QUAN + QUAL", indicated that quantitative evidence was first analysed, but both types of evidence carried equal priority in the analysis. As will be explained in the next subsection, phase II involved the collection and analysis of test content evidence from the expert panel about the draft test. Drawing on evidence from mixed methods in this phase was necessary to produce stronger inferences about the validity of the newly constructed test.

In contrast to phase II, phase III piloted the instrument with a sample of pupils, and the scores underwent psychometric analysis. As shown in the table, the primary evidence in this phase was thus quantitative, and the small case letters "qual" signalled that evidence from qualitative methods supplemented inferences made by the dominant orientation. See Table 3.2 for a summary of the design in the four phases.

Table 3.2

Summary of mixed-method design focus in research

Mixed-Method Focus in Phase	Validation Goal(s)
Phase I QUAL ↓	Collect evidence based on test content for the development of the scientific reasoning construct and the identification of typical performance.
Phase II QUAN + QUAL ↓ Phase III QUAN + qual ↓	Collect evidence based on test content, response processes, and internal structure to develop the draft instrument (Draft PSRT).
Phase IV QUAN + QUAL	Collect evidence based on test content, response processes, internal structure to assess the construct validity of the finalised instrument (PSRT).

Note. Phases II and III were concurrent. Notation adapted from Johnson and Onwuegbuzie (2004) and Morse (1991). "QUAL" refers to qualitative, "QUAN" to quantitative, "→" stands for sequential, "+" stands for concurrent. The use of capital letters refers to the priority or weight of the methodological orientation. For instance, "QUAL + QUAN" would denote both orientations having equal priority in the research phase.

To strengthen inferences from the mixed-methods analyses, the study also considered an emerging group of techniques that demonstrate potential for greater data integration in instrument development and validation (Onwuegbuzie et al., 2010; Onwuegbuzie & Combs, 2010). Known as “crossover analysis”, these techniques involve the use of one or more analytical methods (e.g., effect sizes) commonly identified with one orientation (e.g., quantitative) to analyse the data obtained from another orientation (e.g., qualitative research). Three research phases of this study drew on five of the ten crossover analytic strategies (see Table 3.3 for descriptions) highlighted by Onwuegbuzie and Combs (2010).

Table 3.3

Cross-over analytic strategies in research

Crossover	Phase(s)	Description
Integrated data reduction	II	Thematic analysis of quantitative data, collected from expert panel, about Draft PSRT
Data comparison and data integration	II and III	Comparing and integrating findings of data from phases II and III to review Draft PSRT
Warranted assertion analysis	I, II, III and IV	Reviewing qualitative and quantitative data to yield meta-inferences or overall conclusions about construct validity of PSRT

The remaining sections of this chapter discuss the phase-specific aspects of the validation framework, such as the specific sources of validity evidence and the methods for collecting the evidence. Table 3.4 presents a summary of the mixed-method validation framework for instrument development and validation in this study.

Table 3.4

Summary of mixed-method validation framework in the four-phased study

Research Phase	Method	Analytic Technique
Phase I: Initial test design	• Review of substantive theories and curricular documents (conceptual, measurement, and test design)	• Nomological network analyses
Phase II: Expert panel review Phase III: Pilot study	• Review by expert panel • Field-test Draft PSRT	• Survey analyses • Rasch modelling • Content analyses of participants' responses
Phase IV: Main study	• Review by two-member expert panel • Administer PSRT	• Rasch modelling • Content analyses of participants' responses

3.2.3 Validation Framework in Phases

Phase I

Validation Goal*:* The goal of this phase is to initiate and gather content validity evidence from the development and assembly of items into the Draft PSRT.

Methods to Collect Validity Evidence: Sireci's (1998a, 1998b) content validity framework has four test features: domain definition; domain representation; domain relevance; and the appropriateness of the test-development process. Jointly, these features guide the gathering and evaluation of evidence in this study.

A *domain definition* refers to the observable indicators or behaviours indicative of a personality trait, defined by a theoretical and psychological construct, which the items in a test instrument elicit (Yang & Embretson, 2007). Researchers such as Gorin (2007) recommend the use of detailed construct definitions, including descriptions of cognitive processing and theoretical relationships among trait dimensions (content areas or abilities). Accordingly, the literature review in the previous chapter described the substantive theories

and conceptual basis of the scientific reasoning construct investigated in this study, while Chapter 4 focuses on the conceptualisation process.

This subsection identifies and explains why nomological network analysis was the qualitative method chosen to gather validity evidence in the first research phase of the study. A nomological network is a set of principles, rules, properties, observed behaviours, or theoretical ideas associated with a broader theory (Cronbach & Meehl, 1955). Besides the substantive theories underlying the proposed scientific reasoning construct, the network in this research included measurement concepts, test theories, and curriculum emphases. Analysing the relationships in the network allowed the empirical measurement of hypothetical notions and the interpretation of measured outcomes (Masterson & Rainer, 2009). For instance, one important analytical strategy is drawing up a table of specifications (TOS) to specify the scoring criteria for mark awarding (Nitko & Brookhart, 2007). The TOS ensures a *balance* of items of the intended level of intellectual challenge from an *adequate* selection of knowledge and practices postulated to form the scientific reasoning construct (Sireci & Faulkner-Bond, 2014).

Besides characterising items according to the dimensions and components of the construct, the TOS is used to gather evidence for the next two test features of content validity. *Domain representation*, which is the second feature, refers to the extent to which the characterisation of items in the PSRT aligns to what is reflected in the TOS (Sireci, 1998b). The third feature of content validity is *domain relevance*, which addresses the degree to which each item is relevant to the targeted domain. Specific details about how the TOS and other analytical strategies (e.g., drawing up the study's test framework*)* collected validity evidence are delineated in the next chapter.

The fourth and last type of content validity evidence in Sireci's (1998b) framework is *appropriateness of the test-development process*. This type of evidence comes from methods put in place to corroborate or triangulate evidence that the test measures the construct as intended. Sireci and Faulkner-Bond (2014) described checks on item

properties and adoption of appropriate test features as the two broad categories of these methods. Item property checks implemented included: item reviews by content and assessment experts in phases II and IV (Downing & Haladyna, 1997); pilot testing of items to gauge their suitability and statistical properties in phase III (Linn, 2006); and the use of Rasch modelling techniques, such as differential item functioning, to check whether some items are inordinately more difficult for some subgroups of the population than others in phases III and IV (Ercikan et al., 2010; Haladyna, Downing, & Rodriguez, 2002). Besides item property checks to exclude construct-irrelevant biases, possessing appropriate test features also furnished validity evidence. These features included suitable item formats for eliciting reasoning, and appropriate item organisation in the test to ensure no overlapping content. See Table 3.5 for the four aspects of test content validity and the methods to collect the necessary evidence.

Table 3.5

Test content validity and methods to collect validity evidence in phase I

Mixed-Method Design Focus: QUAL	
Validity Source: Test Content (Sireci, 1998a, 1998b)	Method to Collect Validity Evidence
	Nomological network analyses of:
(i) Domain definition (ii) Domain representation of construct and items (iii) Domain relevance	(a) Substantive theories and curricular emphases • to conceptualise the scientific reasoning construct • to identify and describe the behaviours that underlie the construct (b) Test design and measurement theories • to develop test framework, which outlines the item format, target distribution of marks, and item exemplars • to develop TOS • to develop items and assemble initial draft instrument (Draft PSRT) • to select measurement model
(iv) Appropriateness of the test-development processes	(c) To adopt appropriate item features in the development of items and the assembly of Draft PSRT

Phase II

Validation Goal: The goal of this phase is to obtain content validity evidence and preliminary judgement from an expert panel review.

Methods to Collect Validity Evidence: The four test features of content validity, as described earlier, are used as a framework to capture and analyse quantitative and qualitative evidence from the expert panel. In this study, the panel gauged the degree of correspondence between the construct and the curriculum to the testing content. Engaging experts working in groups or individually to evaluate the degree of alignment of tests is a common practice in content validation exercises (Martone & Sireci, 2009; Sireci & Faulkner-Bond, 2014). Table 3.6 shows the validation plan for this phase.

Table 3.6

Validation plan in phase II: expert panel review

Mixed-Methods Design Focus: QUANT+ QUAL			
Validation Source	Specific Research Question	Method of Collecting Validity Evidence	Data Analysis and Crossover Strategy (Denoted with *)
1. Test content about: (i) the scientific reasoning construct in the domain of science (ii) the items in the Draft PSRT	1. What are the experts' judgement about: (a) domain definition, representativeness and relevance of the construct in the domain of science? (b) relevance of intended types of thinking skills and processes required by the construct? (c) content representativeness and relevance of the items, accuracy of the science tested, item quality? (d) relevance and match of intended types of thinking skills and processes to the curriculum?	1. Review of construct in test framework and items in Draft PSRT by 18 experts using: • paper-and-pencil questionnaire (proforma) • recorded interviews, either face-to-face or online	• Descriptive statistics to summarise quantitative review data about construct and items •*[Integrated data reduction] Thematic analysis of quantitative data • Thematic analysis of qualitative data from recorded interviews

Phase III

Validation Goal: The goal is to obtain test content and internal structure validity evidence.

Methods to Collect Validity Evidence: As part of the test-development process, the Draft PSRT underwent field-testing with Year 7 pupils to check the quality, suitability, and properties of the test items, as well as the appropriateness of the administration procedures and scoring. The performance data of the pilot items were analysed psychometrically using Rasch modelling. This psychometric approach will be elaborated briefly in the discussion on the next phase. See Table 3.7 for the validation plan.

Table 3.7

Validation plan in phase III: pilot study

Mixed-Methods Design Focus: QUANT + qual			
Validation Source	Specific Research Question	Method of Collecting Validity Evidence	Data Analysis and Crossover strategy (Denoted with *)
1. Test content and response process	How does the performance of the pupils inform the design of the questions in the Draft PSRT?	Field-test of Draft PSRT with 138 Year 7 pupils in four selected secondary schools	• Conduct fit analysis of test, item and person using Rasch item response theory (IRT) • Observations of key item response patterns to identify specific problems (e.g., misconceptions) • *[Data comparison] Compare data (qualitative and quantitative) from phase II and quantitative data from phase III • *[Data integration] Integrate findings of data (qualitative and quantitative) from phase II and data (qualitative and quantitative) from phase III to revise items in PSRT

Phase IV

Validation Goal: The goal of this phase is to obtain validity evidence from the three validity sources – test content, internal structure, and response processes – through the administration of the final instrument (PSRT).

Methods to Collect Validity Evidence: Item analysis using Rasch modelling investigated the dimensionality and invariance of the scientific reasoning construct measured in the PSRT. Other Rasch tests conducted to gain internal structure evidence included examining the threshold ordering of polytomous items, fit of items, and differential item functioning (DIF). DIF statistics, which show whether the items function significantly differently for subgroups of test-takers (e.g., different genders and schools), provide important evidence for varied sources of validity (Haladyna & Rodriguez, 2013).

Response process-related evidence came from pupils' responses on the PSRT. Examination of the answers to the items assessed whether the intended processes being used by respondents are those demanded by the items. See Table 3.8 for the validation plan in phase IV.

Table 3.8

Validation plan in phase IV: main study

Mixed-Methods Design Focus: QUANT + QUAL			
Validation Source	Specific Research Question	Method of collecting Validity Evidence	Data Analysis and Crossover Strategy (Denoted with *)
Test content, internal structure, and response process	1(i) Is the PSRT measuring a unidimensional construct?	Administer PSRT to 431 Singapore pupils	• 1(i) and 1(ii) Conduct fit analysis using Rasch modelling

(continued)

Table 3.8 (continued)

Mixed-Methods Design Focus: QUANT + qual			
Validation Source	Specific Research Question	Method of collecting Validity Evidence	Data Analysis and Crossover Strategy (Denoted with *)
	Specifically: • What are the item characteristics in the PSRT? • How does the PSRT function regarding the interaction of the person-gender, person-school, and person-booklet in each item? 1(ii) Is the construct measured by the PSRT the postulated theory of scientific reasoning? Specifically: • Do the item locations and threshold estimates conform to expectation?		• 1(iii) Qualitative analysis of pupils' performance in the PSRT items • *[Warranted assertion] Draw meta-inferences or broader inferences from qualitative and quantitative methods

3.3 Validation and Ethical Considerations

The preceding discussion provided a detailed blueprint for the systematic collection and analysis of validity evidence from three of the five *Standards*-recommended sources. Other than these sources, which had a data-centric focus, final considerations to promote the validity of the inferences about the quality of the PSRT were the rights and fair treatment of the participating pupils. The following discussion on the research ethics and practices adopted in the conduct of this study concludes this chapter.

Guiding all ethical practices was the principle of informed consent (AERA, APA, & NCME, 2014). This principle broadly encompassed the initial procedures: providing information on the requirements of the research, as well as the necessary documentation, to potential participants and, by extension, other stakeholders. To this end, a communication package was prepared consisting of pamphlets, letters, and consent forms for pupils, parents, teachers, and schools. These documents spelled out the aims of the research and the proposed testing protocols. Besides the package, other pre-test measures were enacted to ensure pupils understood their rights and entitlements. These included briefing all pupils prior to the test to reiterate the objectives of the test and the intended use of their test scores. They were also reassured that their performance was confidential and would have no impact on their school work. Consideration of pupils' needs extended to administrative practices during and after the test. For instance, pupils could request withdrawal from the test at any juncture, even when halfway through sitting the test. Post-test practices included the scoring and reporting of results. Among these practices, there were two major concerns, which the study paid close attention to while carrying out the four research phases.

The first concern was ensuring pupils and stakeholders understood there would be no release of individual results. Instead, schools would receive reports outlining broad findings, and general observations of their pupils' performance. The usage of these reports would be at the schools' discretion, which could include sharing them with interested parents. Communicating the stance that pupils' performance would not be released took place in meetings with the schools, as well as through the letters sent to parents and guardians.

The second ethical concern was the requirement to put in place a systematic and rigorous anonymisation strategy to keep the identity of the participating schools and pupils confidential. Measures adopted included assigning unique numerical identifiers to every pupil in the data collection and analysis stages and encrypting and securely storing the data. Permission for the research plan was given by the Central University Research Ethical Committee (CUREC) before data collection and recruitment of expert reviewers from

Singapore. See Appendix 4 for the part of the CUREC form which discusses other ethical considerations.

In conclusion, this chapter has addressed the basis and methods for the validation of a newly constructed test of scientific reasoning for primary school children. Using established understanding and recommended approaches by practitioners in educational and psychological testing, these validation practices were applied in this research to ensure strong inferences could be made to support interpretation of test scores from the newly constructed test. The next chapter reports the activities and results of the study's first phase.

Chapter 4 Initial Test Design

4.1 Introduction

As explained in Chapter 1, this research constructed and validated a scientific reasoning test for primary pupils on the premise that there were gaps in current understandings of the proficiencies involved. The literature review in Chapter 2 presented a new conceptualisation of these proficiencies drawn from various research fields and theoretical frameworks focused on the cognitive development of children. Chapter 3 then discussed the validation methods for evaluating a test instrument created based on the novel scientific reasoning construct proposed. The groundwork described in these three chapters led to the first phase of the research which developed the initial draft of the PSRT, to be described in this chapter.

Broadly, this chapter discusses the theoretical and practical considerations of item and test design during the drafting process. These include decisions and activities related to the development of behavioural descriptors corresponding to the knowledge types and practices in the postulated construct. Also highlighted in the chapter are relevant perspectives from learning theories and measurement guidelines and techniques established to inform item ideation and design. For instance, one well-established item-writing guideline is that test items should be directly meaningful to pupils' everyday experiences or education (Nitko & Brookhart, 2007).

One way of conforming to this guideline in this research was to incorporate realistic and familiar test contexts within items to reduce test anxiety and motivation, thereby enhancing test validity (Haladyna & Downing, 2004; Haladyna & Rodriguez, 2013). Accordingly, real-world contexts used in the design of many new items in the PSRT came from information sources such as newspapers. In the news article example shown in Figure 4.1, the notion of suitable breeding conditions provided the background for a test item assessing pupils' understandings of the epistemic commitments of scientific experimentation.



Looking ahead, Mr Teo added that a key performance indicator would be the ability to successfully get the pandas to breed.

However, he added that such processes were "not controllable". "As you can see, we are facing challenges when it comes to procreation," he said, laughing.

Item: Breeding a male panda with a female panda to get a baby is not easy.

The questions below are about conditions affecting the breeding of pandas. Which of the questions can be answered with a scientific experiment?

(a) Will the amount of food each animal eats affect breeding? Yes/No
(b) Will playing music to the animal affect breeding? Yes/No
(c) Will the feelings of the male panda for the female panda affect breeding? Yes/No

Figure 4.1. An item in the Draft PSRT crafted using ideas gathered from a news event.

The chapter concludes with the test framework section, which discusses considerations for assembling the newly created items into the Draft PSRT. These considerations include the scheme used to characterise item features, as well as the guidelines for organising items with specific features within the test.

4.2 Validation in Initial Test Design

This phase initiated and collected test content evidence from the development and assembly of test elements for an initial draft of the PSRT. The validation framework articulated in the previous chapter identified four test features comprising content validity, proposed by Sireci (1998a, 1998b). As shown in Table 4.1, activities in this phase collected and analysed evidence about these four features: domain definition; domain representation; domain relevance; and appropriateness of the test-development process. To this end, there is analysis of relevant nomological networks to explain, measure, and link the construct of scientific reasoning to other relevant theories (Cronbach & Meehl, 1955). Insights from

identified substantive, measurement, and test theories then guided the development of the construct and test materials. For example, as will be elaborated in a later section, the design parameters of the PSRT (e.g., item format and cognitive demand) drew primarily from cognitive learning theories. The following two sections discuss the analytic process and the activities to create the test and items.

Table 4.1

Relationship between content validity and nomological network analytical activities

Test Feature in Content Validity (Sireci, 1998a, 1998b)	Explanation	Nomological Network Analysis	Activity in Phase I
Domain definition	Descriptions of consistent behaviours attributed to a distinct trait defined by a theoretical construct	Review substantive theories and curricular documents	Develop construct and behavioural descriptors
Domain representation	Alignment between test content and the intended assessment objectives	Review substantive and measurement theories, and test design practices	Identify criteria in a test framework, *and* assemble the Draft PSRT using TOS
Domain relevance	Alignment between test content and the construct of interest		
Appropriateness of test-development process	Operational procedures to ensure test content measures the construct of interest. These procedures consist of (1) checks on item properties[10] and (2) use of appropriate item features	Review substantive and measurement theories, and test design practices	Develop test content such as items, instructions to administer the test, and item assembly procedures across five booklets

[10] This validity evidence came from activities in a few phases: for instance, an expert panel review (phase II) and a psychometric analysis of pupils' performance in the pilot study (phase III).

4.3 Development of Definitions and Behavioural Descriptors

The first step in the development of an initial draft of the PSRT was to define and develop descriptions of the knowledge types and science practices espoused in the scientific reasoning construct. Development of these descriptors followed a combination of four approaches: theory-based, empirically-driven, policy-orientated and expert-devised (Baird, Andrich, Hopfenbeck, & Stobart, 2017). In the first, theoretical perspectives incorporated in the definitions and descriptors came from relevant fields, such as science education and cognitive science research. For instance, research findings informed current understanding of age-appropriate and instructionally relevant procedural and epistemic knowledge for primary pupils (see, e.g., Roberts & Gott, 2004; Sandoval & Çam, 2011).

Empirical evidence from implemented practices and experience also provided guidance. These included existing frameworks that demonstrated the assessment of scientific reasoning (see, e.g., Adey & Csapó, 2012; Lawson, 1985; Shayer et al., 1981) as well as insights from historical and current curricular initiatives and reforms.[11] In the third approach, development also referred to the knowledge and competencies expected of primary pupils and outlined in the science curriculum documents of Singapore and England.[12] Finally, I produced the initial draft of the definitions and descriptions, drawing on 12 years of teaching experience in secondary education followed by a decade of service as a science assessment specialist on Singapore's examination board, designing test items for primary science national assessments. The following subsections explain how a blend of these four

[11]Documents of initiatives consulted included the UK's now-defunct National Strategies (professional development programmes instituted in schools in England from 1997 to 2011); Framework for K-12 Science Education (NRC); National Science Education Standards (NSES); Benchmarks for Science Literacy; Science Framework for the 2009 National Assessment of Educational Progress (NAEP); TIMSS 2015 Science Framework; and PISA 2015 Science Framework.

[12]As a former colony, Singapore's curriculum, teaching and assessment have traditionally been influenced by Britain's relatively mature educational developments (Gopinathan, 1991). The historical association suggests that a comparison of the two curricula could give rise to useful insights. A summary of the science programme of study in the current national curriculum of England is in Appendix 3 of the thesis.

approaches defined and described the types of science knowledge and science practices posited in the construct of scientific reasoning.

4.3.1 Types of Knowledge

This subsection describes the three types of knowledge assessed in the PSRT. As mentioned in the literature review, the first of these three is science content. This refers to the body of scientific principles, facts, concepts, and theories that provide an account of phenomena in the material world (Lawson, 2002). The next is procedural knowledge, or understanding of experimental concepts and methods (Gott, Duggan, Roberts, & Hussain, 2015). The third kind of knowledge is epistemic knowledge, which focuses on understanding the criteria and basis for justifying knowledge and beliefs (Duschl & Osborne, 2002).

Since the participating pupils were from Singapore, this study assessed the three knowledge types by referring to the national primary science curriculum. Reference to the curriculum was important because the PSRT was a "knowledge-rich" test (see section 1.4). In the curriculum, there are three "domains" of learning: (1) knowledge, understanding and application; (2) skills and processes; and (3) ethics and attitude*s* (Ministry of Education Singapore, 2013). Each domain signifies crucial learning about the nature and practice of inquiry science for primary school pupils (Chin & Poon, 2014; Ministry of Education Singapore, 2013). A closer analysis suggested that the material in these domains broadly reflects Bloom's cognitive, psychomotor and affective educational taxonomies (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956; Chin & Poon, 2014; Krathwohl, Bloom, & Masia, 1964). Like these taxonomies, each domain grouped together similar subject matter for learning in the form of statements called "learning outcomes", which prescribed what pupils are expected to know and be able to do. See Table 4.2 for a list of these domains.

As will be argued later in this subsection, the learning outcomes in the Singapore primary science curriculum provided greater clarity about the nature and range of content knowledge to be attained than about procedural and epistemic knowledge. Therefore, other substantive sources were consulted, which will also be discussed in this same subsection.

Table 4.2

Domains in the primary science curriculum of Singapore

Knowledge, Understanding, and Application	Skills and Processes	Ethics and Attitudes
Scientific phenomena, facts, concepts, and principles	Skills	Curiosity
Scientific vocabulary, terminology, and conventions	Observing	Creativity
Scientific instruments and apparatus, including techniques and aspects of safety	Comparing	Integrity
Scientific and technological applications	Classifying	Objectivity
	Using apparatus and equipment	Open-mindedness
	Communicating	Perseverance
	Inferring	Responsibility
	Formulating hypothesis	
	Predicting	
	Analysing	
	Generating possibilities	
	Evaluating	
	Processes	
	Creative problem-solving	
	Decision-making	
	Investigation	

Note. Adapted from Ministry of Education Singapore (2013).

Content Knowledge: Analysis of the prescribed curriculum showed that the domain of *knowledge, understanding, and application* contains the content knowledge. This domain has five broad disciplinary themes, which outline essential scientific concepts for building advanced understanding. Each theme contains learning outcomes organised into biological and physical science topics to "communicate a more coherent and integrated understanding of science" (Chin & Poon, 2014, p. 33). This thematic approach in the curriculum thus ties in seemingly disparate phenomena and facts by incorporating them into a common narrative (Poon, 2014). For instance, the theme of "cycles" emphasises the cyclical and predictable processes underlying the life cycles of animals and physical changes such as states of matter and the water cycle (Ministry of Education Singapore, 2013, p. 24). Accompanying

every theme are key "takeaways" to help users unpack its basic ideas. See Table 4.3 for these takeaways in the disciplinary themes.

Table 4.3

Essential "takeaways" of themes in primary science curriculum of Singapore

Theme	Essential "Takeaways"
Diversity	There is a great variety of living and non-living things around us. We can classify living and non-living things based on their similarities and differences to understand them better. Maintaining the diversity of living things around us ensures their continual survival.
Cycles	There are repeated patterns of change around us. Observing cycles helps us to make predictions and understand things around us.
Systems	A system has different parts. Each part has its unique function. Different parts / systems interact to perform function(s).
Energy	Energy is required to enable things to work or move. There are different forms of energy, and they convert from one form to another. Some sources of energy can be depleted, and we play an important role in energy conservation.
Interactions	There are interactions among humans, living, and non-living things in the environment. We can interact with the environment and make positive or negative impacts. We play an important role in conservation to ensure continuity of life and availability of resources.

Note. Adapted from Ministry of Education Singapore (2013).

Procedural Knowledge: The second domain of *skills and processes* contains learning outcomes that draw on a suite of 14 skills and processes listed in Table 4.2 (Ministry of Education Singapore, 2013). A closer inspection of these skills and processes reveals some similarities with Gagné's (1968) basic and integrated process skills discussed in the literature review chapter (Table 2.1). However, unlike Gagné's approach, the goals of skill and process development in the Singapore curriculum relate to the attainment of disciplinary knowledge. Therefore, the phrasing of the outcome statements in the second domain identified both the skill or process *and* the applicable content knowledge (Lee, Kim, & Yoon,

2015). In Figure 4.2, a learning outcome from a physical science topic of "energy forms and uses" illustrates how this incorporation is done. Here, pupils learn how variables impact on shadow formation through the process of *investigation.* [13] Pupils also need to *communicate*[14] (a skill in the curriculum) by representing their data in various forms.

Investigate the variables that affect shadows formed and communicate findings. - shape, size and position of object(s) - distance between light source-object and object-screen

Figure 4.2. A physical science learning outcome demonstrating the incorporation of content and skills. Adapted from Ministry of Education Singapore (2013, p. 50).

Another example shown in Figure 4.3 is from the life science topic of "interactions within the environment". This learning outcome focuses on the mode and consequence of taking measurements to understand the impact of biotic and abiotic factors on the environment.

Observe, collect and record information regarding the interacting factors within an environment.

Figure 4.3. A life science learning outcome demonstrating the incorporation of content and skills. Adapted from Ministry of Education Singapore (2013, p. 45).

The two examples in Figures 4.2 and 4.3, as well as analysis of the other statements, suggested that learning outcomes in the second domain relate to the collection, justification, representation, and interpretation of evidence through the application of skills and processes. This argument implies the embedding of procedural knowledge within the outcome statements.

[13] The curriculum defined this process as "to find out by carrying out experiments" (Ministry of Education Singapore, 2013, p. 77).
[14] This is the skill of "transmitting and receiving information presented in various forms – written, verbal, pictorial, tabular or graphical" (Ministry of Education Singapore, 2013, p. 9).

Without explicit guidelines from the curriculum, I consulted the science curriculum documents of the English school system, as well as psychological and science education research, to clarify the nature and range of procedural knowledge appropriate for primary pupils. The documents from the English system articulated the procedural knowledge required from pupils in a newly renamed strand – *working scientifically*, from *scientific inquiry* in the previous programme of study (Department for Education, 2013). The documents also specify how this knowledge should progressively develop as pupils advance in grades and proficiency. Brief details, along with a summary of the science programme of study in the current national curriculum of England are given in Appendix 3.

Other sources consulted for evidence of early competencies came from comprehensive reviews (see, e.g., Zimmerman, 2007) and studies in recent child development research. In well-designed studies, young school children can differentiate between hypotheses and evidence (Mayer et al., 2014). They can also grasp how the set-up of experiments for testing hypotheses leads to conclusive and inconclusive conclusions (Bullock et al., 2009). By the age of about eight years, there is a preference for controlled experiments over confounded ones, even though the children are not yet capable of devising a COV strategy without assistance (Bullock et al., 2009). However, young children often fail to interpret empirical evidence involving two or more variables (Kuhn et al., 1988). Children disregard or distort data when evidence contradicts their prior beliefs or knowledge (Amsel & Brock, 1996). Overall, the research suggests that primary school children have basic experimentation and evidence evaluation proficiencies (Mayer et al., 2014). Helpfully, science education research indicates there are basic procedural ideas that can be developed into more advanced understanding (Gott & Duggan, 1996; Gott et al., 2015).

This research then curated and drew from a set of 10 core procedural knowledge typically expected of primary science pupils from these curricular and research sources for assessment in the PSRT. As shown in Table 4.4, this set encompassed basic notions about variables, measurement, management of data, and instrumentation.

Table 4.4

Curated set of 10 basic procedural knowledge appropriate for primary-aged science pupils

	Basic Procedural Knowledge
1	Concept of variables including dependent, independent, and control variables;
2	Concepts of measurement e.g., quantitative (measurements), qualitative (observations), the use of a scale, categorical and continuous variables;
3	The ways of assessing and minimising uncertainty, such as repeating and averaging measurements;
4	The mechanisms to ensure the replicability and accuracy of data;
5	Common ways of abstracting and representing data using tables, graphs, and charts, and their appropriate use;
6	The control of variables strategy and its role in experiments;
7	Measurement scale to obtain sensible values and sensible range of values for quantities;
8	Relationship between the choice of instruments and the required range, interval, and accuracy;
9	The nature of an appropriate design for a given scientific question, e.g., experimental, field-based or pattern seeking;
10	Knowing about anomalous data and having a strategy for handling these.

Note. Adapted from Department for Education (2015); Gott and Duggan (1996); Gott et al. (2015).

Epistemic Knowledge: The third domain of *ethics and attitudes* in the curriculum lists seven attitudinal and dispositional learning outcomes (Ministry of Education Singapore, 2013). These statements reference three broad epistemic notions from the perspective of the pupil-learner. As shown in Table 4.5, the first of the three epistemic notions requires learners to appreciate the inter-dependent nature of knowledge and data. For instance, pupils need to adopt an “open-minded” attitude toward new evidence, which may change their current understanding. The second advocates the view that science is a collective, social endeavour. As such, pupils can contribute to the scientific enterprise by questioning and remaining “curious” and “creative”. They must “persevere” in the search for evidence and be “responsible” when using evidence. The third epistemic idea deals with the

justification of evidence. Therefore, pupils must be "objective" in the search for evidence to validate findings. They must also show "integrity" when managing and reporting data.

Table 4.5

Descriptions of the seven outcomes in the domain of attitudes and ethics

Outcome	Description
Curiosity	Desire to explore the environment and question what they find.
Creativity	Suggest innovative and relevant ways to solve problems.
Integrity	Handle and communicate data and information with integrity.
Objectivity	Seek data and information to validate observations and explanations objectively.
Open-mindedness	Accept all knowledge as tentative and be willing to change their view if the evidence is convincing.
Perseverance	Pursue a problem until a satisfactory solution is found.
Responsibility	Show care and concern for living things and awareness of the responsibility they have for the quality of the environment.

Note. Adapted from Ministry of Education Singapore (2013).

For greater clarity about aspects of epistemic knowledge that were appropriate for primary pupils, this study reviewed the developmental psychology and NOS literature. Findings about the epistemological development of primary pupils are mixed. For instance, some studies suggest that they have weak epistemic understanding and consistently view scientific knowledge as irrefutable (Sandoval & Çam, 2011). Pupils up to secondary school levels are also largely unaware that scientific knowledge is socially constructed. Rather, they believe that such knowledge comes as truth derived from experiments (Driver et al., 1996). Contrary to such findings, research (see, e.g., Sodian, Zaitchik, & Carey, 1991) has shown that five-year-old children can demonstrate basic epistemic ideas (e.g., requirement to check results) although they have neither the specialist vocabulary nor the appropriate knowledge to discuss scientific ideas such as "experiment". These collective findings led Sandoval (2005) to conclude that pupils' acquisition of the appropriate epistemological beliefs and practices can be uneven, even with instruction.

The review also identified a set of basic and formal epistemological ideas that pupils ought to know before they leave school (Osborne, Collins, Ratcliffe, Millar, & Duschl, 2003; Sandoval, 2005). In its basic form, Sandoval (2005) suggested four epistemological ideas for science learners. Essentially, these ideas posit that: constructed scientific knowledge needs to be examined, using a diverse array of scientific methods and evaluation criteria. Also, there are many kinds of scientific knowledge, such as theories, hypotheses, and laws, and each differs by their ability to explain and predict. Finally, scientific knowledge asserts varying levels of truth depending on evidence, usage, and beliefs.

In another basic formulation, Driver et al. (1996) identified three areas that pupils require for understanding: the goals of scientific investigations; the nature of different forms of scientific knowledge (such as experimental and theoretical knowledge); and the influence of social norms on scientific work. Formal epistemology elaborates on these basic ideas and draws from the “consensus” or various widely acceptable accounts of NOS (Erduran & Dagher, 2014b). Recent work emphasises the community and reasoning aspects of epistemological knowledge (Duschl & Grandy, 2013). Key to this current work are the criteria for justifying scientific claims, the types of investigative methods and instrumentation, and the use of models to present data and explain findings (Duschl, 2007; Erduran & Duschl, 2004).

Other than ideas about the epistemology of science which pupils should master, Sandoval (2005) identified methodological shifts for eliciting inferences about their abilities in the NOS literature. Assessments to test epistemic knowledge have conventionally focused on collecting pupils’ self-reports and views (see, e.g., Lederman et al., 2002). These assessments require pupils to describe what they think scientists do and how they view the role of experiments and theories. However, recent studies have drawn out similar ideas in problem-solving scenarios using science content. For instance, Driver et al. (1996) investigated pupils’ epistemic notions with the use of electric circuits. Sandoval (2005) argued that this approach – of presenting young children with knowledge that they can relate and respond to – draws out their implicit understanding.

Another methodological initiative noted by Sandoval (2005) is the use of argumentation discourse to enhance epistemic understanding. Findings and developments from this field suggest that the use of verbal and written explanations and arguments is a fruitful approach to assessing important aspects of epistemic understanding (Jiménez-Aleixandre & Erduran, 2007). For example, some recent studies with primary-aged pupils reported that paper-and-pencil tests using multiple-choice items could elicit basic epistemic knowledge involving simple claims and evidence (see, e.g., Osterhaus et al., 2017).

Given the above discussion, the paper-based PSRT assessed age-appropriate epistemic notions by incorporating familiar content knowledge. There was no direct testing of advanced epistemic knowledge (e.g., "What is evidence?"); instead, only three basic epistemic ideas (see Table 4.6) adapted from the guidelines in the curriculum and the literature sources discussed earlier are tested. Where feasible, the item design required the demonstration of reasoning and justification (e.g., identify supporting evidence).

Table 4.6

Three basic epistemic knowledge appropriate for primary-aged science pupils

	Basic Epistemic Knowledge
1	Relationship between claim (explanation/argument) and evidence E.g., Claims and evidence are different types of scientific knowledge; evidence can support as well as reject claims; strength of claims depend on the evidence.
2	Criteria for evaluating and using evidence E.g., The way evidence is collected affects its precision and reliability; facts, principles and theories are important to explain the evidence; relevance of evidence to claims need reasoned arguments; measurement error affects the plausibility of evidence; Models and modelling are useful to explain evidence.
3	Influence of social contexts in evidence use and evaluation E.g., Individual or group interest can prejudice the interpretation and use of evidence; the evaluation and use of evidence involve many people; people (e.g., scientists) may disagree over the use and evaluation of evidence.

Note. Adapted from Driver et al. (1996); Kind (2013); Osborne et al. (2003); Sandoval (2005).

4.3.2 Science Practices

According to the Singapore primary science curriculum, the eleven skills in the second domain have three uses. These uses, as shown in Table 4.7, are to: (1) engage with an event, phenomenon or problem; (2) collect and present evidence; and (3) reason: make meaning of information and evidence (Ministry of Education Singapore, 2013). A discussion of the curriculum-listed skills and their corresponding uses follows after the table.

Table 4.7

A comparison of the skills and uses in the Singapore's primary science curriculum and the science practices in this study

Primary Science Curriculum		Science Practice in Current Study	
Skill	Use	Science Practice	Brief Description
Formulating hypothesis Generating possibilities Predicting	Engaging with an event, phenomenon, or problem	Giving scientific explanation	Working with scientific ideas, models, and evidence in hypotheses and explanations
Observing Using apparatus and equipment	Collecting and presenting evidence	Designing and evaluating investigation	Developing appropriate strategies and methodologies to collect scientific evidence, through experimental or other means
Comparing Classifying Inferring Analysing Evaluating	Reason: make meaning of information and evidence	Interpreting and analysing data and evidence	Interpreting and analysing data and other scientific evidence to identify and critique outcomes such as conclusions, hypotheses and arguments

Note. Adapted from Ministry of Education Singapore (2013).

The three science practices in the scientific reasoning construct investigated in this study have similar uses as those identified in the curriculum. As outlined in the literature review chapter (see section 2.5), the science practices are *giving scientific explanation*, *designing and evaluating investigation*, and *interpreting and analysing data and evidence*. These

practices are the activities that scientists engage in when attempting to answer queries about the functioning of the natural and material world.

The table shows that the basis for drawing parallels between the uses of the skills and the science practices is the set of skills identified in the curriculum. For instance, as described in the table, the curriculum identified the formulation of hypothesis, generation of possibilities, and prediction as the skills for "engaging with an event, phenomenon or problem". These identified skills are necessary for the development and evaluation of scientific explanations, which is also the focus of the science practice "giving scientific explanation". Similarly, an examination of the other skills and the corresponding uses showed alignment with the focuses of the other two practices of this study. Therefore, this examination suggested that Singapore primary pupils would be familiar with the competencies of the science practices.

4.3.3 Sub-Practices and Attainment Levels

This subsection explicates the specific competencies in each of the three science practices. Organised into broad ideas called *sub-practices*, the descriptors of competencies outlined progressions, or *levels,* of increasingly developed and sophisticated attainment. This approach to arranging competencies is modelled after organised teaching and assessment sequences, known as learning progressions, and acknowledges that young learners of science require sustained engagement with ideas and concepts corresponding to their general development and maturation (Leach & Scott, 2003). Researched extensively for over 20 years, conceptions of learning progressions have shifted considerably (Gotwals & Alonzo, 2012; Ng, in press).

Early versions underscored the natural learning and developmental trajectories of learners, while current perspectives focus on incorporating instruction and emphasise the transferability of learning across contexts (Forster & Masters, 2010). For instance, a recent definition viewed learning progressions as continua of coherent and increasingly complex

learning sequences informed by the disciplinary domain and impacted on by curriculum and instructional practices (National Research Council, 2007).

In this research, the content and sequencing of descriptors in each sub-practice followed the broad recommendations of Harlen (2006). Harlen argued that school children do not necessarily develop competencies in a developmentally typical and predictable manner, and therefore advocated that existing progresssion frameworks be carefully adapted or treated as provisional. Recent writings and research have emphasised the tentative and evolving features of learning progressions, echoing Harlen's stance of almost a decade earlier (National Research Council, 2012; Pellegrino, 2017).

Harlen also recommended that descriptions of competencies in a progression model be guided by evidence drawn from a few sources, such as experts, research, curriculum documents, and initiatives. In keeping with her recommendations, the descriptors developed in this study are "coarse-grained statements suitable for summarising attainments" (Harlen, 2006, p. 104) to be subjected to future refinements. Initially drawn from the literature and some of the curricular initiatives, the level descriptors in each sub-practice delineate the range of significant behaviours to target corresponding reasoning according to an adapted model of development for young children (Curriculum Council, 2005; Leach & Scott, 2003). Shown in Figure 4.4, this model is a roadmap charting the progression of four discrete and hierarchical attainment stages from naïve understanding to the transfer and wider application of knowledge. Table 4.8 explains these stages.

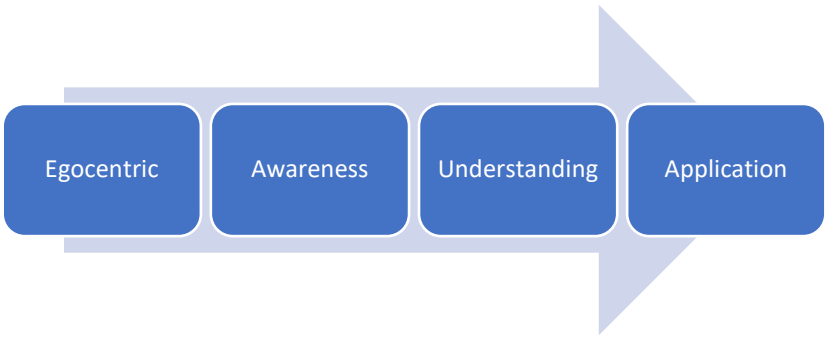


Figure 4.4. Four attainment stages to guide the development of level descriptors in each sub-practice. Adapted from Curriculum Council (2005).

Table 4.8

Four stages of learning attainments

Stage	Description of Learning Attainment
Egocentric	Ideas are viewed as disparate entities; ideas remain largely unconnected; learners are unable to interpret ideas
Awareness	Learners are aware of context-specific concepts and processes; perspective remains simplistic and idiosyncratic; there is developing capacity to explain, interpret evidence, identify causal factors, and generalise
Understanding	Learners are developing greater understanding of concepts and processes; there is increasing ability to infer, generalise, and conclude; there is growing cognisance of the complexity of inter-relationships between ideas
Application	Learners have well-developed understanding of concepts and processes; there is ability to generalise, draw conclusions, and evaluate complex issues; there is ability to propose solutions to complex issues, and anticipate and resolve potential and future impact

Note. Adapted from Curriculum Council (2005).

Items in the Draft PSRT were deliberately constructed to elicit behaviours outlined in the chosen competency level of specific sub-practices. Evidence from the performance in the questions was then used to determine whether the pupils had achieved the targeted levels of attainment. The rest of this subsection describes the sub-practices and levels in each of the three practices. The identification of specific competencies underlying each sub-practices are drawn largely from historical and current curricular initiatives and reforms highlighted at the beginning of this section 4.3 (see also footnote 11).

Practice 1: Giving Scientific Explanations

This practice refers to how pupils work with scientific ideas, models, and evidence to recognise, suggest, and evaluate explanations. Therefore, modelling and providing explanatory accounts of scientific phenomena are the two major competencies involved. Accordingly, the first of its two sub-practices (sub-practice 1.1) measures pupils' abilities to identify, use, and generate models when explaining or hypothesising about phenomena.

The five levels in this first sub-practice seek to develop and assess distinctive reasoning associated with modelling competencies such as analogical and inductive reasoning (Erduran & Duschl, 2004). See subsection 2.5.2 for a discussion of these forms of reasoning. Accordingly, the paper-and-pencil draft and final versions of the PSRT utilised two-dimensional symbolic and visual modes, as well as verbal modes, to represent models in test items (Gilbert, 2004). Symbolic modes consisted of units of measurement or the use of arrows in food webs to denote the abstract idea of energy flow. Visual modes consisted of simple pictograms of electronic devices assembled in circuit diagrams, classification systems, schematic diagrams of complex processes or mechanisms, and diagrammatic renditions of physical and living phenomena. Verbal modes employed written descriptions of the parts or relationships in models (e.g., “batteries are arranged in series”). Table 4.9 shows the learning outcomes in the five levels of this sub-practice.

Table 4.9

Learning outcomes in the five levels of sub-practice 1.1: Identify, use and evaluate models and representations in explanations and hypotheses

Level	Learning Outcome
1.1.1	Represent things in the real world using simple models
1.1.2	Use simple models to identify scientific ideas
1.1.3	Use abstract ideas or models of more than one step when describing processes or phenomena usually in familiar situations
1.1.4	Evaluate abstract ideas or models used to explain processes or phenomena or use abstract ideas or models in new situations
1.1.5	Develop new models or justify the selection of model; Make explicit connections between abstract ideas and/or models in explaining processes or phenomena

Note. Adapted from Qualifications and Curriculum Development Agency (2009); National Research Council (2012); NGSS Lead States (2013); and Schwarz et al. (2009).

Competencies in sub-practice (1.2) involve the recall and application of knowledge. Among its four levels, the lowest (1.2.1) describes the use of observational skills or simple ideas to

identify evidence. Typically, items in the PSRT that targeted this level invoked lower-level mental operations (Koufetta-Menicou & Scaife, 2000). Items corresponding to higher levels involved more challenging cognitions. For instance, the fourth level assessed knowledge application in problem-solving contexts. Although items might include models such as diagrams and graphs, the predominant reasoning of interest is hypothetico-deductive in nature. Such items testing the logical application of general principles are commonly encountered by primary pupils, as their initial learning emphasises rule-based learning such as basic content and declarative knowledge (Goswami, 2010). The learning outcomes in this second sub-practice are presented in Table 4.10.

Table 4.10

Learning outcomes in the four levels of sub-practice 1.2: Recall and apply appropriate scientific knowledge

Level	Learning Outcome
1.2.1	Draw on observations and simple ideas (scientific or otherwise) to offer answers to questions; Respond to suggestions (in items) to identify some evidence (in the form of information, observations, or measurements) needed to answer a question or select an answer to a simple question from options
1.2.2	Use scientific ideas or knowledge to answer or suggest solutions to simple problems with short responses (e.g., single answers or short phrases) or select an answer from options
1.2.3	Use scientific ideas or knowledge to describe or explain processes or phenomena; Answer or suggest solutions to complex problems
1.2.4	Apply appropriate scientific knowledge linked logically or inferred from models and evidence to construct and evaluate explanations

Note. Adapted from Qualifications and Curriculum Development Agency (2009); National Research Council (2012); NGSS Lead States (2013); Scaife (2000); and Gotwals, Songer, and Bullard (2012).

Practice 2: Designing and Evaluating Investigations

This practice refers to the way pupils ask questions, hypothesise, and develop appropriate strategies and methodologies to collect scientific evidence through experimental or other

means. Given that primary-aged pupils do not have much exposure to laboratory learning, the three sub-practices in this practice focus on basic experimentation competencies involved in the planning of investigation, manipulation of variables, and the collection of precise and reliable data. These competencies are elaborated in the following discussion.

The first sub-practice (2.1) focuses on the planning of effective scientific investigations or the evaluation of approaches. Its five level descriptors ranged from basic competencies at the lower end, such as the selection of appropriate equipment, to the justification of investigative strategies at the higher end. Overall, the learning outcomes focus on planning aspects such as methodology, instrumentation, and data collection (Gott & Duggan, 2003). The types of reasoning elicited depend on the assessment objectives, item context, and item anatomy. See Table 4.11 for the learning outcomes in this sub-practice.

Table 4.11

Learning outcomes in the five levels of sub-practice 2.1: Plan for effective scientific investigations or evaluate approaches

Level	Learning Outcome
2.1.1	Select or describe appropriate equipment or methods to address specific questions or ideas under investigation
2.1.2	Explain why particular pieces of equipment or information sources (e.g., methods) are appropriate for the questions or ideas under investigation
2.1.3	Collect data choosing appropriate ranges, numbers, and values for measurements and observations
2.1.4	Formulate scientific questions by synthesising information from a range of sources or evaluate whether these ideas can be investigated scientifically
2.1.5	Justify choice of strategies for investigating different kinds of scientific questions, using scientific knowledge and understanding

Note. Adapted from Qualifications and Curriculum Development Agency (2009); National Research Council (2012); NGSS Lead States (2013); and Gott and Duggan (2003).

The second sub-practice describes five levels of competencies related to the identification and manipulation of experimental variables. While primary pupils are not expected to have an in-depth understanding of the variable types and their consequent impact on the level of

measurement, they are required to appreciate the differences between independent and dependent variables (Phipps & Duggan, 1995). Therefore, the main form of reasoning tested in this sub-practice is the COV strategy, which is a reasoning closely associated with formal-operational development (Adey et al., 2012). Depending on the testing objectives, design, and context, items assessing COV competency could measure several fundamental types of reasoning, such as seriation, classification, and combinatorial. Table 4.12 presents the learning outcomes in the levels of this second sub-practice.

Table 4.12

Learning outcomes in the five levels of sub-practice 2.2: Identify and manipulate variables within an investigation

Level	Learning Outcome
2.2.1	Recognise one or more key variables in simple investigations
2.2.2	Identify one or more key variables in unfamiliar or complex investigations, or select the most suitable to investigate
2.2.3	Understand the role of key variables in fair tests Identify and explain the role of key variables in investigations using scientific knowledge and understanding
2.2.4	Identify key variables in complex contexts, explaining why some cannot readily be controlled and planning appropriate approaches to investigations to take account of this
2.2.5	Justify choice of strategies for investigating different kinds of scientific questions, using scientific knowledge and understanding

Note. Adapted from Qualifications and Curriculum Development Agency (2009); National Research Council (2012); NGSS Lead States (2013); and Gott and Duggan (2003).

The third and final sub-practice describes competencies related to the collection of precise and accurate data as evidence. These competencies included making whole-number measurements, identifying appropriate measurement intervals, and explaining how sources of errors can be reduced in data collection methods. Like many competencies from other sub-practices, a few reasoning types are involved, including proportional and quantitative reasoning. Although more commonly taught and assessed in the mathematics curriculum,

these reasoning types are necessary to identify associations between physical quantities and scientific concepts introduced to primary science pupils such as rates (e.g., evaporation), mass, and volume (Adey & Csapó, 2012). Learning outcomes in this sub-practice also register rudimentary competencies related to measurement errors. These errors, whether attributed to human or instrumentation faults, come about during the measurement process (Gott & Duggan, 2003). See Table 4.13 for the learning outcomes associated with this third sub-practice.

Table 4.13

Learning outcomes in the four levels of sub-practice 2.3: Collect precise and reliable data during the gathering of evidence

Level	Learning Outcome
2.3.1	Make some accurate observations or whole-number measurements relevant to questions or ideas under investigation Identify differences in repeated observations or measurements or provide straightforward explanations for differences
2.2.2	Make observations or measurements, identifying the ranges and intervals Repeat observations or measurements where appropriate, selecting suitable ranges and intervals
2.2.3	Explain how data collection methods minimises or increases error Explain how sources of error are considered in order to collect reliable data
2.2.4	Choose and justify data collection methods that minimise error, and produce precise and reliable data

Note. Adapted from Qualifications and Curriculum Development Agency (2009); National Research Council (2012); NGSS Lead States (2013); and Gott and Duggan (2003).

Practice 3: Interpreting and Analysing Data and Evidence

This practice refers to the way how pupils interpret and analyse data and scientific evidence to identify outcomes and draw conclusions. The two major competencies are the analysis of data, and the evaluation of evidence for the purpose of appraising conclusions, hypotheses, and arguments. Competencies delineated in this practice are analogous to the

higher-order mental processes referred to in traditional learning taxonomies, such as Bloom's (1956) cognitive taxonomy (see subsection 4.4.3), or learning approaches such as Gagne's (1968) hierarchy of processes discussed in Chapter 2 (see subsection 2.2.2).

Level descriptors in the first of two sub-practices (3.1) outlined the drawing of inferences to support claims from the evaluation of evidence represented in various graphical or pictorial forms. As shown in Table 4.14, the learning outcomes across the five levels in this sub-practice outlined an increasing proficiency to draw explanatory insights from evidence using theories such as the recognition of constraints in the evidence, and the need to draw on other sources to substantiate conclusions (Kuhn, 2011; Kuhn & Franklin, 2006).

Table 4.14

Learning outcomes in the levels of sub-practice 3.1: Process, transform, and analyse data

Level	Learning Outcome
3.1.1	Identify straightforward patterns in observations or in data presented in various formats, including tables, pie and bar charts, and graphs
	Draw direct conclusions from data presented in various formats
	Describe what has been found in experiments or investigations, linking cause and effect
3.1.2	Identify less straightforward patterns in data presented in various formats
	Draw less direct conclusions from data presented in various formats
3.1.3	Draw valid conclusions or predictions that utilise more than one piece of supporting evidence or data, including numerical data and line graphs
	Interpret data from a variety of formats, recognising obvious consistencies and inconsistencies
3.1.4	Select and manipulate the most relevant data and information and use these to contribute to conclusions or predictions
	Draw conclusions that are consistent with the evidence collected and explain these using scientific knowledge and understanding
	Make valid comments on the quality of data

Note. Adapted from Qualifications and Curriculum Development Agency (2009), National Research Council (2012), NGSS Lead States (2013), and Gott and Duggan (2003).

The second and last sub-practice (3.2) describes competencies related to the evaluation of evidence to construct reasoned discourses, such as explanations and scientific argumentation. Descriptors drew heavily on learning progressions designed to monitor the proficiency trajectories of argumentation in science (see, e.g., Berland & McNeill, 2010; Lee et al., 2014; Osborne et al., 2016). For instance, a commonality among these learning progressions is that they generally begin with the formulation of claims and culminate in the rebuttal or appraisal of claims (Ng, in press). Table 4.15 shows the learning outcomes of this sub-practice.

Table 4.15

Learning outcomes in the five levels of sub-practice 3.2: Evaluate evidence to construct and critique conclusions, hypotheses, and arguments

Level	Learning Outcome
3.2.1	Identify scientific evidence used to construct, support or refute claims, conclusions, hypotheses, or arguments
3.2.2	Identify and explain how the evidence has been used to make the required conclusions or claims; Describe evidence used to support or refute ideas, hypotheses or arguments
3.2.3	Suggest reasons based on scientific knowledge and understanding for any limitations or inconsistencies in evidence collected; Identify relationships between variables, using them to inform conclusions and make further predictions
3.2.4	Explain how data can be interpreted in different ways and how unexpected outcomes could be significant; Critique claims, conclusions, hypotheses, and arguments
3.2.5	Critically interpret, evaluate, and synthesise conflicting evidence

Note. Adapted from Qualifications and Curriculum Development Agency (2009), National Research Council (2012), NGSS Lead States (2013), Osborne et al. (2016), and Ng (in press).

In conclusion, the level descriptors of the sub-practices in each of the three science practices outlined the complete repertoire of competencies or behaviours exhibited when reasoning with content, procedural, or epistemic knowledge. As mentioned at the start of

this subsection, these descriptors are subjected to further refinements. Accordingly, the descriptors of competencies characterising items in the Draft PSRT underwent expert review in the next phase. To assist experts in this review, every descriptor in a sub-practice was accompanied by a short explanatory or clarification note to guide interpretation. The following chapter reports the experts' findings about the accuracy and utility of these descriptors. See Appendix 5 for a compilation of the descriptors and their explanatory notes [15] in the seven sub-practices across the three science practices. Table 4.16 summarises the seven sub-practices in the scientific reasoning framework.

Table 4.16

The three practices and their sub-practices in the scientific reasoning construct

Practice	Sub-Practice
Practice 1: Giving scientific explanation	1.1 Identify, use and evaluate models and representations in explanations and hypotheses 1.2 Recall and apply appropriate scientific knowledge
Practice 2: Designing and evaluating investigative approaches	2.1 Plan for effective scientific investigations or evaluate approaches 2.2 Identify and manipulate variables within an investigation 2.3 Collect precise and reliable data during the gathering of evidence
Practice 3: Interpreting and analysing data and evidence	3.1 Process, transform, and analyse data 3.2 Evaluate evidence to construct and critique conclusions, hypotheses, and arguments

After developing the construct and behavioural descriptors, the next steps were construction of test items and assembly of the instrument. This involved reviewing the conceptual frameworks of learning theories, as what is valuable in and about learning determines the function, form, and substance of assessment (Baird, 2014). Also, there was consultation of

[15] In Appendix 5, the explanatory notes are placed alongside the descriptors in an adjoining *Look for* section.

practical frameworks, which are workable mechanisms or processes originating from personal experiences, long-standing practices of authorities, or collective wisdom of practitioners (Lester, 2005). The practical frameworks referred to in this study are the “best practices” of item design, item generation, and validation methods widely practiced and advocated by the education and assessment communities. Other practical frameworks consulted were established methods of test assembly (e.g., matrix-design principles for booklet design). In effect, phase I was an instrument development phase, which not only considered diverse perspectives and practices from theory and evidence but also took on board the emic or “insider” knowledge of participants and experts (Onwuegbuzie et al., 2010). The following sections report the impact of learning theories and practical frameworks on the items in the Draft PSRT.

4.4 Development of Test and Items

Learning theories were the key conceptual frameworks consulted to build items after the development of the construct and behavioural descriptors. Learning theories are not only important for understanding the learning process, but also serve as the basis to prove that learning has occurred (Baird, Hopfenbeck, Newton, Stobart, & Steen-Utheim, 2014). Baird et al. (2017) argued recently for greater alignment among learning theories, assessment design, and assessment theories to ensure that assessments accurately measure desired behaviours or performances, which Baird (2014) termed “valuable learning” (p. 21). Their recommendation echoed clarion calls repeatedly issued about the gap in understanding of the impact of cognitive models on the testing of school children (National Research Council, 2001; Shepard, Penuel, & Davidson, 2017).

This section begins by explaining the need for metaphorical language in accounts of the learning process. The subsequent subsection briefly considers three predominant theories and their associated metaphors. These theories have distinctive postulations about how people learn, which have consequential bearing on assessments. This section then

concludes with a discussion of the rationale for adopting the tenets of cognitive learning theories to guide the assessment design and implementation of the PSRT.

4.4.1 Centrality of Learning Metaphors

Hager (2006, p. 110) asserted that "a single account of learning is unlikely to be viable" on the basis that the relevance and usage of each account depends on the context where learning is taking place. Moreover, descriptions or explanations about learning in the English language almost always evoke the use of figurative devices, such as metaphors (Reddy, 1978). The use of metaphorical statements to denote scientific relationships and processes is particularly pervasive in the learning of science. To illustrate, the following descriptions are found in introductory science texts for young learners: "Imagine the many millions of chloroplasts in just one lettuce leaf, each a tiny factory for producing sugars and starches; seeds (of fruits) can be dispersed to new locations, where they will not have to compete with the parent plant for soil … and sunlight" (Darian, 2000, p. 174). The preceding description also highlights the ability of metaphors to distil significant and easy-to-grasp similarities or analogies between or within ideas, processes, people, objects, or events (Scheffler, 1960). Essentially, metaphors are not mere linguistic tools but serve as crucial components of conceptual thoughts and actions.

Ground-breaking research by Lakoff and associates further suggests that human thought processes are largely metaphorical (Lakoff & Johnson, 1980/2003, 1999). Thus, a practical way of making sense of how people think about learning is to investigate the use of metaphors in language (Lawley & Tompkins, 2000). As will be shown in the following subsection, metaphors highlight the underlying assumptions associated with three contemporary views of learning, with significantly different implications for assessment practices.

4.4.2 Learning Theories and Assessment Implications

Broadly, there are three groups of learning theories associated with formal learning in school: behavioural, cognitive, and sociocultural. Each makes unique assumptions about the nature of the learning content and the conditions under which learning occurs (Penuel & Shepard, 2016).

Behaviourist theories carry various metaphorical descriptions such as *folk learning* (James & Lewis, 2012), *learning is being taught* (Watkins, 2003), and *common sense* (Hager, 2006). Theories grouped under this learning perspective do not accord emphasis to underlying mental processes and further regard the *mind as a mere container* for disconnected bits of knowledge (Bereiter, 2002). Learning presumably occurs with knowledge *acquisition* – a process independent of the context and the learner. A notable characteristic of this perspective is its advocated use of rewards and penalties to reinforce desired behaviours. These assumptions imply that assessment approaches usually involve assessing the ability to recall discrete information or demonstrate specific skills. Typically, performance occurs under controlled testing conditions, and there is comparison to some pre-defined criteria or norms. Scoring of performance is objective since learning is observable and does not require interpretation by markers (James, 2006a).

Cognitive theories posit learning as *individual sense-making* (Watkins, 2003). In this orientation, learning presumably takes place by active construction of mental representations of the external reality. James (2006b) opined that this perspective views knowledge as a commodity to possess – a stance it has in common with behaviourist approaches. However, the focus of all cognitive theories is the workings of the mind and how it processes, assimilates, and assesses knowledge to problem-solve (Baird et al., 2017). The mind's engagement with specific reasoning (e.g., induction) to construct and deepen understanding is likened to a *computer functioning across networks* (Bereiter, 2002; Searle, 1990). Given these assumptions, the focus of assessments is the extent of cognitive understanding, and the capacity to transfer that understanding to other contexts and solve complex problems (Pellegrino, 2017).

Other implications include timed assessments to elicit higher-order cognition and the involvement of assessor judgement (James & Lewis, 2012). Haladyna and Rodriguez (2013) asserted that these theories underpin most current understanding and practices in educational assessment. For instance, the notion of learning progressions, which arises from cognitive science research, postulates the existence and attainment of developmental sequences of domain knowledge and ability (Black, Wilson, & Yao, 2011). A common testing practice that reflects this idea of progressive mastery is the manner of structuring questions within the same assessment so that test-takers encounter simpler, i.e., less cognitively demanding questions first, before progressing to more complex ones later. This practice encourages test-takers to attempt as much of the assessment as they can.

Like the other two learning theories, sociocultural learning has associations with various metaphors, such as *situated* (Wenger, 1998) or *building knowledge as part of doing things with others* (Watkins, 2003). In this perspective, learning happens at the nexus between an individual and the larger community of interaction (Vygotsky, 1994). The process is socially constructed, mediated by cultural tools (e.g., books or language), and significantly shaped by the environment. Assessments subscribing to this view occur “in situ” in the same learning instances; the evaluation focuses on how well test-takers harness their time, knowledge, and resources in problem-solving tasks. Some examples of such assessments in schools include coursework that takes place over a period of time, or projects that attempt to address real-world problems such as investigating how methods of rubbish disposal can impact on the environment (James, 2017).

In summary, major learning theories view learning in two ways: as *acquisition* or as *participation* (Sfard, 1998). However, no single group of theories can account for the complexities of every learning situation, such as those encountered in formal learning (Hager, 2006). The implicit assumption is that the underlying content in formal learning is important to the community-at-large and socially valued (Baird, 2014).

The adoption of the appropriate learning theory and associated assessment features comes from an initial identification of the learning purpose and objectives (Briggs, 2017). For instance, developing cognitively complex thinking and reasoning, such as intellectual competencies in science is a key focus of educational efforts in the 21st century (DeBoer, 2000). Consequently, the assessment of scientific attainment predominantly draws on cognitive learning theories (Haladyna, 2013). Similarly, in this study, cognitive theories shaped four primary design parameters of the PSRT and items. Specifically, these parameters are the test content, item format, cognitive demand and item features. The next subsection discusses these parameters in turn.

4.4.3 Impact of Cognitive Learning Theories on Item and Test Design

Test Content: In this study, scientific reasoning is the ability to coordinate between claims and evidence, by drawing on three types of knowledge and three science practices. Therefore, PSRT items assessed *content*, *procedural*, and *epistemic knowledge* when engaged in *giving scientific explanation*, *designing and evaluating investigations*, and *interpreting and analysing data and evidence*. As outlined in the literature review (section 2.5) and section 4.3 of this chapter, insights from psychology and philosophy prescribed the knowledge types and developed the behavioural descriptors of the science practices. This subsection extends the discussion by describing the approach of operationalising the defined construct to test and item development.

The TOS guided the design of items by delineating the range and proportion of sampled knowledge and science practices in the PSRT, as well as indicating the intended cognitive demand – one of three possible levels – of every item. As opposed to the “what” (knowledge) and “when” (scientific practice) of reasoning, cognitive demand factors in the extent of mental processing or “how much” (depth) reasoning is assessed by the items. Characterising the PSRT items in the TOS ensured a *balance* of items of the intended level of intellectual challenge from an *adequate* selection of knowledge and practices postulated to form the scientific reasoning construct (Sireci & Faulkner-Bond, 2014). In effect, the TOS

directed the emphasis of the assessment by stipulating the scoring criteria of the tested content (Nitko & Brookhart, 2007). Appendix 6 contains the items of the Draft PSRT. Details such as the actual proportions of sampled knowledge in the TOS are provided later in the test framework (section 4.5).

Cognitive Demand: Cognitive theories also influenced item design and test construction by fore-grounding the notion of cognitive demand. Cognitive demand is the combined mental processing that an item intends to elicit from pupils (Haladyna & Rodriguez, 2013). Cognitive taxonomies – such as Bloom's taxonomy and its revised version, the Revised Bloom's taxonomy, are well-known examples of schemes for categorising the types of cognitive expectations imposed by items and tasks (Nitko & Brookhart, 2007). As an example, the first taxonomy by Bloom and collaborators classified items according to six hierarchical levels – "knowledge", "comprehension", "application", "analysis", "synthesis", and "evaluation" – to signal increasing degrees of complexity (Bloom et al., 1956).

Unfortunately, despite 60 years of research, there exists no empirical support for the validity of these taxonomies (Haladyna, 2013). Confounding investigations about the relationship between item performance, such as item difficulties, and the order of the levels in the taxonomies is the varying intellectual development of test-takers (Haladyna & Rodriguez, 2013). Novices and intellectually strong pupils experience significantly different cognitive demand on identical items. Cultural differences among educational systems, such as the regularity of exposure to testing and the amount of outside-of-school instructional exposure (e.g., tuition), also impact on pupils' perceptions of item difficulty (Baird, Caro, & Hopfenbeck, 2016).

Moreover, classifying the overall demand of items with taxonomic terminology (e.g., comprehension) does not provide prescriptive information about the depth and range of mental processing, nor the amount and type of recalled or applied knowledge required of test-takers (OECD, 2017). Gorin and Embretson (2013) advised that taxonomic labels cannot identify the source of cognitive complexity in items, and lack predictive power.

Therefore, it is not uncommon to categorise items requiring many pieces of easy, recalled knowledge as being cognitively challenging.

Acknowledging the issues with the application of current taxonomies, testing specialists recommend the adoption of simplified classifications for pragmatic reasons (Haladyna & Rodriguez, 2013). This study consequently adapted a model developed by the Organisation for Economic Co-operation and Development (OECD) to categorise the cognitive demand of items in the 2015 PISA science assessment for 15-year-olds. OECD's model referenced Webb's (1999, 2007) Depth of Knowledge (DOK) framework and, for operational feasibility, ordered items according to "low", "medium", and "high" cognitive levels (OECD, 2017). In contrast, Webb's (1997, 1999, 2007) conceptualisation categorised items into one of four levels of cognitive demand (DOK) – "recall", "skill/concept", "strategic thinking", and "extended thinking".

See Table 4.17 for a comparison of the cognitive demand levels in Webb's and the OECD's categorisation schemes with the adapted model. As shown in the table, modifications to the current model took into account the typical level and range of mental processing expected of primary science pupils. Also, the adapted model was deliberately kept simple to enable ease and consistency of application. Subject-matter specialists, such as school teachers and curriculum authorities, used the adapted model to provide feedback in the next phase (phase II) about the suitability and match between the actual cognitive levels elicited from items and the intended levels.

Item Format: A third way in which cognitive theories affect test items is by informing the item format, which refers to the structure and arrangement of item components that pose questions, elicit responses, and facilitate scoring (Haladyna, 2004). Research investigating the fidelity of the format to the tested construct has shown that formats that facilitate the demonstration of pupils' thinking and reasoning provide stronger evidential basis for inferring the quality of these abilities (Haladyna, 2013). Therefore, the potential of each of the three basic items formats to produce evidence of complex thinking abilities varies.

Table 4.17

A comparison of cognitive demand levels across three categorisation schemes

Webb's DOK	Science PISA 2015	Current Study
Level: Recall Recall of information, such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or apply a formula.	*Level: Low* Carry out a one-step procedure; for example, recall a fact, term, principle or concept, or locate a single point of information from a graph or table.	*Level: Low* Carry out a one-or-two-step procedure involving a recall of information, or read data directly from a table or graph.
Level: Skill/concept Engage mental processes beyond habitual responses; e.g., "classify", "organise", "estimate", "make observations", "collect and display data", and "compare data". Also includes activities such as explaining the purpose and use of experimental procedures, and carry out the procedures.	*Level: Medium* Use and apply conceptual knowledge to describe or explain phenomena, select appropriate procedures involving two or more steps, organise/display data, interpret or use simple data sets/ graphs.	*Level: Medium* Identify or select knowledge to describe or explain phenomena; apply procedures involving more than two steps; organise/display data; interpret data from a table/graph.
Level: Strategic thinking Reason, plan, and use evidence. Explain thinking, and make conjectures, or justify responses. Also includes activities such as citing evidence and developing a logical argument for concepts, explaining phenomena regarding concepts, and using concepts to solve problems.	*Level: High* Analyse complex information or data; synthesise or evaluate evidence; justify; reason, given various sources; develop a plan or sequence of steps to approach a problem.	*Level: High* Analyse complex phenomena or data; synthesise or evaluate data or evidence; justify claims or arguments.
Level: Extended thinking Complex reasoning, planning, developing, and thinking over an extended period. Be able to relate ideas within and among content areas. Also includes activities such as designing, conducting, and critiquing experiments; making connections between a finding and related concepts and phenomena; and combining and synthesising ideas into new concepts.		

Note. Adapted from Webb (1999, p. 31) and OECD (2017, p. 41).

The three basic item formats are selected-response objective scoring (SROS), constructed-response objective scoring (CROS), and constructed-response subjective scoring (CRSS). Each SROS item provides a pre-defined list of options, and scoring does not involve markers' judgments. When designed appropriately, these options expose misconceptions or partial understanding. CROS items involve the objective scoring of short written answers. Like SROS items, CROS items can also produce evidence of understanding. However, CRSS items, which contain open-ended prompts to encourage production of extended answers for markers' evaluation, are most likely to give strong inferences about reasoning. See Figure 4.5 for three items that exemplified two of these formats – SROS and CRSS.

For pragmatic reasons, this study used a combination of all three formats to capitalise on the testing advantages afforded by each. Other than the ability to draw out complex and critical thinking abilities, other advantages include broad domain sampling, scoring reliability, and ease of construction (Nitko & Brookhart, 2007). See Table 4.18 for a list of the salient testing benefits associated with each item format. Details of the specific combination of item formats implemented in the PSRT are discussed in the next section (section 4.5).

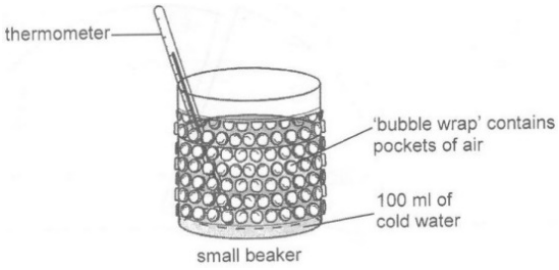
Table 4.18

Testing advantages of the three basic item formats

Advantage	Item Format		
	SROS	CROS	CRSS
Domain sampling	Broad	Narrow	Can be broad
Scoring reliability	Can be high	Can be high	Not as high
Rater effects (e.g., severity)	None	None	May pose a problem
Relative ease of construction	Difficult	Less difficult	Difficult
Elicitation of problem solving and complex reasoning	Possible	Limited	Highly possible

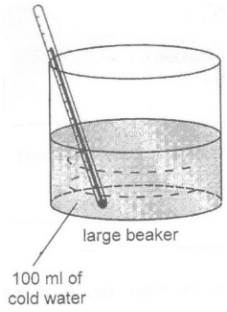
Note. Adapted from Haladyna and Rodriguez (2013).

John wants to study whether the material, 'bubble wrap,' affects how cold water gains heat from the room. The diagram below shows one set-up of his experiment.

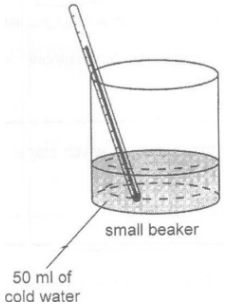


Item 1: Which other set-up shown below, P, Q, or R, must he use in his experiment? Choose your answer by ticking (✓) the box. [1 mark]

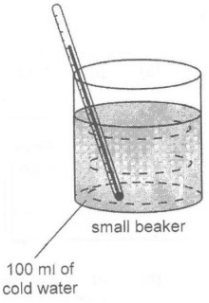
P ☐



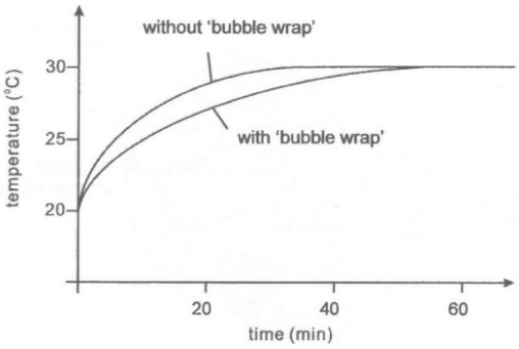
Q ☐



R ☐



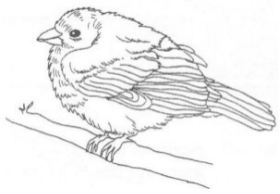
Item 2: John recorded the readings of the thermometer over time. His results are shown below.



Based on his results, John concluded that the beaker with the 'bubble wrap' gains heat more slowly. Is he correct? Circle your answer.

Yes / No / Cannot Tell [1 mark]

Item 3: A bird with a thick layer of feathers is shown below. There are air pockets among the feathers.



Based on the results of John's experiment, explain how the air pockets keep the bird warm in cold air. [1 mark]

Figure 4.5. A three-item question in the Draft PSRT. Items 1 and 2 have three-options SROS formats while item 3 has a CRSS format.

Item Features: Cognitive theories also influence item and test design by informing the impact of item features on test-takers' expected behaviours. In this study, a test item is a "device for obtaining information about a test-taker's domain of knowledge and skills or a domain of tasks that define a construct" (Haladyna & Rodriguez, 2013, p. 3). As the most fundamental unit of a test instrument, a test item consists of a few components (Osterlind, 1998). Common item components are the stem, stimulus, response, key, and distractor. A rich pool of information exists from decades of cognitive science research about how these components impact on test-takers' cognition and test performance (Leighton, 2013). This section briefly defines these components in the context of the study before elaborating on their possible effect.

Operationally, an *item* assesses a scientific concept, principle, understanding, or skill, irrespective of its format. A correct *response* receives a score. Its *stem* is a piece of text that introduces the item and contains a formulated problem. A graphical, visual, or textual *stimulus* may accompany an item. The stimulus provides additional content information for test-takers to act on in addressing the problem. More specifically, every SROS item contains a *key* and a few *distractors*. The former is the correct answer and the latter are the remaining plausible but incorrect options. A *question* is made up of one or more items.

Research from cognitive science, and in particular from the information-processing perspective, suggests that the presentation and content of item components affect respondents' comprehension, memory retrieval, data processing, and reporting abilities (Karabenick et al., 2007; Tourangeau, 1984). Studies such as those of Leighton and Gokiert (2008) and Gorin and Embretson (2006) have produced more specific details; both identified construct-irrelevant item features, which unduly increased cognitive loads and promoted inappropriate inferences. Leighton and Gokiert (2008) analysed science items designed for 13- and 16-year-old pupils and found them lacking in clarity, curriculum relevance, and technical quality. They traced these issues to three broad groups of features: (a) organisation and presentation of item components, such as poor graphic placement and lack of consistent document formatting; (b) textual information, such as ambiguous wording

or specialist terminology; and (c) item context, such as missing or irrelevant content in the stem. Gorin and Embretson (2006) found that complex (e.g., age-inappropriate vocabulary) or longer text passages, which require additional cognitive processing, might hinder textual encoding and the choice of appropriate response. Studies have consistently found that linguistic features such as negatively worded statements or the inclusion of onerous and unfamiliar wording, tax test-takers' working memory and present interferences to reasoning (see, eg., Clark, 1969; Kyllonen & Christal, 1990).

Established item-writing principles and assessment standards, drawn from guidelines and accrued knowledge of the research and assessment communities, steered the ideation and crafting of items in the PSRT. Table 4.19 provides the rationale for some of the adopted rules and guidelines that facilitated the assessment of cognitive abilities in this study.

Table 4.19

Adopted guidelines for item design and their rationale

Guideline	Rationale
Contexts are realistic and novel	To assess higher-order reasoning, contexts should neither be overly familiar (e.g., from text-books) nor require out-of-curriculum knowledge
Item interdependence, with no cueing to the answers of other items	To prevent "test-wise" behaviours
Clear and unambiguous formatting and layout	To prevent mis-reading or misinterpretations
Parsimonious use of words	To reduce unnecessary reading and processing load
Data in tables, graphs, and pictures are accurate, realistic, and salient	To prevent misinterpretations and activation of incorrect mental representations
Assess application of significant learning outcomes in sampling of test content	To ensure the demonstration of deep conceptual understanding
Require suitable number of steps in execution	To target age-appropriate processing abilities
Draw out important reasoning	To allow the demonstration of rule-based thinking such as inductive or deductive reasoning

Note. Adapted from Ahmed and Pollitt (2001); Downing (2005); Haladyna et al. (2002), and Stagnaro-Green and Downing (2006).

For the evaluation of the draft items, an expert panel, convened in the next phase to examine cognitive demand, also scrutinised the items' technical quality and content accuracy, as well as providing editorial advice. Revised items underwent a further review using empirical findings from pupils' performance in the pilot phase. The subsequent two chapters (Chapters 5 and 6) report the panel's initial findings and the results of the pilot phase. The next and final section of this chapter discusses the test framework.

4.5 Test Framework

The preceding discussion of the processes involved in the crafting of items began with the development of the construct and behavioural descriptors. This section describes the considerations involved in the assembly of the initial items into the Draft PSRT. It begins with the item characterisation approach to align the assessment objectives of the items to the scientific reasoning construct. The remainder of the section discusses the organisation of items in the instrument and the criteria for mark distribution.

4.5.1 Characterisation of Items

In the PSRT, items characterised by one of three levels of cognitive demand measured scientific reasoning by assessing pupils' scientific understanding in implementing specific practices. Figure 4.6 shows a three-dimensional grid model representing the knowledge, skills, and depth of assessment in every item.

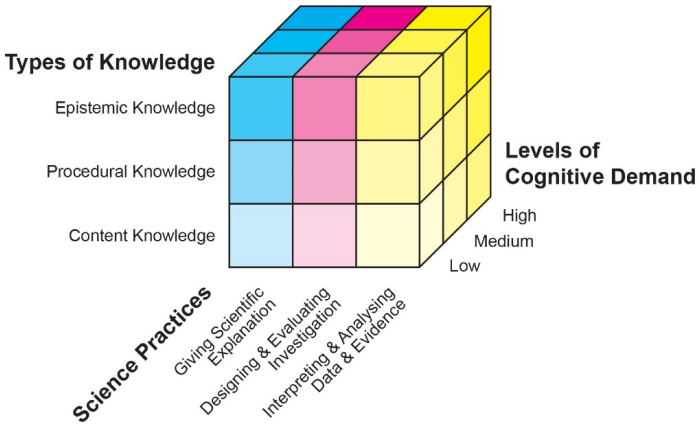


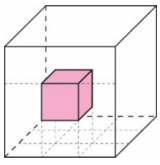
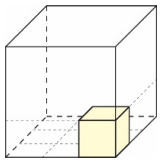
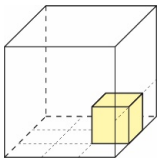
Figure 4.6. A grid model for characterising items along the axes of knowledge type, science practices, and levels of cognitive demand. Each of the three colours (blue, pink, and yellow) signifies a science practice.

As shown in Figure 4.6, the three axes of the cube permit 27 possible ways to describe an item. Cubes of the same colour share a common science practice, and the intensity of the colour relates to the level of demand. For instance, the darkest yellow cube at the top right-hand corner of the figure represents an item of high cognitive demand, which measures epistemic understanding in the practice of interpreting and analysing data and evidence. As indicated in Table 4.20, the grid model is the basis of the characterisation scheme of three items, which formed a question (Figure 4.5) in the Draft PSRT. The scheme identifies key features of an item: knowledge and practice being measured, its cognitive demand, the mark allocated, and its item format. Sub-practices are also identified. This study posits that these features function collectively to determine the difficulty of the item. See Appendix 6 for the item characterisation scheme of every item in the Draft PSRT.

The next subsection elaborates on the arrangement of items in the instrument for administration, and the final number and type of questions for inclusion in the test. The chapter concludes with the rationale for apportioning and assembling items with specific features to reflect the assessment focus in the TOS.

Table 4.20

Characterisation scheme of a three-item question shown in Figure 4.5

Feature	Item 1	Item 2	Item 3
Knowledge type	Procedural	Content	Content
Practice*	2	3	3
Sub-practice and level	2.1.1	3.1.2	3.1.3
Cognitive demand	Low	Low	Medium
Mark	1	1	1
Item format	SROS	SROS	CRSS
Grid-model placement			

*Note: *Practice 1 – giving scientific explanations, practice 2 – designing and evaluating investigative approaches, practice 3 – interpreting and analysing data and evidence.*

4.5.2 Assessment Structure

Five Test Booklets: A test must contain an adequate number of items to ensure the reliable and valid measurement of a complex construct composed of multiple aspects (AERA, APA, & NCME, 2014). In the validation of a new test, the minimum number of items required for achieving acceptable measurement precision typically cannot be completed by pupils in a single administration. Therefore, this study assembled the Draft PSRT into five paper-and-pencil test booklets, each of which contained subsets of items from the item pool, assembled according to multiple-matrix sampling principles (Sirotnik & Wellington, 1977). These broad-ranging principles guided the development and choice of booklet design, such as the breadth of the construct domain to sample, the number of reporting scales to include, and the testing duration (Gonzalez & Rutkowski, 2010).

Though more commonly deployed in large-scale educational assessments and surveys, this booklet design leveraged key advantages afforded by the deployment of Rasch modelling as the main data analytic method (van der Linden, Veldkamp, & Carlson, 2004). For instance, Rasch modelling facilitates representative testing of important aspects of scientific reasoning without compromising testing rigour. Also, there is flexible item assembly and construction, as booklets need not contain the same number of questions or amount to the same total marks for psychometric analysis. Moreover, pupils experience less testing fatigue and de-motivational effects, as they need not attempt all questions to enable the drawing of robust inferences about their abilities (Johnson, 1992). Chapter 6 provides a full discussion of the advantages conferred by Rasch modelling.

Matrix Sampling Design: The five-booklet assembly followed a balanced and incomplete design (van der Linden et al., 2004). In this design, there was an initial grouping of questions into 11 *question blocks,* each of which consisted of two to three unique questions. Every block appeared an equal number of times in sequential order within the five booklets, and each booklet contained a unique subset of the blocks. Other than the anchor block of questions, every test booklet contained four question blocks. Organising the questions thus facilitated the assignment of questions into various booklets, as well as control over the

representation of tested content across the booklets (Gonzalez & Rutkowski, 2010). Table 4.21 shows the organisation of question blocks across the booklets.

Table 4.21

Arrangement of anchor block and ten question blocks across five Draft PSRT booklets

Booklet	Anchor Block and Question Blocks										
1	Anchor	I	II	III	IV						
2	Anchor				IV	V	VI	VII			
3	Anchor	I				V			VIII	IX	
4	Anchor		II				VI		VIII		X
5	Anchor			III				VII		IX	X

Note. I to X are question blocks, each of which contained two to four questions.

Of the 11 question blocks in the PSRT, one block was common across every test booklet. This common block, also known as the *anchor block,* helped specify a common-person and common-item frame of reference for comparison with questions in the other booklets that were not administered to all pupils. The remaining ten question blocks rotated among the booklets, with each appearing in two of the five booklets. Linking the different booklets in this manner was necessary for comparison of item difficulties and pupils' abilities using Rasch modelling, as the type (e.g., item format and response options) and number of total items varied among booklets (van der Linden et al., 2004).

As shown in Table 4.22, the number of items in each block varied, from a low of five (i.e., anchor block and block I) to a high of 12 in block IV. There were two to four questions in each block. Each question could contain one or up to seven items. The items in a multi-item question can share the same stem or stimulus.

Table 4.22

Number of items in anchor block and ten question blocks across five Draft PSRT booklets

Booklet	Block											Total
	AN	I	II	III	IV	V	VI	VII	VIII	IX	X	
1												
No. of items	5	5	7	5	12							34
2												
No. of items	5				12	7	8	9				41
3												
No. of items	5	5				7			10	11		38
4												
No. of items	5		7				8		10		8	38
5												
No. of items	5			5				9		11	8	38

Note. AN stands for anchor block, I to X are question blocks. Each block had two to four questions.

In total, the number of items per booklet ranged from 34 to 41. As far as possible, efforts were made to prepare booklets with approximately equal test length, since each pupil attempted only one of the five booklets. This approach ensured all pupils could be equally engaged during the testing duration. The test was an hour long in accordance with the usual practice in school-administered class assessments in Singapore.

Item Format and Cognitive Demand: The Draft PSRT contained a total of 87 items across 32 questions. About 80% of these items were of the constructed-response format for stronger inferences about pupils' reasoning, understanding, and application of scientific concepts. See Table 4.23 for the distribution of items by format in the Draft PSRT.

Table 4.23

Distribution of the 87-item Draft PSRT by level of cognitive demand and item format

Cognitive Demand	Number of Items			Total
	SROS	CROS	CRSS	
Low	7	10	17	34
Medium	8	2	23	33
High	3	0	17	20
Total	18	12	57	87

As indicated in Table 4.23, about 20% of the items in the Draft PSRT were of "high" cognitive demand. Items of "low" and "medium" levels constituted about 40% of the test each. The reason for distributing items according to these levels was to elicit maximal participation and information from the mixed-ability pilot sample. See Appendix 7 for a detailed breakdown of the cognitive demand and item format for each item in the Draft PSRT.

Number of Dichotomous and Polytomous Items: Among the 87 items, 11 had three response options. These two-mark polytomous items were of the CRSS format. The remaining items were dichotomously-scored and carried one mark.

4.5.3 Tables of Specifications

Curriculum Outcomes: Items in the Draft PSRT tested disciplinary content from the topics specified in the primary science curriculum of Singapore (Ministry of Education Singapore, 2013). These topics also served to contextualise the assessment of procedural and epistemic knowledge. The first item in Figure 4.5 is an example of how the topical focus on heat exchange facilitated the demonstration of an understanding about fair testing – an important aspect of procedural knowledge. This study did not specify the extent of topical sampling in the Draft PSRT to emulate a similar practice adopted by national end-of-course primary assessments in Singapore. The rationale for the practice was to encourage schools to teach the whole of the curriculum and prevent "teaching-to-the-test" behaviours.

Efforts were made to ensure the sampling of as many topics as possible across the five booklets. Therefore, the items in the Draft PSRT tested topics from all five curricular themes of the primary science curriculum described in an earlier subsection of this chapter (see 4.3.1 and Table 4.3). About half of the marks in the PSRT tested the themes of energy and interactions, as observed in the summary in Table 4.24. These two themes covered more advanced topics such as energy conversion and animal adaptations in response to environment changes.

Table 4.24

Mark distribution in the Draft PSRT according to the disciplinary themes in the curriculum

	Disciplinary Themes in Curriculum					Total
	Diversity	Cycles	Systems	Energy	Interactions	
Number of Marks	8	28	12	16	34	98

See Appendix 8 for the mapping of each item onto the curricular themes. Besides steps taken for broad curricular coverage, other measures to ensure appropriate testing included checking there were no overlapping or trivial learning outcomes tested within each booklet.

Knowledge Types, Science Practices, and Cognitive Demand: The approximate balance of knowledge types and science practices in the Draft PSRT aligned with current emphasis in the national and school assessments for primary science in Singapore. For instance, the Draft PSRT assigned almost two-thirds of all available marks to the assessment of *content knowledge*. This testing approach for novice science learners is consistent with pedagogical principles promoting the initial acquisition of basic understanding of the material world as a “stepping stone” for future learning in science (Harlen, 2006). Correspondingly, the science practice of *giving scientific explanations* received almost half of the total marks in recognition of this fundamental skill.

The knowledge and science practice with the least assessment focus were *epistemic knowledge* and *designing and evaluating investigation*. This reduced focus accorded with the teaching emphasis in the Singaporean science curriculum: primary science pupils do not have enough understanding of the content to appreciate the epistemic requirements of science and have little exposure to investigative work. Table 4.25 presents the targeted mark distribution in the Draft PSRT instrument.

Table 4.25

Distribution of marks according to knowledge types and science practices

Knowledge Type	% of Marks	Science Practice	% of Marks
Content	60–80	Giving scientific explanation	40–50
Procedural	15–30	Designing and evaluating investigation	15–30
Epistemic	<10	Interpreting and analysing data and evidence	20–40
Total	100%	Total	100%

Table 4.26 provides a summary of the detailed TOS, which is shown in Appendix 9. This summary table outlines the overall mark distribution according to knowledge type, science practices, and cognitive levels.

Table 4.26

Mark distribution according to knowledge type, science practice, and cognitive demand

Knowledge Type	Cognitive Level	Marks in Science Practice			Total
		Giving scientific explanation	Designing and evaluating investigation	Interpreting and analysing data and evidence	
Content	Low	18	0	1	19
	Medium	7	0	4	11
	High	8	0	1	9
				Sub-total	39
Procedural	Low	1	5	5	11
	Medium	0	8	13	21
	High	0	0	7	7
				Sub-total	39
Epistemic	Low	2	1	2	5
	Medium	0	4	1	5
	High	0	7	3	10
				Sub-total	20
				Total	98

As shown in Table 4.26, there were few or no marks allocated across all levels of cognitive demand for certain combinations of knowledge types and science practices. For instance, no items assessed *content knowledge* in the practice of *designing and evaluating investigation*. This is because primary pupils have yet to acquire the necessary experimental skills to probe, contribute declarative knowledge, or further current understanding of the material world. In contrast, slightly more than one-third of the marks went to the testing of *content knowledge* in the practice of *giving scientific explanation*. This emphasis took reference from the curriculum focus on developing basic knowledge and skills. The detailed TOS in Appendix 9 provides information about the mark allocation at the item and question-block levels.

4.6 Conclusions

This chapter has outlined the initiation and collection of test content validity evidence from the process of item construction and test assembly. Documenting the test conceptualisation and design process comprehensively is an important aspect of construct validation (Markus & Smith, 2010). The first activity to develop an initial draft of the PSRT was to define and describe in detail the three knowledge types and competencies of the three science practices espoused in the scientific reasoning construct. This study used a combination of four approaches – theory, policy, empirical evidence, and expert opinion – to draw up these definitions and descriptors of competencies, which jointly served as evidence of domain definition (Sireci, 1998a, 1998b).

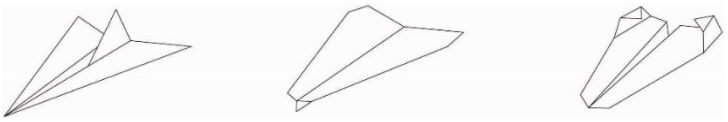
Guided by cognitive learning theories, the next step, following the development of the construct and behavioural descriptors, was to determine the four test design parameters: testing content, item format, cognitive demand, and item features. In turn, these parameters steered the characterisation and design of each item. For instance, the characterisation scheme of an item contained information about the targeted knowledge type, science practices (and levels of sub-practices), and cognitive demand, as well as the assigned item format and mark weighting. Assembly of the items into the Draft PSRT followed test provisions in the TOS to ensure there was a balanced and adequate coverage of the curriculum and construct. In the next phase (phase II), the assembled draft test was subjected to an expert panel review. The panel's findings form the focus of the next chapter.

Chapter 5 Expert Panel Review

5.1 Introduction

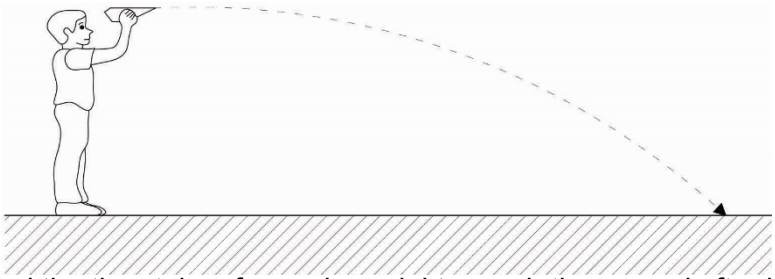
In this phase, an expert panel convened to review the scientific reasoning construct, including alignment with the curriculum, and the draft items designed to measure this construct. As an illustration of the review process, the draft item shown in Figure 5.1 received positive ratings from the panel for its alignment with the curriculum and the scientific reasoning construct. The panel also rated it as well written and endorsed it for possessing appropriate cognitive demand. However, a problem surfaced with the item, which led to it being excluded from subsequent versions of the test instrument.

Allan experimented with three different models of paper aeroplanes A, B and C.



Model A Model B Model C

He launched each model horizontally with the same force at the same height as shown below.



He recorded the time taken for each model to reach the ground after launching.

Model	A	B	C
Time (s)	3.4	4.9	2.1

Based on Allan's results, which one of the models, A, B, or C, should he use to design a toy aeroplane that can fly the furthest distance? Give a reason for your answer.

[Key: B. Longest time in air]

Figure 5.1. An item in the Draft PSRT that had a reasoning flaw detected during the review.

The item was unacceptable because, to arrive at the answer, test-takers had to assume that the duration the paper aeroplane was in the air was directly proportional to the distance it travelled. This was flawed reasoning, as there was no information about the relationship between the flight duration and the arc drawn in the question. A member of the panel detected this reasoning error during the review process.

This chapter discusses the role of the expert panel and the criteria they used to review the construct and the draft items. The discussion also covers the materials and procedures for gathering the panel's review and presents their judgements on the conceptual validity of the construct, including the strengths and weaknesses of the draft questions.

5.2 Validation by Expert Panel Review

The main purpose of this phase was to maximise the content validity of the Draft PSRT using the judgement of an expert panel (DeVellis, 2012). Almost all content validity investigations of tests conducted since the 1960s have involved subject-matter experts convened to provide independent and unbiased assessments (Newton & Shaw, 2014; Roach, Niebling, & Kurz, 2008). As explained in Chapter 3, the panel assembled in this study evaluated the test items according to Sireci's (1998a, 1998b) content validity framework, which comprises four test features: domain definition, domain representation, domain relevance, and appropriateness of the test-development process.

The *domain definition* operationalises the theoretical construct of scientific reasoning into clearly articulated, identifiable behaviours (Yang & Embretson, 2007). Defining the domain entailed describing in detail the components that made up the construct (Gorin, 2007). See Chapter 4 (section 4.3) for detailed descriptions of these components.

The second test feature is *domain representation*, which refers to the degree to which the items in the PSRT adequately measure the targeted scientific reasoning construct specified in the TOS (Martone & Sireci, 2009; Sireci & Faulkner-Bond, 2014). The third feature, *domain relevance,* is the alignment between the construct of scientific reasoning and the

individual items. Finally, the fourth feature, *appropriateness of the test-development process,* refers to measures adopted during the item- and test- development process to ensure the test produced valid and reliable inferences. In this study, the measures took reference from well-established item and test-development practices, such as item-writing and editorial guidelines (Downing & Haladyna, 1997; Webb, 2007). In total, the panel performed nine interrelated, complementary activities, and in the process gathered an important body of content validity evidence related to the test features. A summary of the relationship between the activities and the test features is shown in Table 5.1.

As part of established practices in content validity studies, members of the expert panel were carefully selected to ensure their qualifications, skills, and experience matched the area of testing (Haladyna & Rodriguez, 2013). The next section describes the profiles and relevant professional backgrounds of the panel convened for this study.

5.3 Expert Panel

The recommended size of an expert panel for a content validation study is between 10 and 20 people (Sireci & Faulkner-Bond, 2014). Accordingly, the expert panel in this study consisted of 18 members. Except for a single member, the experts currently live in Singapore and are familiar with the teaching, learning, and assessment of the primary science national curriculum in the country. Fifteen of them work in seven of the participating primary schools, and the remaining two have associations with the examination board[16] in Singapore.

[16] Singapore Examinations and Assessment Board (SEAB).

Table 5.1

Nine review activities conducted by the expert panel

No.	Review	Activity	Evidence of Content Validity
1	Construct	Experts checked the relevance, clarity, and appropriateness of the domain definition and descriptors.	Domain definition
2	Construct–item alignment	Experts checked the extent of correspondence between the domain definition and descriptors (reasoning practices) to the test and items.	Domain representation and domain relevance
3	Item writing	Experts checked written items against established practices and guidelines for violations.	Appropriateness of the test-development process
4	Cognitive demand	Experts checked items to see if the likely cognitive demand elicited matched the intended demand.	Domain representation and domain relevance
5	Content	Experts checked items for accuracy of scientific content, alignment to the curriculum, and accuracy of the assessed scientific knowledge, skills and practices.	Appropriateness of the test-development process
6	Editorial	Experts checked items for clarity, technical quality and any grammar, spelling, or punctuation errors.	Appropriateness of the test-development process
7	Sensitivity and fairness	Experts checked items for stereotyping of persons or insensitive use of language.	Appropriateness of the test-development process
8	Verification of the correct answer	Experts checked for accuracy of intended answers. Experts could suggest other possible answers.	Appropriateness of the test-development process
9	Answer justification	Experts analysed the rationale behind responses from the standpoint of test-takers.	Appropriateness of the test-development process

Note. Adapted from Haladyna (2004); Haladyna and Rodriguez (2013).

The group from schools was composed of teachers and senior curriculum or management personnel, such as a principal, two heads of department, and three senior teachers.[17] They consented to participate in the study during a series of face-to-face meetings held in Singapore from March to May 2016.

The sole member not from Singapore was an English examiner and item designer for key stage science tests, and a former assistant head teacher. She was invited to provide expert judgement based on her experience with the English primary curriculum. All members of the panel signed consent forms permitting the reporting of their professional profiles and reviews in this research.

Background and Professional Experiences: The expert panel was composed of 15 women and three men. At the time of the review, more than half of them were between 30 and 40 years old, and five were in their forties. The oldest member was 69 years old, and the two youngest were in their mid-twenties. All members of the panel held postgraduate teaching qualifications and are currently or had previously been science teachers. Four members – the two affiliated to the examination board, the principal, and one head of department – possessed Master's degrees in education and science. The English reviewer held a doctorate in education.

The average teaching experience of the 17 Singapore panel members was approximately nine years. Half of them reported more than 10 years' experience in constructing and vetting science questions for school and national examinations. The English reviewer reported 19 years of marking and coding science key stage tests at school and national levels. Table 5.2 summarises their teaching and test-evaluation experiences.

[17] Senior teachers have rich curriculum and teaching expertise. They serve as mentors and role models to less-experienced or new teachers (Ministry of Education Singapore, 2001)

Table 5.2

Professional profile of expert panel

Profile	Number of Experts	Number of Years
Masters or doctorate degrees in education	5	-
Average teaching experience	-	9.72 years
Test development and evaluation for examinations:		
Average experience	-	9.67 years
≥20 years	2	-
15–19 years	3	-
10–14 years	5	-
5–9 years	4	-
<5 years	5	-

5.4 Materials

Test Framework: This document, as described in Chapter 4 (section 4.5), outlined the operationalisation of the theoretical construct of scientific reasoning into items in the Draft PSRT. Specifically, it provided information about the design parameters of the items, as well as the criteria for organising and assembling questions in the test booklets. It also contained the TOS, the distribution of marks, and the number of items by item format.

Draft PSRT and Characterisation Scheme: This instrument contained a total of 87 items in 32 questions, assembled into five paper-and-pencil test booklets following a balanced and incomplete sampling matrix design. The rationale for using this design and the specifics involved in item-booklet organisation are detailed in Chapter 4 (subsection 4.5.2). Briefly, the design reduces threats to testing rigour, even though pupils did not answer all the questions (Johnson, 1992). Appendix 6 contains the items in the Draft PSRT and their characterisation scheme (as Table 1 in Appendix 6).

Proforma: The third and final document was the proforma, which was a survey to collect the judgement of the expert panel from their review of the other two documents. The format of the proforma was modelled on researcher recommendations that instruments for collecting and analysing judgements be as simple as possible (Haladyna, 2004; Rovinelli & Hambleton, 1977). Accordingly, the proforma had a simple bipolar four-point Likert-like scale, with eight statements to gather feedback about the construct and 13 statements about the test items (Oppenheim, 2005). The statements related to the nine review aspects highlighted in Table 5.1. The lower two points of the scale indicated disagreements, while the higher points signalled agreement ("1" – "strongly disagree", "2" – "disagree", "3" – "agree", "4" – "strongly agree").

Following the recommendation of Krosnick and Fabrigar (1997) to consider midpoints only if they are conceptually demanded, there was no midpoint in the scale because it was unclear whether a middle position in judgement was meaningful for this initial phase of the research. Krosnick and Fabrigar's (1997) recommendation was derived from mixed findings in research regarding the value of a midpoint for improving the reliability and validity of survey-based measures (Alwin, 2007; Krosnick, 1991).

Other than rating responses, the proforma also gathered qualitative feedback, such as comments and suggestions in blank spaces after every question. Rating scales are commonly used in content validity studies to measure raters' opinions, using declarative sentences arranged on a response continuum of agreement such as "strongly disagree" to "strongly agree" (DeVellis, 2012). The main reasons for the popularity of rating scales are ease of construction and the ability to capture differences in attitudinal traits or attributes quickly (Oppenheim, 2005). Tables 5.3 and 5.4 show the statements in the proforma, and the specific evidence elicited by each, according to Sireci's (1998a, 1998b) four features of content validity.

Table 5.3

Eight statements about construct and relationship to specific evidence of content validity

No.	Statement About Scientific Reasoning Construct	Evidence of Content Validity
1	Scientific reasoning and its scales (knowledge type, practice, and sub-practice) represent an important domain of content *of* and *about* science.	Relevance of domain to science
2	Scientific reasoning and its scales specify important reasoning skills and processes that pupils in primary schools should acquire.	Relevance of domain to curriculum
3	Each sub-scale contributes towards the overall notion of scientific reasoning.	Completeness of domain description
4	Scientific reasoning and its scales are clearly written.	Clarity of descriptors about domain
5	The developmental progression outlined in the descriptors of each of the three practices is logical and reasonable.	Appropriateness of domain description
6	Where appropriate, the reasoning skills and processes in the Draft PSRT are taught to primary pupils in the curriculum of your country.	Opportunities to learn the domain in school
7	Where appropriate, the reasoning skills and processes in the Draft PSRT can be learnt by primary pupils.	Age-appropriateness of domain
8	The distribution of knowledge types and science practices assessed in the Draft PSRT broadly follow the national assessments of your country.	Appropriateness of the test-development process

Table 5.4

Thirteen statements about items and relationship to specific evidence of content validity

No.	Statement About Items in Draft PSRT	Evidence of Content Validity
1	The concept(s) assessed represent important content *of* and *about* science.	Appropriateness of the test-development process
2	The science content in the question is scientifically accurate.	Appropriateness of the test-development process
3	The concept(s) assessed are aligned to the curriculum.	Appropriateness of the test-development process
4	The sub-scales have been correctly characterised in the question.	Domain representation and domain relevance
5	The level of cognitive demand is identified correctly.	Domain representation and domain relevance
6-7	<u>Item Stimulus</u>	
	(a) Contains only words that are essential for responding.	Appropriateness of the test-development process
	(b) Vocabulary and sentence structure are grade-appropriate.	
8-10	<u>Item Stem</u>	
	(a) Text is minimal in length, written as plainly as possible.	Appropriateness of the test-development process
	(b) Vocabulary and sentence structure are grade-appropriate.	
	(c) Target content is evident from stem.	
11-13	<u>Visuals (pictures, charts and graphs)</u>	
	(a) Visual(s) used are necessary for responding.	Appropriateness of the test-development process
	(b) Visual(s) clearly depict or provide the intended information and are as simple as possible.	
	(c) Visual(s) contain only text that is essential for responding.	

5.5 Procedure

This section discusses the preparatory and closure procedures enacted before and after the review, respectively, for four months from June to September 2016. During this period, all meetings with the expert panel took place through Skype, as I was based in Oxford, UK. Meetings occurred either on an individual basis or in groups.

5.5.1 Before and During Review

Before the August 2016 review, every member of the expert panel met with me at least once for briefings about the process of documentation and review. As the members from the seven schools were not familiar with the review procedures of new tests, a preparatory package consisting of the draft domain definition and descriptions, as well as two exemplar items, was sent in advance via mail in June 2016. Discussions about the materials in the package took place three to four weeks later in meetings with every school. The discussions largely consisted of clarifications about the definition and construct descriptors, and the alignment between the descriptors and the items. Feedback from the school members during these meetings also helped refine the descriptors.

Subsequently, all members received the test review package consisting of the test framework with the finalised descriptors, proforma, and the Draft PSRT by email in August 2016. The instructions included the duration of the review period and the procedure (arrangement of online meetings or email) to take should members have any queries or need clarification. There was no issuance of the mark schemes in this phase, as the review was solely designed to establish potential problems with the items, such as poor phrasing or unsuitable stimuli. Also, as this phase occurred at approximately the same time as the pilot study (phase III, Chapter 6), pupils' responses from the pilot augmented feedback from the expert panel about the items. Consolidating the responses from both phases provided the range of acceptable and unacceptable answers as evidence of correct and incorrect understanding, which helped with the diagnosis of problematic items (Pollitt & Ahmed, 2008).

At the end of the one-month review period, members submitted their completed proformas to me by email. During the review period, there were no queries from the expert panel.

5.5.2 After Review

Panel members[18] met with me, either individually or in groups, after submitting their proformas. During these closure meetings, which were recorded, participants clarified their written inputs and shared their final thoughts, remarks, and ideas with me. I also took the opportunity to thank the reviewers for their participation.

5.6 Data Analysis

This section discusses the data-analytical methods used to organise, summarise, and interpret the content validity evidence collected in the proforma. The section also outlines how the qualitative information collected during closure meetings with the expert panel augmented the interpretations of the evidence from the proforma.

5.6.1 Rating Scales in the Proforma

There are various methods of analysing and summarising data from rating scales in content validity investigations (Webb, 2007). Sireci and Faulkner-Bond (2014) recommended that besides providing summarised descriptions of domain representation (e.g., TOS), there should also be reports of "congruence/alignment statistics" (p. 103). These congruence statistics can take various forms such as proportion of agreement or outputs from statistical testing. For example, Popham (1992) reviewed 32 studies that evaluated content in teacher licensure tests and reported two common ways of summarising content ratings. One way is to compute the majority index, which is the proportion of people on the panel who endorse the survey statement after comparison with the criterion (e.g. alignment to curriculum). The other way is to calculate the arithmetic average of the ratings. Based on the endorsement values adopted in each of the 32 studies, Popham suggested that survey statements should

[18] All but one reviewer participated in the closure meeting. It has not been logistically possible to organise a meeting with this reviewer.

achieve agreement rates of 75–80% to be considered congruent with the criteria. Popham added that in contrast to the majority index, the arithmetic averages of ratings are more sensitive and conservative measures of approval from the responses of reviewers. Averaging ratings is a common approach in studies that evaluate both broad alignment between the domain and test specifications, and fine-grained associations between specific test elements (e.g. cognitive demand) and the domain (Sireci & Faulkner-Bond, 2014; Sireci & Geisinger, 1995).

Other statistical methods for summarising content ratings include computing the content validity index (CVI) and Aiken's item content-relevance index (Aiken, 1980; Polit & Beck, 2006). The CVI is a variant of the majority index; it is calculated by first dichotomising the four-point nominal scale into "agree" and "disagree", and then noting the proportion of reviewers rating each survey statement as acceptable out of the total number of reviewers. This index is a common indicator in studies of many disciplines, such as nursing and education (Polit & Beck, 2006). There are two types of CVIs: item-level index and the mean index of a group of items measuring the same content area (Lynn, 1986). Regardless of type, researchers have advocated minimal levels of between 0.8 and 0.9 as indicators of congruity, depending on the purpose of the research (Davis, 1992). Though quick to calculate and easy to understand, this method of collapsing categories into only two also removes potentially relevant information (Polit & Beck, 2006). Aiken's index varies between zero and one, and besides indicating the proportion of reviewers rating survey statements above the midpoint of the rating scale, it can be used to test specific hypotheses concerning the mean values of the population's ratings. Unfortunately, the procedure for computing this index is technically complex (Penfield & Giacobbi Jr, 2004).

Ratings About the Construct: Based on the above discussion, this study reported the means and standard deviations of the ratings for each statement about the construct and tabulated the frequency of these specific ratings. An additional procedure for the mean ratings about the construct was carried out to determine their accuracy as estimates of the true population using confidence intervals. This additional procedure – the score method –

also permitted the testing of directional and non-directional hypotheses concerning the value of the unknown population mean (Penfield, 2003). Appendix 10 explains the score method and its computation. The CVI could not be applied to ratings related to the construct, as the collected responses from the panel were all positive.

Ratings About the Items: In contrast to the ratings about the construct, this study did not summarise the ratings of statements about the Draft PSRT items with the measures of central tendency, such as the mean and standard deviation. Information about the mean and mode are redundant, as decisions about retaining, amending, or removing test items required an overall assessment of the items. Instead, the study computed the CVIs for ratings of statements about the items. As an indicator of inter-rater agreement, the CVI was one of many factors involved in item retention and modification decisions. Also, some of the information provided by the panel had more primacy in these decisions. For instance, as demonstrated through the rejected item in the introduction to this chapter, a single input about the doubtfulness of the science tested captured in an individual rating statement outweighed the collective positive feedback about the item's quality and relevance to the curriculum. Other information critical for decision-making included alignment with the curriculum (as a measure to show that pupils knew the tested content and reasoning), and appropriate item quality (to ensure there were no construct-irrelevant biases). Table 5.5 shows the statistics to summarise the panel members' ratings.

Table 5.5

Statistics summarising the panel members' quantitative ratings in the proforma

Reported Descriptive and Quantitative Statistics	
Ratings in Proforma About the Construct	Ratings in Proforma About the Items
Means Standard deviations Confidence intervals (from score method) Frequency tabulation	CVI

The treatment of scores from the ratings outlined above mirrored the recommendations of Allen and Seaman (2007) and Carifio and Perla (2007); that is, to take a sensible and appropriate approach to analysis and interpretation.[19] For instance, the means and standard deviations provided a useful way of describing the "typical" opinion of a reviewer but were inadequate for drawing firm conclusions about the underlying construct measured in the rating statement. A final analysis conducted with the ratings from the proforma was the cross-over strategy of data reduction. As mentioned in the methodology chapter (subsection 3.2.2 and Table 3.3), this referred to the thematic analysis of quantitative data collected from the responses of the panel. Braun and Clarke (2006) defined thematic analysis as, "a method for identifying, analysing and reporting patterns (themes) within data" (p. 79). Accordingly, the analysis and grouping of responses about the construct and the items targeted the production of inferences about the various aspects of content validity.

5.6.2 Qualitative Feedback from Proforma and Closure Meetings

Qualitative data in the proforma consisted largely of comments, explications, or clarifications by the panel members about specific survey statements or ratings that they awarded. Together with the transcribed data from the closure meetings, the information provided further understanding of the panel members' opinions about the construct and its items.

In summary, the study used a variety of data-analytical methods to summarise and interpret the expert panel ratings recorded in the proforma. Collectively, these methods ensured coherent and efficient processing of the ratings, as well as greater integration of qualitative and quantitative data in this phase for stronger inferences about the content validity of the Draft PSRT. The next section discusses the findings of this phase followed by a discussion of how panel members' input clarified the construct and helped identify problematic items.

[19] This approach acknowledges concerns about the scale nature of these scores and the type of analyses that should be performed. Still, some studies have shown that Likert scales can indeed be analysed effectively as interval scales (Maurer & Pierce, 1998; Vickers, 1999).

5.7 Results and Discussion

This section reports and discusses the evidence of content validity gathered from the panel members' quantitative and qualitative responses in the proforma and during the closure meetings.

5.7.1 Content Validity Evidence of the Scientific Reasoning Construct

Quantitative Evidence: All 18 members rated the eight statements about the scientific reasoning construct positively. More than half of the panel members gave strong endorsement to six of the eight statements. The statement that received the strongest agreement was about the relevance of the construct in science. The two statements with the weakest ratings pertained to the clarity of the construct's descriptors and the appropriateness of the proposed progression in reasoning practices. See Figure 5.2 (next page) for a summary of the ratings awarded by the panel.

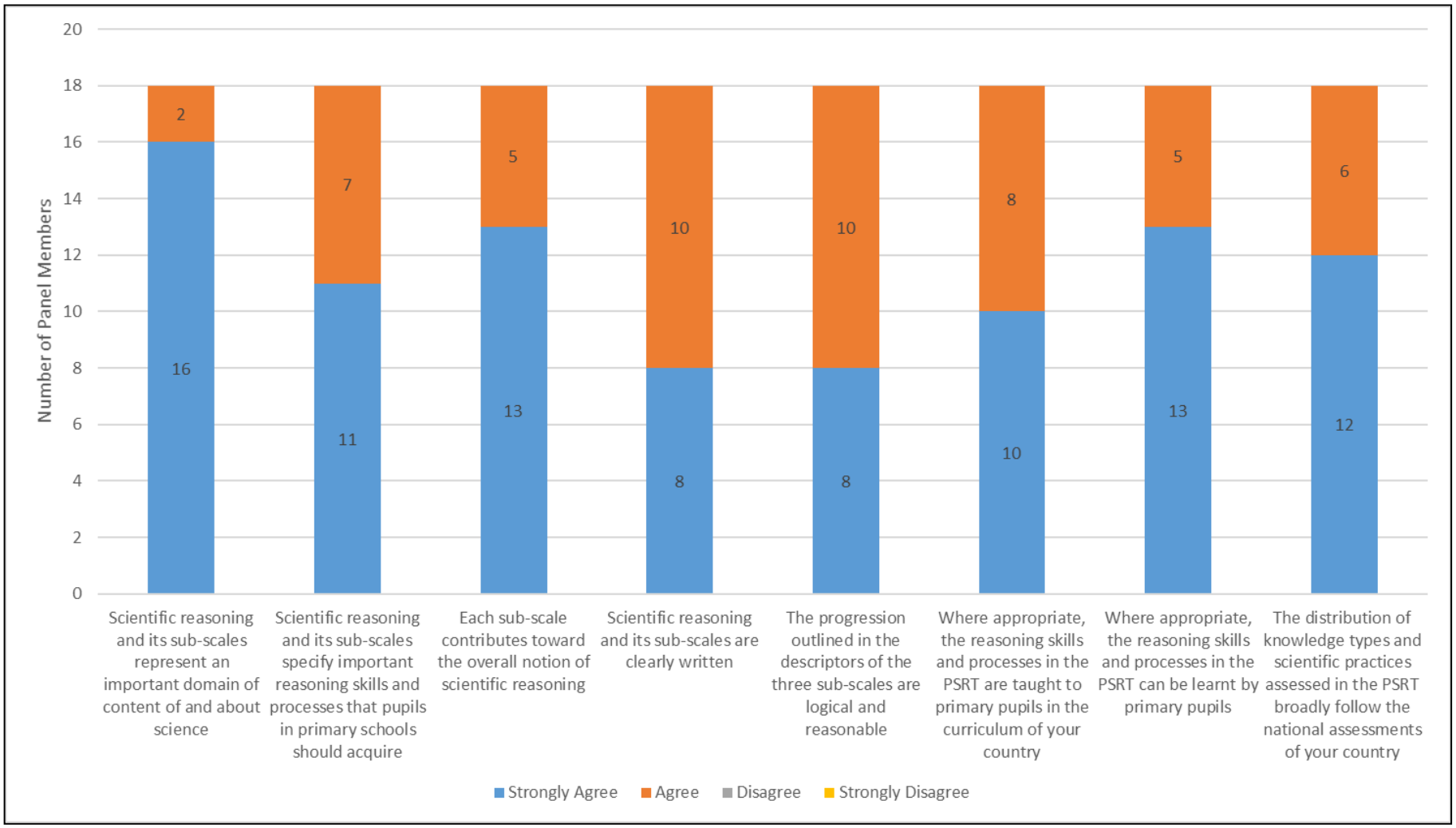


Figure 5.2. Extent to which members of the panel agreed with the statements about the scientific reasoning construct.

As shown in Table 5.6, the mean and standard deviation showed the smallest spread of scores in the first statement, which had the most agreement, and the widest spread in the two statements with the least endorsement. In general, these results showed that the panel highly agreed with the statements.

Table 5.6

Means, standard deviations, and 95% score confidence intervals for eight statements about the scientific reasoning construct

	Statement	Mean	SD	95% Confidence Index	
				Lower Limit	Upper Limit
1	Scientific reasoning and its scales (knowledge type, practice, and sub-practice) represent an important domain of content *of* and *about* science.	3.89	0.32	3.44	4.34
2	Scientific reasoning and its scales specify important reasoning skills and processes that pupils in primary schools should acquire.	3.61	0.50	3.15	4.07
3	Each scale contributes towards the overall notion of scientific reasoning.	3.72	0.46	3.26	4.18
4	Scientific reasoning and its scales are clearly written.	3.44	0.51	2.98	3.91
5	The developmental progression outlined in the descriptors of each of the three practices is logical and reasonable.	3.44	0.51	2.98	3.91
6	Where appropriate, the reasoning skills and processes in the Draft PSRT are taught to primary pupils in the curriculum of your country.	3.56	0.51	3.09	4.02
7	Where appropriate, the reasoning skills and processes in the Draft PSRT can be learnt by primary pupils.	3.72	0.46	3.26	4.18
8	The distribution of knowledge types and science practices assessed in the Draft PSRT broadly follow the national assessments of your country.	3.67	0.49	3.21	4.13

To assess the validity of the panel's good overall endorsement, this study adopted a criterion-based validity test using hypothesis testing as the basis for acceptance (Penfield, 2003; Penfield & Miller, 2004). The 95% confidence intervals of the mean rating to the eight statements in Table 5.6 were thus computed. The criterion used is the midpoint between the response categories of 1 and 4, which is 2.5. By comparing with the criterion, the results indicate that the lower limits of the confidence intervals were greater than the null hypothesis value of 2.5. Therefore, there was rejection of the null hypothesis in favour of the alternative directional hypothesis that $\mu > 2.5$, using the specified Type 1 error rate of $\alpha = 0.05$. Overall, the favourable endorsement of all eight statements was statistically significant.

Qualitative Evidence: Qualitative feedback about each of the eight statements further elucidated the panel members' opinions. Unsurprisingly, statement 1 garnered just two comments, affirming the relevance of the construct in science. Statement 2 received comments representing divergent views; two teacher members believed that some of the scientific reasoning knowledge and practices could not be attained by academically weaker pupils, or even by primary pupils, while the principal said the existing curriculum covered substantial aspects of the knowledge and practices. There were no comments on the third statement.

Statement 4, which sought to elicit views about the clarity of the descriptors, received the most remarks and suggestions. Though panel members broadly agreed that the descriptors were clear, they highlighted specific descriptors that they found difficult to interpret. For instance, four members the panel identified the descriptors for sub-practice 1.2, under the broader practice of *giving scientific explanation,* as vague (see Appendix 5 for the descriptor tables of the seven sub-practices). One of two other commonalities among the feedback from the panel about statement 4 was the difficulty in distinguishing between procedural and epistemic aspects of scientific knowledge. One reviewer attributed the confusion between forms of knowledge to the wording in the test framework document. The reviewers also commented on the formatting and presentation of the descriptors of sub-practices, which were considered lengthy. One reviewer was in favour of the *Look for* section in the

descriptor tables, as it provided “further explanation on the learning outcomes”. Another reviewer suggested providing an overview of all the sub-practices for ease of reference.

Statement 5 elicited concerns about the reasonableness of the proposed progression of manifested behaviours as pupils advanced in their capabilities in the various reasoning sub-practices. While comments on statement 5 were positive overall and indicated broad agreement, many reviewers expressed the need for greater specificity and clearer delineation between stages in some of the sub-practices. As one reviewer explained, such enhancements would “improve coherence and usability”.

Statement 6 sought views about whether the reasoning skills and processes measured in the test were in the curriculum. Feedback on this statement was notable for the wide range of opinions expressed. A significant number of panel members opined that the curriculum and its learning outcomes did not outline the reasoning skills and practices espoused in the Draft PSRT, and that pupils were instead implicitly taught these skills and practices. A few added that these reasoning skills were not consistently taught across Singapore schools. The English reviewer highlighted that the teacher often serves as a key factor in the acquisition of these skills and practices. She explained: “there may still be the desire to ensure content is covered and this is then at the expense of the skills and reasoning”.

In addition, feedback from the Singapore reviewers expressed a preference for reasoning skills and practices to be articulated more clearly in the curriculum document. As another reviewer explained, this articulation could help teachers impact on the development of “pupils’ capability to self-regulate and learn independently”. Notably, the lack of explicit guidelines in Singapore’s primary curriculum about types of reasoning skills and practices in use resonates with comments from the last statement, as will be described shortly.

Feedback to statement 7 acknowledged that primary pupils could learn the reasoning skills and processes in the PSRT, given the right contexts and opportunities. However, one reviewer said teaching and learning epistemic knowledge may prove challenging, as many teachers and pupils are “fixated on ‘one right answer’ to questions”. For this reason, he

approved the relatively low proportion of questions assessing epistemic knowledge in the Draft PSRT.

The eighth and final statement related to panel members' opinions between the weighting of the scales in the Draft PSRT versus the national assessments. Notably, the English reviewer and one Singapore panel member stated opposing views. The former agreed that there is a general alignment between the distribution of knowledge type and practices assessed in the Draft PSRT and England's national assessments, elaborating that: "the frameworks in primary school weave the procedural and epistemic through the content. There is a mixture of low cognitive demand of some content and the higher cognitive demand is often where the epistemic knowledge types are assessed." In contrast, the Singapore panel member felt that national assessments in Singapore rarely test for epistemic understanding. He added that there is limited assessment of procedural and experimental design, which in most instances is constrained to about 30% of the assessments. However, his judgement that there is a misalignment seemed isolated, as another member opined that the Draft PSRT questions were modelled largely after the style of Singapore's national examinations.

Discussion of Results Using Cross-over Analysis: Augmented by the above qualitative insights from the panel about the construct, further analysis framed the quantitative ratings thematically to Sireci's (1998a) four test features of content validity. The feedback about the construct addressed two of these four features: domain description and appropriateness of the test-development processes. The first five statements addressed the relevance, completeness, appropriateness, and clarity of the domain description on scientific reasoning. It was clear that all members agreed that the construct was relevant to the learning of science and the primary science curriculum. They also generally agreed that the descriptors of the knowledge types and reasoning practices were comprehensive, helpful, and explicit, and that the proposed progression of the reasoning practices was logical and reasonable. However, the relatively lower ratings and comments on statements 4 and 5 indicated that some of the descriptors needed modification or reconsideration. Reviewers cited conflicting

terminology, discrepancies between the descriptors and their corresponding examples, and conceptual overlaps between descriptions as some reasons for amendments.

Statements 6 and 7 solicited panel members' views about opportunities for school exposure, as well as the age-appropriateness of the knowledge types and reasoning practices outlined in the domain description. Lower ratings and the accompanying qualitative feedback from the Singapore panel suggested that opportunities were variable across schools and depended heavily on teachers' experience and knowledge. Overall, there was consensus that primary pupils can learn the espoused knowledge and practices, and that a deeper understanding of epistemic notions (e.g., science is not a positivistic activity) is required.

The eighth statement checked panel members' opinions about the weighting of knowledge types and practices between the Draft PSRT and national assessments. Their ratings provided validity evidence for the appropriateness of the testing approach. Feedback was somewhat mixed. Other than a single comment that the national assessments in Singapore seldom assess epistemic knowledge, the ratings and feedback from the remaining Singapore panel members suggested the weighting was broadly aligned. From the perspective of the English reviewer, there was alignment in weighting between English national assessments and the Draft PSRT. The next subsection discusses the findings from panel members' inputs about the individual items in the Draft PSRT.

5.7.2 Content Validity Evidence of the Draft PSRT Items

Quantitative Evidence: Panel members evaluated each of the 87 draft items on 13 criteria or statements. Calculation of the CVI began by collapsing the four-point ratings into two categories of "agree" and "disagree", and then noting the proportion of reviewers who agreed with each statement about the items. In a practice aligned with Popham's (1992) recommendation to classify items with a CVI above 80%[20] as congruent with the criteria, ratings below this value were flagged for attention in the tables. Table 5.7 shows the CVIs

[20] As a gauge, CVIs of 0.83 and below meant that three or more panel members disagreed.

of the first seven items in the test, arranged by question blocks. The CVIs of all the items in the test are provided in Appendix 11. Analysis of the CVI across the items showed that the focus of panel members' disagreements centred on the extent of correspondence between the targeted construct of scientific reasoning in the item characterisation schemes and the test items. Specifically, reviewers disagreed with the assigned scales, comprising knowledge types and science practices, as well as the cognitive demand of the items. In all, 26 items were flagged (from an item pool of 87) for receiving a CVI of below 0.8 in at least one rating.

Table 5.7

CVIs of panel's feedback on 13 statements about the first seven items in the Draft PSRT

Statement	CVI of First Seven Items* in Draft PSRT						
	1(a)	1(b)	2(a)	2(b)	3	4	5
1 The concept(s) assessed represent important content of and about science	1	1	1	0.94	0.94	1	0.94
2 The science content in the question is scientifically accurate	1	0.94	1	1	1	1	1
3 The concept(s) assessed are aligned to the curriculum	1	1	1	0.94	0.89	1	1
4 The sub-scales have been correctly characterised in the question	0.83	0.94	0.78	0.67	0.78	0.94	1
5 The level of cognitive demand is identified correctly	1	1	1	0.89	1	0.78	0.94
Item Stimulus							
6(a) Contains only words that are essential for responding	1	1	1	1	1	1	1
6(b) Vocabulary and sentence structure are grade-appropriate	0.94	0.94	1	1	1	1	1
Item Stem							
7(a) Text is minimal in length, written as plainly as possible	0.94	0.94	1	1	1	1	1
7(b) Vocabulary and sentence structure are grade-appropriate	1	1	1	1	1	1	1
7(c) Target content is evident from stem	1	1	1	1	1	1	1
Visuals (pictures, charts and graphs)							
8(a) Visual(s) used are necessary for responding	1	1	0.83	1	1	1	1
8(b) Visual(s) clearly depict or provide the intended information and are as simple as possible	1	1	1	0.83	1	0.94	1
8(c) Visual(s) contain only text that is essential for responding	1	1	0.94	1	1	1	1

*Note. *These first seven items are in the anchor block. Cells highlighted in red and orange indicate CVIs of below 0.8.*

Indices of Domain Representation and Domain Relevance: Domain representation is the extent of correspondence between *all* the items on the Draft PSRT and the targeted aspect of the construct (Sireci & Faulkner-Bond, 2014). Following the example shown in Sireci (1998b), the mean CVI from all the items in statements 4 and 5 are the indices of domain representation for the sub-scales of knowledge types and practices and cognitive demand respectively. As shown in Tables 5.8 and 5.9, these indices were 0.92 and 0.90, which meant that on average, for every item, about 17 out of 18 panel members agreed that the scales (knowledge type and practices) and cognitive demand, were aligned to the targeted construct. Both indices met the 0.8 and 0.9 congruency criteria recommended by Popham (1992) and other researchers (e.g., Davis, 1992) respectively.

Table 5.8

Number of items by CVI obtained in statement 4

	CVI of 87 Items on Statement 4 (Scales)							Mean CVI
	1*	0.94*	0.89*	0.83	0.78	0.67	0.56	0.92
No. of items	44	22	4	3	10	2	1	

*Note. *Items with CVIs at and above 0.89 indicate an acceptable level of alignment to the targeted construct.*

Table 5.9

Number of items by CVI obtained in statement 5

	CVI of 87 Items on Statement 5 (Cognitive Demand)									Mean CVI
	1*	0.94*	0.89*	0.83	0.78	0.72	0.61	0.56	0.39	0.90
No. of items	38	20	7	9	4	3	2	1	1	

*Note. *Items with CVIs at and above 0.89 indicate an acceptable level of alignment to the targeted demand.*

Domain relevance addresses the degree to which *each* item on the PSRT matches the descriptions in the TOS (Sireci & Faulkner-Bond, 2014). This study targeted a pre-defined level of around 0.9 to evaluate the relevance of individual draft items. This higher standard was set to highlight more items for scrutiny at this preliminary stage of instrument development. However, striving for 100% agreement may be excessively demanding, because there were many reviewers in this initial validation study and disagreements were therefore expected (Polit & Beck, 2006). Based on the pre-defined level, the proportion of items out of the total with CVIs at and above 0.89 were the indices of domain relevance for scales and cognitive demand. As indicated in Tables 5.8 and 5.9, the indices of domain relevance were 0.80 for scales and 0.75 for cognitive demand. These indices meant that only 80% and 75% of the 87 items respectively were found to have achieved acceptable levels of alignment to the scales (knowledge type and practices) and cognitive demand. The lower indices were unsurprising, as members had rated many items on statements 4 and 5 less favourably.

Like panel members' earlier feedback on the scientific reasoning construct, their qualitative feedback on the items explained their quantitative ratings. The next section summarises the key issues gathered from the panel's qualitative inputs, as collected via the proforma and closure meetings.

Qualitative Evidence: Most of the comments in the proforma were elaborations about poor ratings on the items. Based on the earlier discussion about the items' ratings, it was unsurprising that these remarks centred on the characterisation of the scales and cognitive demand. Other concerns reflected in the proforma included accuracy of the tested science content, clarity of phrasing and choice of vocabulary. See Appendix 12 for a summary of the key issues highlighted by the expert panel at the item level. During the closure meetings, many panel members reiterated these same issues. At this juncture, they shared four other issues not reflected in their comments in the proformas.

First, two panel members from a school discussed how the choice of an item format impacted on the presentation of evidence for assessing complex reasoning. They referenced the few stand-alone multiple-choice items in the draft items and commented that pupils' selection of the key/options was not an adequate basis for understanding their thinking. In contrast, items with formats that required pupils to explain their choices and answers provided avenues for examining their reasoning. Second, in response to two draft items that assessed the concept of volume displacement, the English reviewer explained that Piagetian and related concepts such as mass conservation and seriation, are not in England's primary science curriculum. She was aware that the Singapore primary science curriculum emphasises these concepts earlier than England on the established basis that the ability to conserve is one of the signposts of progression into formal Piagetian stages.

She shared her view that English primary pupils are likely to learn these concepts in mathematics as extension activities (e.g. how to work out the volume of an object). Third, some panel members expressed reservations about some items as the tested learning outcomes received heavy emphasis in classroom teaching. They were uncertain that performance on these items could give accurate insights about pupils' actual reasoning. Items flagged by these members included those testing the COV strategy and concepts such as the prey-predator relationship.

Lastly, closure meetings with panel members from schools brought up matters related to curriculum organisation, teaching, and pedagogical implications, as well as possible impact on assessment practices. Overall, they expressed that the research had enabled them to appreciate the notion of scientific reasoning. Besides discussions about how curriculum time could be re-organised to expedite teaching and learning of the knowledge types and practices, some panel members also shared how the research had an immediate impact on their assessment practices with pupils. In summary, qualitative evidence from the proforma and closure meetings substantiated the ratings given by the panel. The next subsection of this chapter discusses the findings from a cross-over analysis conducted on the quantitative ratings of the items.

Discussion of Results Using Cross-over Analysis: The quantitative ratings addressed three of Sireci's test features of content validity (see Table 5.4, p. 138 of this chapter). Ratings from six of the 13 statements (1, 2, 3, 6, 7, and 8) elicited evidence about the appropriateness of the test-development processes. Specifically, the first three statements checked the relevance, accuracy, and curricular alignment of the scientific concepts assessed in each Draft PSRT item. Most of the items received endorsement for the relevance of the scientific concepts tested in the items. There were instances where some panel members expressed doubts over the accuracy and the curricular alignment of the concepts. However, no more than two reviewers disagreed with each of the first three statements about the scientific concepts for affected items. Likewise, disagreements about the technical quality did not exceed more than three reviewers for identified items in statements 6, 7, and 8.

Statements 4 and 5 collected evidence of domain representation and domain relevance. As discussed earlier in this subsection, the indices of domain representation suggested that on average, for every item, 17 of the 18 panel members agreed that its sub-scales and cognitive demand aligned to the suggested characterisation schemes in the TOS. However, based on the CVI criteria set for acceptable levels of alignment to the scales (knowledge type and practices) and cognitive demand, only 80% and 75% of the 87 items achieved these levels respectively. From the overall evidence presented in the findings thus far, there were two probable reasons why higher levels of domain relevance were not achieved in the Draft PSRT. First, evidence about the appropriateness and clarity of the construct suggested there were issues with the domain description that had, in turn, imposed alignment issues for the reviewers at the item level. The second area where misalignment between item and targeted construct could have occurred was at the item level. Items could have been poorly designed, assessed inaccurately, or included out-of-curriculum scientific concepts, knowledge types, or practices. The next and final subsection explains how the content validity evidence from the expert panel was used to determine or consider items in the Draft PSRT for inclusion or removal from the final test.

5.7.3 Item Outcomes Based on Validity Evidence from the Expert Panel

Three possible item outcomes arose from the feedback of the expert panel. Draft items could receive recommendations for amendments or retention (without changes). They could also be deleted from the test. The criteria for deciding the outcome of an item depended on its panel-awarded ratings and comments, as will be described in this subsection.

Item Deletion: Four types of items were deleted from the Draft PSRT based on feedback from the panel. The first type was items that received feedback expressing doubts about the scientific concepts present in the stimulus or stem. This issue affected five items, four of which were single-mark dichotomous items. Three stand-alone, multiple-choice items were also deleted due to concerns raised by some reviewers that the item format was unable to provide insights into pupils' reasoning. An additional three items from a question were taken out because the English panel member flagged that primary pupils in England were unfamiliar with the scientific concepts tested. The removal of these items came about as the study then intended to collect evidence of item invariance (items functioned similarly across different pupil samples) by conducting the test in English schools.[21]

Finally, 13 items from five questions were removed due to concerns that they could produce limited information about how pupils reasoned. This was because the assessment objectives (e.g., COV strategy) were commonly tested and thus received undue instructional emphasis from "teaching to the test" practices in primary science classrooms. In total, 23 items, which represented about 27% of the Draft PSRT, were removed based on feedback from the panel to reduce threats arising from construct-irrelevant variance (e.g., testing out-of-curriculum materials), and to ensure maximum inferences from content validity evidence (e.g., using appropriate item format). After the removal of these items, the remaining 64 items in the test were considered for either retention or modification.

[21] The research plan to collect data from English pupils was dropped because the study could not obtain permission for access to the selected schools in time.

Item Retention or Amendments: Generally, among the remaining draft items, those that received positive endorsement from the panel were retained. Items that received either poor ratings and/or negative comments were flagged for amendments. Specific recommendations for modifications depended on the nature of the issues highlighted by the ratings and comments. These recommendations were further shaped by suggestions for item improvement if these were provided by the panel members. Some of these suggestions included changes to the items' designated features in the characterisation schemes (e.g., cognitive levels), proposals for editorial changes, enhancement to visual materials (e.g., graphics), and inclusion/exclusion of marking points. For instance, nine of the remaining items were flagged for possible amendments as a significant proportion of the panel members disagreed over the match between the intended and actual cognitive demand. Changes to adjust the assessment demand (e.g., by simplifying the task) in some of these nine items also took on board other suggestions for improvement from the panel members. See Appendix 12 for the item-specific outcomes of the Draft PSRT.

Given the validation approach of this mixed-methods study, a point to note at this juncture was that recommendations related to item retention and amendment remained preliminary, subject to other validity evidence from the pilot study in the next phase (phase III). As explained in Chapter 3 (subsection 3.2.2), evidence from other methods was needed during the development process to corroborate or challenge interpretations about the robustness of the instrument. Therefore, phases II and III were conducted in tandem so that validity evidence produced from each could be compared. The inferences from such a comparison guided the finalisation of decisions related to retention or modification.

In the validation framework described in Chapter 3 (subsection 3.2.3), the pilot study investigated the psychometric properties of the draft items. Statistical information from the pilot could, for example, impact on the specifics and extent of changes proposed for a draft item in the preliminary recommendation made in this phase. Also, the qualitative responses of the pilot sample could identify hitherto unknown issues with the draft items, that might otherwise be retained. Table 5.10 summarises the outcomes to the draft items in this phase.

Table 5.10

Summary of item outcomes for the 87-item Draft PSRT from findings of phase II

Outcome	Number of draft items	
	1-mark	2-mark
Recommendation for retention	29	6
Recommendation for amendments	28	2
Rejected:		
• Doubts about science content	4	1
• Stand-alone multiple-choice items	3	0
• Not taught in the key stage primary science curriculum	2	1
• Learning outcomes unduly emphasised during instruction	12	1
Total	78	11

5.8 Conclusions

Overall, this phase presented content validity evidence for the Draft PSRT from the expert panel. Analysis of the evidence was carried out using Sireci's (1998a, 1998b) four test features of content validity. From the perspective of domain description of the scientific reasoning construct, the expert panel strongly endorsed its relevance to science learners and the primary science curriculum. The expert panel also generally agreed that the detailed domain description of the knowledge types and reasoning practices were wide-ranging, useful, explicit, and could be learnt by primary pupils. The panel also agreed that the outlined progression of the reasoning practices was logical and reasonable. However, the panel highlighted clarity issues with some of the descriptors, which required modification or reconsideration.

Inferences about the appropriateness of the testing approach came from three types of validity evidence in this phase. First, there was some evidence of alignment between the weighting of knowledge types and practices to the national assessment. This evidence is important because of the production of stronger inferences from pupils' performances on

the finalised PSRT, given their familiarity with the national assessments.[22] Second, evidence from the panel suggested that the scientific concepts assessed in most of the Draft PSRT items were relevant, accurate, and aligned with the science curriculum. Third, the technical quality of the items received general endorsement from the panel. In all, evidence from the panel suggested that the testing approach from the perspective of assessment – weighting, curricular content, and item design – was appropriate.

The evidence collected from the ratings for domain representation suggested adequate representation for the scales and the cognitive demand in the items in the Draft PSRT. Setting a high CVI criteria reduced domain relevance for individual items. However, it was necessary at this phase of the instrument development. Items that did not meet the targeted levels of domain representation and domain relevance were treated in one of two ways, depending on the recommendations from the panel. The first way was to adjust the characterisation schemes of these items, and the second was to examine the design of the item and, if required, amend the item.

In summary, the panel review data provided valuable content validity evidence for making inferences about the strengths and weaknesses of the draft instrument and its items. While the inferences based on the panel's feedback were tremendously useful, they were not necessarily the only evidence considered in final decisions about item outcomes. For instance, unlike the identification of items for deletion, decisions about keeping or amending items considered other sources of validity evidence in phase III – the pilot study. As will be reported in the next chapter, the pilot study examined the psychometric performance of the draft items to produce evidence from test content as well as two other sources: response processes and internal structure. Jointly, inferences from the expert panel and pilot study shaped decisions about retention and amendments leading to the assembly of the final test instrument for use in the main study.

[22] At the time of the research, the targeted Year 6 pupils in the main study had just completed the end-of-course national assessments, which are high-stakes examinations.

Chapter 6 Pilot Study

6.1 Introduction

This phase piloted the draft version of the PSRT. It was conducted in tandem with the expert review presented in the previous chapter. The objectives of the pilot were twofold. First, it enabled a trial of the procedure and activities involved in the administration of the test. On trial were the test duration and sequence, and the clarity of instructions to pupils during the test, as well as the layout and print quality of the booklets. The second objective was to conduct a preliminary analysis of the performance data to ascertain the viability of the data analysis approach and to check the general functioning of the Draft PSRT and its items.

This chapter begins with a discussion of the validation goals of the pilot study. Subsequently, it describes the school and participant profiles, the materials used, and the procedures for testing. It then discusses in depth the rationale and advantages of adopting Rasch measurement theory as the psychometric method in this study, and subsequently the results of the pilot are presented. Valuable evidence produced from Rasch modelling can support inferences about the abilities of test-takers from their performance on the test items. These inferences contribute to greater understanding of the underlying construct measured by the Draft PSRT. The concluding section of this chapter reviews the Draft PSRT by comparing and integrating the findings of this phase with those from the expert review. The following section complements the validation plan (see Table 3.7) by detailing the methods used for the collection of validity evidence in this phase.

6.2 Validation in the Pilot Study

Trialling the Draft PSRT produced validity evidence based on the validity sources of test content and internal structure. Among the four test features of content validity defined in Sireci's (1998a, 1998b) theory (see Chapter 3, subsection 3.2.3), evidence about the suitability of the testing procedures, and the fair and consistent treatment of pupils during the conduct of the test, evaluated the fourth test feature, which was the appropriateness of

the test-development process. The subsequent scoring and analyses of pupils' performance in the test produced evidence from the second source based on internal structure. As explained in Chapter 3 (subsection 3.2.2), this validity source refers to the relationships among the draft items or parts of the test, such as the functioning of the items. In the pilot study, this evidence came from psychometric investigations using Rasch modelling procedures. See Table 6.1 for a summary of the methods used and the targeted inferences.

Table 6.1

Sources and methods used to collect validity evidence in the pilot study

Source of Validity Evidence	Method	Validity Inference
Appropriateness of the test-development process	Administered Draft PSRT	Suitability of testing procedures and conditions
Internal structure	Conducted Rasch modelling procedures	Preliminary information about item functioning in accordance with the construct

6.3 Schools and Participants

One hundred and forty-two 13-year-old secondary one (Year 7) pupils from four schools took part. As shown in Table 6.2, two schools had 38 pupils each participating. Fewer pupils from the remaining schools took part, with the smallest number from one school being 30 pupils ("A" in Table 6.2). Overall, there were twice as many girls as boys in the pilot sample.

Table 6.2

Distribution of the 142 participating pupils according to their gender and school

Gender	Secondary School				
	A	B	C	D	Total
Girls	17	14	36	27	94
Boys	13	24	0	11	48
Total	30	38	36	38	142

Note. "A" refers to secondary school A and so on.

One of the secondary schools (i.e., C) admitted only girls; the other three were co-educational schools. Pupils in these schools were predominantly of mixed ability. The choice of pupils from this grade to pilot the test differed from the grade level of the sample in the main study, who were all Year 7 pupils. The difference was due to a lack of access to the younger pupils who were then undergoing final preparations for upcoming national examinations. Therefore, recruitment took place among pupils who had finished primary school the year before. At the time of testing, these Year 7 pupils from four secondary schools had completed half a year of secondary science instruction. Their performance on the PSRT was thus expected to differ from the targeted sample of Year 6 pupils because of developmental differences and exposure to slightly more advanced scientific concepts.

6.4 Materials

A description of the Draft PSRT and its development process, using requirements in the TOS was detailed in the test framework (Chapter 4, section 4.5). The items in the Draft PSRT can be found in Appendix 6. The detailed TOS is shown in Appendix 9.

6.5 Procedure

I visited the schools to administer the one-hour Draft PSRT to the pupil participants. As part of the pre-test preparation, the teachers acted on my behalf to seek and obtain the necessary consent. They also prepared seating plans that organised their pupils in individual seats to minimise contact and ensure impartial administration of the test. Just before the test, I briefed the pupils to explain the purposes of the research and the test conduct. During the briefing, pupils were reassured that their test results were confidential and had no impact on their school work. They were also told that there would be no repercussions should they withdraw at any point during the test.[23] I concluded the briefing with a short question-and-answer session to address pupils' queries.

[23] There were no instances of pupils withdrawing during the pilot.

After the briefing, every pupil received one of the five Draft PSRT booklets on a rotation basis. Before the commencement of the test, I instructed pupils to note the duration they took to complete the test, and to check the pagination and print quality of the booklets. Throughout the duration of the test, I was present in the venue to answer queries. The administration process and briefing were repeated in every school for consistency. Table 6.3 shows that the total number of each booklet attempted ranged from a low of 26 (booklet 5) to a high of 30 (booklet 2). The two rightmost columns in the table indicate that, except for booklet 5, approximately twice as many girls attempted every booklet.

Table 6.3

Distribution of the five-booklet Draft PSRT by schools and gender

Booklet	Secondary School				Total	Gender	
	A	B	C	D		Girls	Boys
1	6	8	7	8	29	19	10
2	5	7	8	10	30	21	9
3	6	8	6	8	28	20	8
4	6	8	8	7	29	19	10
5	7	7	7	5	26	15	11
Total	30	38	36	38	142	94	48

After the administration of the Draft PSRT, I sampled approximately 10% of the scripts to formulate a provisional mark scheme. Thereafter, an English science intern from the Postgraduate Certificate in Education (PGCE) course at the University of Oxford applied the draft mark scheme by randomly selecting and marking 10 scripts from every school. During this initial round of sampling, she had difficulty comprehending some of the pupils' responses and required clarification from me as the researcher.[24] The intern also highlighted responses to items that could not be accommodated by the provisional mark

[24] Although these responses were in English, the intern was unfamiliar with the phrasing and sentence structure.

scheme. After discussion between us, modifications were made to finalise the mark scheme, which subsequently guided the intern as she marked all the scripts.

In this next round of marking, the intern expressed uncertainty regarding the credit-worthiness of responses from approximately five to eight scripts per school. In a large majority of these cases, the correct science concepts had not been clearly expressed in the pupils' answers. Given the large number of affected scripts, we marked all the scripts a third time together, and focused on ensuring the marking was accurate and reliable across all schools. Since the marking was implemented as a team, there was no computation of marking reliability indices. As a final check, I randomly sampled the recorded marks of 10% of the scripts and found no instances of inaccurate recording. The next section describes the psychometric approach used to analyse the item responses.

6.6 Data Analysis

There are two broad psychometric traditions in test development (Salzberger, 2013). The first established its conceptual foundation from classical test theory and pioneered procedures to guide test construction and evaluation (DeVellis, 2006). The second, relatively newer tradition is commonly known as modern test theory and encompasses two primary families of theories: item response theory (IRT) and Rasch measurement theory (RMT). The following subsection briefly compares these theories and explains why this study adopted the Rasch model for psychometric analysis. Such a comparison highlights the unique psychometric advantages its use afforded, as part of the methodology in this study, in collecting and analysing validity evidence from responses to the draft and final versions of the PSRT.

6.6.1 Classical and Modern Test Theories

Classical Test Theory: The model representing this classical psychometric perspective is as follows:

observed score = true score + error score

In this model, two unobserved scores impact on the observed score. These are the actual score a test-taker would have putatively achieved and the presumed error from random and systematic sources. In practical settings, the realisation of this simple linear relationship requires several empirical assumptions (Gulliksen, 1987). These include treating both theoretical scores as statistically unrelated, presuming the true scores of test-takers to be constant and accepting the presence of normally distributed errors. The classically-derived model has a long and established record of diverse applications in testing such as item statistics and factor analysis (DeVellis, 2006).

Regardless of the advantages, the classical tradition of test development has significant limitations. First, the two component scores are not separable in an observed score, although the two are theoretically independent. Also, by assuming a mean error score of zero in the population, the classical model produces estimates of population parameters but requires representative samples to do so. Furthermore, score interpretations are test and sample-dependent due to the assumption of fixed true scores of individuals on specific tests (Hambleton, Swaminathan, & Rogers, 1991). Even with the availability of statistical procedures (e.g., equating) to compare persons and tests, strict requirements hamper the establishment of reliable score correspondence (Embretson, 1996). For instance, appropriate equating among test forms can only occur when high score reliabilities and comparable distribution of scores are present. Modern test theory purports to overcome many of these shortcomings (Hambleton & Russell, 1993).

Modern Test Theory: This psychometric approach moves away from hypothesised scores and emphasises the unvarying and stable nature of the underlying trait or ability of a test-taker (Lord, 1953; Rasch, 1960/1993). All unidimensional models based on IRT and RMT assume this psychological property or latent trait of persons to be quantifiable and solely responsible for their responses to items (Wilson, 2005). Other commonalities included claims that model-derived item statistics and ability measures are independent of the examinees and test items respectively. Key differences among the models arise from the paradigmatic treatment of test data (Andrich, 2004; Mellenbergh, 2011).

Briefly, the Rasch models subscribe to the notion that the probability of successful responses relies only on ability and item difficulty. Practitioners claim that desirable model properties facilitate interval-level interpretation of the raw scores when there is appropriate model fit (Wright, 1999). In effect, what these proponents mean is that Rasch modelling "transforms" ordinal data that are organised into classes, and whose distances from class to class do not represent equivalent amount of the trait measured, into scales with equal intervals (Merbitz, Morris, & Grip, 1989). Still, there is contention over the theoretical and empirical basis of the Rasch model (Andrich, 1989, 2004). For instance, some researchers doubt that the simplicity of the model accounts adequately for item responses (see, e.g., Kline, 2000). Other critics question the sensitivity of the fit statistics and the criteria adopted for evaluating model-fit using these estimates (Heene, 2013).

More complex IRT models attempt to manage model-fit issues by factoring in additional item parameters, such as discrimination and guessing, to account for variations in responses (de Ayala, 2009). With more parameters, these models tend to fit the data better (Andrich, 2004). However, robust estimates of these extra parameters require larger sample sizes than simpler models (Şahin & Anil, 2016). Furthermore, there are doubts about the substantive clarity of the additional parameters as they are susceptible to non-postulated factors such as examinee characteristics (Humphry, 2009). Notwithstanding these debates, this study adopted the unidimensional Rasch measurement model [25] for unique psychometric features, which advantaged the collection and analysis of validity evidence in test development. The next subsection introduces this psychometric approach and elaborates its distinctive merits.

6.6.2 Rasch Measurement Theory

The Fundamental Idea in Rasch Measurement: All Rasch models postulate that success in solving an item only depends on a test-taker's latent ability and the difficulty of the item. Modelling these person and item attributes are the parameters for ability (θ) and item

[25] Multidimensional Rasch models also exist. E.g., Briggs and Wilson (2003).

difficulty (δ). Depending on the total marks the items are awarded, there are two basic types of items: dichotomous and polytomous.

(a) Dichotomous (one-mark) item and item location: Mathematically, the probability (*P*) of achieving a correct response depends on the difference between the estimates of these parameters. This relationship is expressed in the following formula for achieving a score *x* on a one-mark (dichotomous) item i in a test with *n* items by a person *v*:

$$P(X_V = 1) = \frac{e^{(\theta_v - \delta_i)}}{1 + e^{(\theta_v - \delta_i)}}$$, i = 1, 2, …, *n;* θ and δ_i are non-negative real numbers

Simply put, the above formula suggests that the greater the ability of the test-taker, or the easier the test item, the probability of answering the item correctly increases in a log-linear fashion. For a dichotomous item, the point (or threshold) where a test-taker has a 50% probability of attaining a successful response is the item's difficulty (or its location).

(b) Polytomous (two-mark and above) item and item thresholds: A polytomous item has more than one threshold; these are points on the scale where the chances of responding correctly on adjacent response categories are 50% (Andrich, 1988). For instance, in a two-mark item intended to invoke three categories of response (zero credit, one mark, and two marks), there are two thresholds. The first divides the two less difficult response categories and shows the location where there is a 0.5 probability of scoring zero or a single mark.

If the item is functioning properly, the next threshold is located higher on the scale (i.e. more difficult); at this point, there is as much likelihood of attaining a mark as there is of gaining two marks. Item responses that show thresholds between higher scoring categories to be more difficult than between lower scoring categories demonstrate ordered thresholds. This means that across the scale, as ability estimates increase, the probability of exceeding the first and second thresholds (i.e., gaining more marks) correspondingly increases. If the thresholds are reversed in this two-mark item, the middle category (i.e., one mark) is not

working properly to differentiate responses during scoring. As a polytomous item with ordered thresholds evidences more than one hierarchical level of complex knowledge or competencies, it offers more information than a dichotomous item (van Wyke & Andrich, 2006). More information about the thresholds of polytomous items is found in the next subsection.

The Total Score as a Sufficient Statistic*:* From the perspective of RMT, the total score of a test is a sufficient statistic, or a statistic that summarises all the information about the ability of the test-taker (Andrich, 1988). Similarly, arising from the relationship between the two parameters in the model, a symmetrical argument can be made about the sufficiency of the total score of an item; it is a statistic containing all the information about the difficulty of the item.

Parameter Invariance: The invariance of the parameter estimates is the measurement foundation of the Rasch model (Wright & Stone, 1979). This measurement property means that acceptable model-data fit permits the drawing of meaningful inferences about the items and test-takers that are independent of each other, i.e., inferences about the relative abilities of the test-takers do not depend on the specific items given to them, and comparisons between item difficulties do not depend on the sample of test-takers responding to them (Andrich, 1989).

Unidimensionality and Local Independence: The Rasch measurement approach carries two primary assumptions (Mellenbergh, 2011). The first assumption of unidimensionality presupposes that a single latent trait accounts for the range of manifested responses (Hambleton et al., 1991). The Rasch model also assumes item responses to be locally independent, or in other words, that the common latent trait fully accounts for the inter-correlations between all items and there are no additional relationships between subsets of items (Baghaei, 2008). When subsets of items in a test relate to each other over and above what is accounted for by a common latent trait, this leads to biased estimates of parameters and impacts on the assumption of the unidimensionality of the test. The next discussion

identifies four pertinent implications of these properties and assumptions of the Rasch model and explains how each contributed to and impacted on the item and test development in this research.

***(a) Smaller sample size and shorter test length for stable calibration*:**

The sample size and the corresponding test length required for robust estimates with acceptable standard errors of measurement (measures of score precision) are smaller for Rasch models (Şahin & Anil, 2016). Linacre (1994) proposed a minimum sample size of 108 individuals for attaining stable item or person parameters within plus/minus half a unit (logit) on a dichotomous Rasch scale with 99% confidence. The recommended minimum, while adequate with a well-designed scale, increases to 243 with poorly-targeted items (i.e., items that are too difficult or too easy for the test-takers).

The sample requirement is more stringent with polytomous items (Linacre, 2002). However, in instances where the precision requirements of the estimates are secondary objectives (e.g., piloting and exploratory test development), a sample of 50 individuals is put forth as the minimum for achieving estimates of appropriate precision in polytomous scales within plus/minus 1 logit with 99% certainty (Linacre, 1994). These recommendations of sample size and item number that are smaller than required in more complex IRT models are convenient for recruitment and development of test instruments. In this study, sample sizes recruited for the pilot (n = 142, Draft PSRT test length = 87 items) and main study (n = 431, PSRT test length = 100 items) phases fell within these guidelines.

***(b) Flexible sample recruitment and item formats*:**

Unlike classical tests, Rasch modelling does not require representative samples if the data fits the model (Embretson, 1996; Hagquist, Bruce, & Gustavsson, 2009). In addition, the analysis of tests with mixed item formats using the Rasch approach does not incur the problems (e.g., unequal weighting on final scores) that beset classical methods (Embretson & Reise, 2000). Instead, the precision of the estimates parameters improves when the sample is of mixed abilities and the item format is varied (Embretson & Reise, 2000).

***(c) Customisable test forms*:**

Validating a new test requires an adequate number of items to test diverse aspects of the construct of interest (AERA, APA, & NCME, 2014). Moreover, the minimum number of items to achieve acceptably precise estimates is often too large for test-takers to attempt in a single sitting. One way of overcoming this limitation is to administer a subset of the items to every test-taker to ensure coverage of all items. A common procedure to achieve full coverage without compromising on measurement rigour is the administration of test forms created from a matrix sampling design (Gonzalez & Rutkowski, 2010).

Analysing data from multiple test forms is possible due to the invariant properties of the estimated item and person parameters from the Rasch model, where the common person links across items are used to estimate the relative item difficulties and represent them on the same scale (Andrich, 1988; Shoemaker, 1971). Accordingly, there were five test forms or booklets in the pilot and final versions of the PSRT. The test framework outlined in Chapter 4 described the test assembly process and how the items related across the booklets (section 4.5).

***(d) Diagnostic opportunities using psychometric and theoretical perspectives*:**

Fit statistics on data analysed using the Rasch model are sensitive to poor item functioning or violations of the model assumptions. A wide range of fit indices are available during test development to signal potential anomalies between the item responses (i.e., data) and the model expectations (Hagquist et al., 2009). Andrich, de Jong, and Sheridan (1997) advised that items with detected psychometric anomalies be further analysed in substantive terms. Thus, utilisation of psychometric data, theoretical perspectives (e.g., construct), and contextual information (e.g., curriculum) promotes rigorous examination of item quality (Andrich, 1988). In turn, there is the facilitation of test development and validation. The next subsection identifies the software used for the Rasch procedures, and describes the suite of statistical tests of model fit of the data collected during the pilot phase. It also discusses the treatment of unanswered items.

6.6.3 Rasch Analysis

The Rasch analysis in the pilot study examined three broad groups of fit statistics to ascertain the degree of fit between the responses on the Draft PSRT and the model. The RUMM2030 software was used for the analysis (Andrich, Sheridan, & Luo, 1997-2012b).

Test Fit Statistics: The first group, as summarised in Table 6.4, consisted of overall tests of fit of the data to the model. A discussion of these fit indices follows the table.

Table 6.4

Test-level fit statistics and statistical tests examined in phase III

Fit Statistics	Description
Mean and standard deviation of the fit residual of items in the test	Compare to approximate expected values of 0 and 1 respectively
Mean and standard deviation of locations of persons	Compare to fixed mean and standard deviation of the locations of items
Mean and standard deviation of the fit residual of persons	Compare to approximate expected values of 0 and 1 respectively
Chi-square test	Check overall chi-square test of fit to establish if there is a statistically significant deviation in data fit at 5% significance
Person Separation Index	Check Person Separation Index to ensure test has adequate reliability and differentiating power

Item and sample fit statistics indicate how well data at the item and person levels fit the model (Andrich, 1988). The means and standard deviations of parameter estimates are expected to be close to 0 and 1 respectively for both items and persons. There was also comparison of the estimated mean location of the persons to the software-fixed mean location of the test items to produce information about the targeting of the difficulty of the items to the sample's average ability. For example, a relatively higher mean person location

estimate signals that the sample is more able than the average difficulties of the items, and vice versa.

The fourth test-level fit index is a chi-square statistic which evaluates the invariance properties of the construct across the scale by summarising the relationship between responses of groups of test-takers of differing abilities and model expectations. A non-statistically significant result indicates that the performance of ability groups along the scale accord to the model expectations across all the items, while a significant result signals a violation of the invariance assumption of the model.

The final test-level statistic – the Person Separation Index (PSI) – serves as the equivalent of the classical test-based reliability index for internal consistency, i.e., Cronbach's alpha coefficient. The PSI is an index of the proportion of true variance in the person estimates relative to the observed variance, which also includes error variance. In addition, it indicates how well the items differentiate the test-takers by ability along the scale and is a criterion to evaluate the strength of the fit statistics (Wright & Stone, 1999). A minimal level of 0.7 is necessary to support claims that the test has adequate power to differentiate between two groups of examinees, as well as providing a moderate level of power for the tests of fit (Andrich, 1982; Fisher, 1992). Higher PSI values, such as those between 0.8 and 0.9, suggest the reliable discrimination of three to four groups and indicate well-powered fit statistics to detect misfit (Andrich, 1982).

Item and Person Fit Statistics: While the first group of indices provided a summary of the overall test-level fit, the second group of reported fit-statistics provide information about each item and test-taker with respect to model expectations. A summary of these item and person-level fit statistics is provided in Table 6.5.

Table 6.5

Item and person-level fit statistics and statistical tests examined in phase III

Focus of analysis	Fit statistics	Description
Item	Fit residual of each item	Check if fit residual meets <± 2.5 criteria
	Chi-square test	Check item-level chi-square probability to establish if there is a statistically significant deviation in data fit at 5% significance
	Analysis of variance (ANOVA) test	Examine item characteristic curves (ICCs) of items and results of analysis of variance across class intervals of groups and across subgroups (schools, genders and booklets)
	Threshold order	Examine the functioning of response categories in polytomous items
	Location of item	Examine location estimates
Person	Location of person	Examine location estimates
	Fit residual of each person	Check if fit residual meets <± 2.5 criterion

Item and person-level fit statistics consisted of the residuals and the chi-square indices. The former are summed data-model deviations for the item and person responses, which are standardised to approximate normal distributions with expected means of 0 and standard deviations of 1. Residual values that are greater than +2.5 or less than -2.5 indicate unacceptable model fit (Pallant & Tennant, 2007). Item and person-level chi-square indices reflects misfit in the performance of items or persons which did not occur by chance.

Item characteristic curves (ICCs) provide visual details of the fit by comparing the observed means of nearly equivalent-size class intervals (i.e., ability groups) along the scale to the model expectations (Hambleton et al., 1991). If there is good item-model fit, the observed means of the class intervals fall on or near to the theoretical curve. Plots of observed class interval means that have steeper or flatter slopes than the model curve indicate misfit. The ICCs also show the location estimates (i.e., difficulty) of the items on the scale. As mentioned in the previous subsection, for a one-mark (dichotomous) item, this is the point

or threshold where the probability of responding in either of the categories is 0.5. An item that is worth more than one mark (polytomous) has more than one threshold. Its item location is the average of the thresholds present for the item. Figure 6.1 shows the ICC output from RUMM2030 for a dichotomous item from the Draft PSRT. As shown in the ICC, the pilot sample was divided into three class intervals (indicated by the black dots).

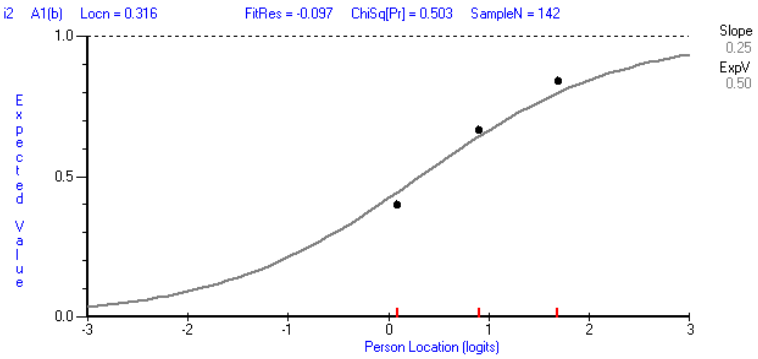


Figure 6.1. The ICC of the dichotomous item – A1(b) from the anchor block. The item had a location estimate of 0.316. The black dots show the average response of the three class intervals and the curve shows the model expectations. Visual inspection suggested there were slight deviations between the observed means and the model. The item fit residual of -0.097, which was close to the expected value of 0, indicated acceptable overall deviation from model.

Other than the item locations, when the observed proportions are plotted for different subgroups of the sample, such as gender and school affiliation, the ICCs also provide useful information about the functioning of the items among these subgroups (Kreiner, 2013). Ideally, given similar or identical ability estimates, the performances of the subgroups should be similar or identical due to the operation of an invariant scale. Visually on the ICC, the observed proportion of correct responses of these subgroups should be approximately identical across the full range of abilities and accord with model expectations. Evidence of deviations, also known as differential item functioning (DIF), suggests violations of invariance (Zumbo, 2007).

Detection of DIF comes from analysis of variance tests which compare both the actual performance across subgroups and their class intervals. DIF occurs in two general forms –

uniform and non-uniform. The former is a statistically significant main effect characterised by a consistent amount of deviation between one or more subgroups and the remaining subgroup(s). In contrast, non-uniform DIF is a significant trait-group interaction and occurs when there are variable differences among the subgroups. See Figure 6.2 for the ICC of the only item in the Draft PSRT with observed uniform DIF among schools.

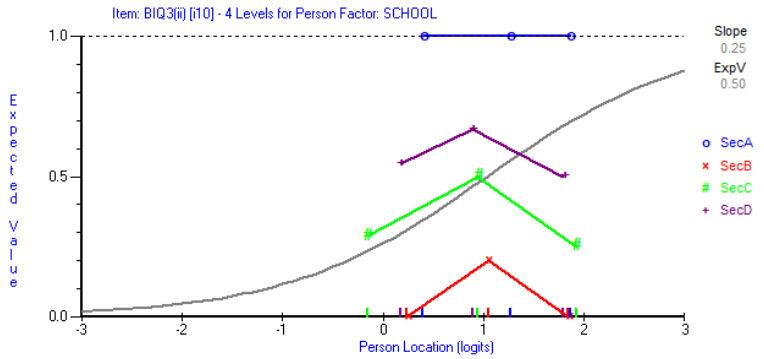


Figure 6.2. Uniform differences observed among subgroups (schools) reflect the presence of DIF in a dichotomous item – BIQ3(ii) from block I. This meant there was a consistent deviation in performance on the item among the four schools.

Post-hoc remediation of DIF typically involves resolving the affected item into distinct items for the subgroups. For instance, an item with an observed gender DIF effect is “split” into two items – one for males and one for females. In the pilot phase, DIF was not examined for the Draft PSRT due to insufficient sample sizes for the subgroups but will be examined for the main study in the following chapter.

Besides the ICCs, the category probability curves (CPCs) also serve as a source of graphical information. CPCs are useful for examining the functioning of the response categories of polytomous items (Wright & Masters, 1982). As described earlier, ordered response categories represent competencies sequenced hierarchically according to theoretical considerations (Andrich, 2010a, 2010b). Thus, aberrations in the ordering of the threshold estimates, which represent the locations where the probability of being located in adjacent categories is 0.5, warrant further examination of the prescribed competencies in the affected response categories (van Wyke & Andrich, 2006). See Figure 6.3 for the CPC

of an item from the Draft PSRT with ordered thresholds. In contrast, the CPC of another item that demonstrated disordered thresholds is shown in Figure 6.4.

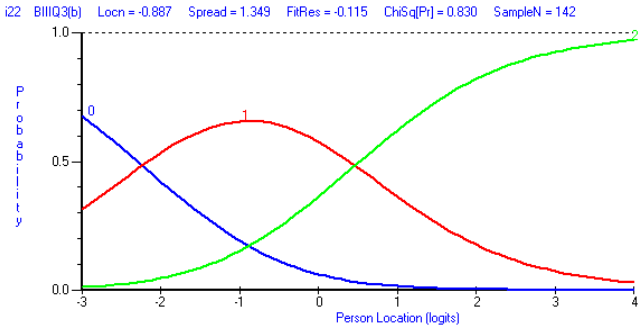


Figure 6.3. The category probability curve (CPC) of the polytomous item – BIIIQ3(b) from block II, demonstrating ordered thresholds. Persons with ability estimates ranging from approximately -2 to 0.5 logits on the scale were more likely to attain the middle response category of one mark compared to the adjacent categories of no mark or two marks.

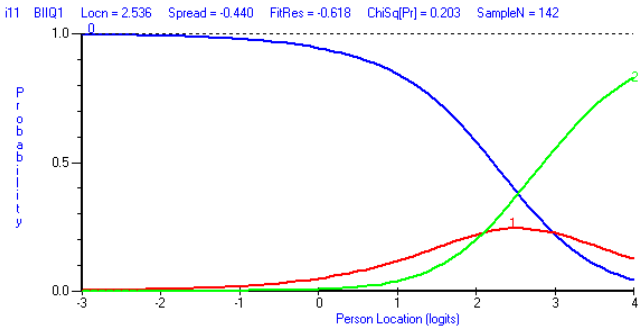


Figure 6.4. The CPC of the polytomous item – BIIQ1 from block II, demonstrating disordered thresholds. The middle scoring category of one mark is not functioning as intended.

A common resolution of disordered thresholds is to collapse adjacent categories when there is sufficient evidence that the responses in these categories are not distinct (Andrich, 2010b). However, this method of resolving disordered thresholds does not always work empirically to improve overall model fit (Pallant & Tennant, 2007). Given the concern about the accuracy of the parameter estimates due to a small sample, there were no statistical attempts to resolve polytomous items found to have disordered thresholds in the pilot study.

Treatment of Missing Data: Missing data occurred because the pupils failed to answer specific items in the Draft PSRT. There was no empirical evidence that pupils in the four secondary schools were unable to complete any of the five test booklets in the one hour allocated. For instance, the average duration recorded by pupils for completing their booklets was below 45 minutes in all schools (see Table 6.6 in the results section). Also, all 142 scripts had evidence of attempts on the items of the last question of the script. Therefore, it was likely that pupils left the items unanswered because of insufficient knowledge and ability.

Studies on the treatment of such omissions within the Rasch measurement framework broadly recommend two strategies (Custer, Sharairi, & Swift, 2012). The first is to ignore the omitted items in the computation, and the second is to score the omissions as incorrect. The results of using these strategies are mixed (DeMars, 2002; Koretz, Lewis, Skewes-Cox, & Burstein, 1993). For instance, Ludlow and O'Leary (1999) reported that treating unanswered items as missing resulted in inflated person estimates and deflated item estimates. The reverse happened when these items were scored as incorrect. Considering these findings, this research adopted the second scoring approach to produce conservative parameter estimates. Moreover, designating a blank response as an incorrect answer is the typical approach applied in science classroom tests to reflect gaps in teaching and learning. Treating these instances of data missing from the Draft PSRT similarly in the analysis could possibly ensure that final inferences about the items and construct would be more relevant to educators and assessors.

6.7 Results

6.7.1 Draft Primary Scientific Reasoning Test (Draft PSRT)

Duration: Pupils across the four schools took approximately 38–43 minutes on average to complete the Draft PSRT (see Table 6.6 for the average and range of duration). These self-reported values suggested that the test length was too short and the number of items insufficient to engage pupils optimally for the hour-long test duration.

Table 6.6

Average and range of reported duration to complete the Draft PSRT per school

	Duration of Completion in Secondary School (Minutes)			
	A	B	C	D
Average	38.6	42.2	40.6	37.9
Range	25–50	20–60	40–58	25–60

Note. "A" refers to secondary school A and so on.

Overall Fit: The total-item chi square to degree of freedom ratio of 1.132 (196.92/174) was not statistically significant (p = .112), suggesting no violation of the invariance condition and thus acceptable fit between the data and the model. A comparison of the mean fit residuals of the 87 items (0.062, *SD* = 0.754) and the corresponding index for the pupil sample (-0.119, *SD* = 0.645) with the expected values of 0 and 1 indicated that there was a higher level of misfit among the pupils' ability estimates, but this misfit was quite marginal.

Reliability and Targeting: At 0.757, the PSI indicated adequate power to differentiate among three groups of pupils. The difference between the mean location estimates of the draft items, which had been fixed at 0 (*SD* = 1.591) and the mean ability estimates of the pupils (0.859, *SD* = 0.771) showed some mistargeting. The person-item threshold distribution in Figure 6.5 pinpoints the primary source of mistargeting below 0 logits where half of the item difficulty estimates were located and very few pupils were located.

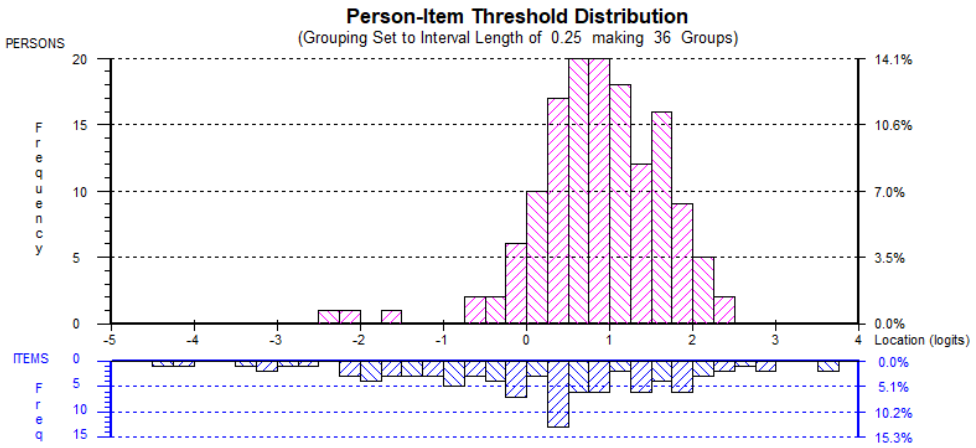


Figure 6.5. Person-item threshold distribution of the Draft PSRT. This graph superimposes the distribution of the items (in blue-patterned bars) and the persons (in red-patterned bars) on the same scale.

6.7.2 Item and Person Fit Statistics of the Draft PSRT

Item-Level Fit Statistics and Tests: Among the 76 dichotomous items in the Draft PSRT, there was only a single instance of a fit residual that exceeded the +2.5 threshold. This observed fit threshold (2.717) indicated unacceptable fit to the model. As shown in Figure 6.6, the ICC of this item demonstrated under-discrimination relative to the model expectation (i.e., the average discrimination across all items). Pupils of the lowest ability performed better than predicted by the model, while the most able pupils underperformed.

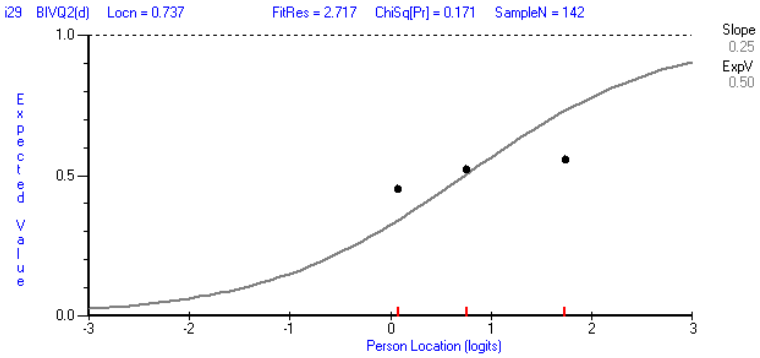


Figure 6.6. ICC of an item from question 2 of block IV showing under-discrimination relative to model expectation.

A second irregularity detected was the presence of uniform DIF in an item whose ICC was shown earlier in Figure 6.2. As shown in red on the graph in the figure, school B demonstrated consistently poorer performance across all three class intervals compared to the other schools. Similarly, there were consistent differences in performance among the remaining schools. Of the 11 two-mark polytomous items in the test, responses to six demonstrated disordered thresholds, which signalled that one or more scoring (response) categories were not functioning as intended.

The CPC of one of these items was shown earlier in Figure 6.4. In this item, the middle scoring category of one mark did not function as expected between the two adjacent categories. Therefore, regardless of the ability estimates of the pupils, the probability of achieving a single mark never exceeded the probabilities of receiving zero credit or two

marks. As will be discussed in subsection 6.8.1, polytomous items with disordered thresholds were either earmarked for changes (e.g., re-crafting) or had their mark schemes refined. The subsection also discusses how results from the psychometric analysis impacted on the preliminary recommendations from the expert review to finalise item retention and modifications.

The evidence described above suggested 60% of the dichotomous items evidenced some form of model-fit issues in their ICCs, and more than half of the 11 polytomous items had disordered thresholds. As no statistical remediation was conducted on the pilot data due to concerns about the accuracy of the item parameters, no further inferences about the construct could be drawn using the location estimates from the pilot data. Appendix 14 reports the location estimates and fit-statistics of all the draft items. Item-level estimates and fit-statistics, along with a brief description of the model fit observed in the ICC and/or CPC, have been reported in an earlier appendix (Appendix 6).

Person-level Fit Statistics and Tests: In the pilot sample, there was only a single instance of a fit statistic with a residual value just marginally above the upper-bound criterion at 2.564. The person was not removed from the analysis since it was uncertain how model fit issues highlighted in the preceding discussion affected the performance of the sample.

Performance and Targeting by Gender: The mean ability estimate of the boys in the pilot sample was 0.982 (SD = 0.73), and 0.797 for the girls (SD = 0.79). The difference in performance was not statistically significant ($F_{1, 140} = 1.841$, $p = .177$). As shown in Figure 6.7, the least able pupils found below -2 logits were girls. However, both genders were represented at the top end of the scale.

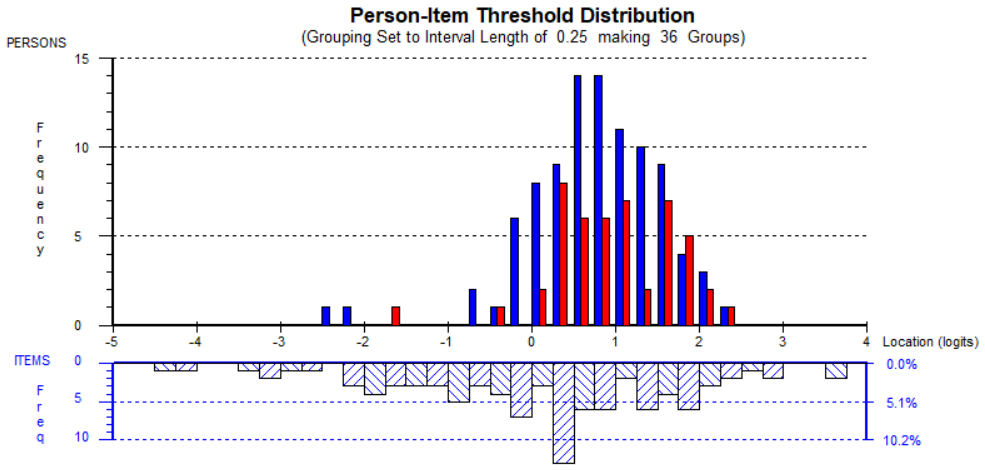


Figure 6.7. Person-item threshold distribution by gender of the Draft PSRT. Blue and red columns represent the location estimate distribution of girls and boys respectively.

Functioning of Booklet: The mean difficulties of the booklets were between 0.711 (*SD* = 0.81) and 0.996 (*SD* = 0.72). Levene's test for equality of variances was insignificant ($F_{4, 137}$ = 0.441) at p = .779. Therefore, post-hoc analyses used Tukey's test to evaluate the pairwise differences among the mean difficulties of the booklets. An insignificant result in the ANOVA ($F_{4, 137}$ = 0.803, p = .525) suggested that statistically the booklets were not significantly different in difficulty. See Table 6.7 for a summary of these booklet-level indices, as well as the statistical comparison of these means.

Table 6.7

Confidence intervals of pairwise differences in mean difficulty estimates between booklets

Booklet	Difficulty Estimate		95% Confidence Intervals of Pairwise Differences			
	Mean	*SD*	1	2	3	4
1	0.944	0.75				
2	0.711	0.81	[-.32, .79]			
3	0.996	0.72	[-.62, .51]	[-.85, .28]		
4	0.913	0.59	[-.53, .59]	[-.76, .35]	[-.48, .65]	
5	0.729	0.97	[-.36, .79]	[-.59, .55]	[-.31, .85]	[-.39, .76]

Note. Comparison performed using Tukey's method at 0.05 significance level.

Performance of Schools: As shown in Table 6.8, pupils from two of the schools had mean abilities above the sample mean estimate of 0.859 (*SD* = 0.77) with the highest mean recorded from school A (1.411, *SD* = 0.54). In contrast, the performance of the remaining two schools was noticeably lower, producing means that were almost two-thirds below school A. Pupils from school D had the poorest performance.

Table 6.8

Confidence intervals of pairwise differences in mean ability estimates between schools

School	Ability Estimates		95% Confidence Intervals of Pairwise Differences		
	Mean	*SD*	A	B	C
A	1.411	0.54			
B	1.066	0.58	[-0.02, 0.72]		
C	0.537	0.94	[0.37, 1.38*]	[0.03, 1.02*]	
D	0.523	0.59	[0.52, 1.26*]	[0.18, 0.90*]	[-0.48, 0.51]

*Note. *Difference in means is significant using Dunnett's C method at 0.05 significance level. A is secondary school A and so on.*

Unlike the results of statistical tests on gender performance and booklet functioning, there were significant differences ($F_{3, 138} = 13.301$, $p < .001$) among the mean abilities of pupils of the four schools. Post-hoc comparisons of the school performance means did not assume equal variances as the variances from the four schools varied widely from 0.292 to 0.884.

Moreover, the non-significant result from the test of homogeneity of variance ($F_{3, 138} = 2.207$, $p = .09$) could be biased by a lack of power due to a small sample. From the results of these post-hoc analyses (see Table 6.8), the mean ability estimates of schools A and B differed significantly from the other two schools. These results showing a broad spectrum of abilities among the pupils in the pilot sample were to be expected as they were chosen through convenience sampling rather than through random sampling.

6.8 Discussion

In this pilot phase, the collection of evidence was related to the two validity sources of test content (appropriateness of the test-development process) and internal structure. Evidence from the trial of the test administration procedures suggests that the test length and duration had not been optimised to draw out maximal inferences about the abilities of the pupils because there was an inadequate number of items in the pilot. On the other hand, there was also evidence that the approach of briefing pupils before the test and organising the test items into five booklets elicited pupils' best efforts and ensured they were not overly taxed by the demands of the test.

The scoring process examined the responses and produced initial evidence to evaluate whether the intended learning objectives had been targeted successfully. Most of the marking issues discovered by the two markers (myself and an intern) were related to uncertainties about the credit-worthiness of responses in items with CRSS formats. The qualitative examination of the responses during marking suggested that most of the items in the 87-item Draft PSRT elicited the required objectives and reasoning.

Rasch analysis of data from the Draft PSRT drew on various fit statistics to evaluate data-model fit at three levels. At the test level, fit statistics provided information about the interaction of the items in the test with the sample and trait. At the individual item and person levels, fit analysis involved both statistical indices and graphical information. Items or persons with poor data fit were checked to understand the underlying issues affecting the responses. Apart from a single dichotomous item with uniform DIF, the test-fit statistics supported initial inferences that the Draft PSRT functioned invariantly across the subgroups of schools, booklets, and genders.

The PSI indicated that the test had adequate power to differentiate the sample reliably into three class intervals, or ability groups, and thus supported inferences about item functioning drawn from the ICCs and fit statistics of the items. The individual item and fit statistics diagnosed a single dichotomous item and person with misfit; the remaining 76 dichotomous

items displayed acceptable fit statistics. Of the 11 two-mark (polytomous) items, six had disordered thresholds. At the booklet level, there were no significant differences in difficulty across the booklets. Other than psychometric results that indicated probable issues with some items and persons, further inferences about the unidimensionality and characteristics of the construct tested could not be drawn due to the small sample size of the pilot.

Overall, the pilot study produced evidence to support and improve the test administration procedure in preparation for the main study. On the basis that there were no marked indications of psychometric issues arising from the analysis of the pilot data, there was reasonable evidence to conclude that RMT was a viable psychometric approach for validating this instrument. The next subsection discusses how the findings from the pilot study and the expert review were integrated to finalise two item outcomes in the Draft PSRT: retention and modification.

6.8.1 Final Item Outcomes Based on Evidence from Phases II and III

As reported in the previous chapter, the expert review phase produced evidence about the content validity of the Draft PSRT. In response to the panel's feedback on the 87-item Draft PSRT, 23 items were deleted (see Chapter 5, subsection 5.7.3, for the details and rationale). The remaining items received preliminary recommendations for retention or amendment in the PSRT (see Appendix 12 for these recommendations), pending validity evidence from the pilot study. The rest of this sub-section discusses how the final outcomes of these draft items depended on their fit-statistics shown in Appendix 14 from the Rasch analysis using responses from the pilot sample. To facilitate the discussion, a summary of these final item outcomes from the findings of phases II and III is presented first in Table 6.9.

Table 6.9

Summary of final outcomes for 64 items in the Draft PSRT from findings of phases II and III

Item Outcomes from Phase II	Psychometric Finding from Phase III	Final Item Outcome	Number of Items	
			1-Mark	2-Mark
Recommendation for retention	Acceptable item-fit statistics	Retention	28	2
	Anomalous item-fit statistics	Amendment	1	4
		Total	29	6
Recommendation for amendment	Acceptable item-fit statistics	Amendment	26	0
	Anomalous item-fit statistics	Amendment	1	2
		Total	27	2

A comparison of the preliminary item outcomes from the expert review and the psychometric results indicated that responses to most of the 35 items that were recommended for retention displayed acceptable item-fit statistics. Therefore, these items were retained without changes for inclusion in the PSRT. Of the remaining five items which evidenced psychometric flaws, four were polytomous items with disordered thresholds. The remaining dichotomous item showed a fit residual that was above the acceptable threshold. After examining the responses to these items, amendments were made to their mark schemes.

Twenty-six Draft PSRT items that received change recommendations had acceptable item-fit statistics. Hence, the recommendations of the expert panel took precedence in the final outcomes, and the items were amended according to the recommendations. Of the remaining three items with recommendations for changes, one was a dichotomous item with DIF, and the other two were polytomous items with disordered thresholds. These items were modified using both the inputs of the panel and their psychometric results. In summary, the final amendments that items received affected either part of the item (e.g., graphics,

question, or phrasing in the stem) or the mark scheme. Details of the final outcomes at the item-level are available in Appendix 15.

6.9 Conclusions

The preceding discussion explained how the findings from phase II (expert review) and phase III (pilot study) informed decisions about the suitability and treatment of the draft items for inclusion in the final test instrument – the PSRT. As well as strengthening test content validity, combining evidence across both phases in decisions about item outcomes mitigated the constraints that limited the inferences drawn from the psychometric results of the pilot data about internal structure. The next chapter reports the findings of the final phase (phase IV), which was the main study, and describes how a comprehensive suite of Rasch indicators and procedures was used to analyse performance data from a larger group of pupils of the targeted grade. Unlike the earlier phases, these analyses also provided validity information from a third source: response processes.

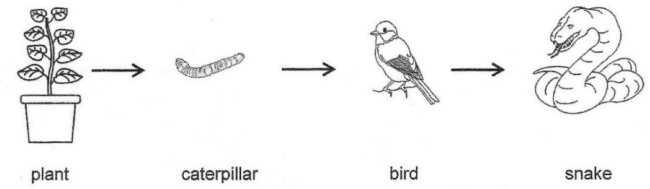
Chapter 7 Main Study – Dimensionality of Construct

7.1 Introduction

In the previous two phases, the Draft PSRT underwent expert review and psychometric piloting. The preceding chapter explained how the empirical findings from these phases resulted in the deletion of slightly more than a quarter of the items in the Draft PSRT. The present chapter and the next discuss the results of activities undertaken in the last phase of the research to finalise and validate the PSRT. Part of the validation involved the participation of primary pupils, who completed booklets of the PSRT.

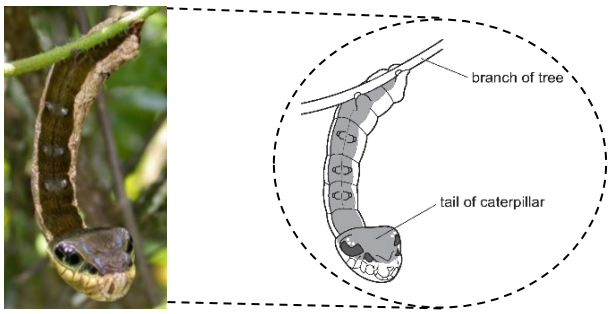
The analysis of pupils' performance used Rasch modelling, which was introduced in the previous chapter to analyse the results of the pilot study. The following discussion of a two-item question, shown in Figure 7.1, illustrates how similar fit indices from this psychometric approach and qualitative analyses of the responses produce inferences about the functioning of its two items.

Study the food chain below.



(a) Identify the producer in the food chain.

When attacked by the bird, the caterpillar raised its tail as shown below.



(b) Explain why the raised tail of the caterpillar will frighten the bird away.

Figure 7.1. A two-item question in the PSRT to illustrate how evidence from various methods of analysis gave insights about the functioning of the items.

The first item targeted the recall of basic content knowledge and was one of the easiest items in the PSRT. It also had good statistical fit to the model. Similarly, fit indices showed that the two-mark mark scheme in the second item performed as intended. However, more able pupils underperformed on the item. Further checks on the scripts found that many in this group did not mention in their answers that the caterpillar's tail resembles a snake.

There were two possible reasons for this omission. First, these pupils may have reasoned that the similarity was obvious and thus redundant. Alternatively, the differences in the depictions of the snake in the food chain (item 1) and the supposed snake (caterpillar's tail) in item 2 may have inadvertently caused pupils to identify the animal mimicked by the tail inaccurately. Jointly, these findings provided information about factors that affected these items' functioning and pupils' reasoning on them, which in turn shaped decisions about the item including possible areas for improvement. The next section complements the validation plan outlined for this phase in Chapter 3, subsection 3.2.3, by detailing the sources and methods for gathering validity evidence in this final phase of the research.

7.2 Validation in the Main Study

In this thesis, validation of the PSRT uses evidence from three validity sources: test content, response processes, and internal structure. The accumulated evidence from these sources provides inferences about whether the test validly measures scientific reasoning. The evidence collection and analytical methods in this phase took place over two sequential stages. Activities to assemble the PSRT in the first stage collected test content evidence, or evidence that the items and test sampled the construct of scientific reasoning appropriately and adequately (AERA, APA, & NCME, 2014; Sireci & Faulkner-Bond, 2014). Briefly, these activities included designing new items to undergo an informal and brief round of expert review. The PSRT test specifications were also drawn up, using feedback from the earlier phases and the informal review. The second and final stage gathered evidence from three validity sources derived from the psychometric and content analyses of pupils' performance in the test.

Overall, the qualitative and quantitative evidence collected from the various validity sources supported inferences about score interpretation and usage. This technique of integrating relevant evidence to yield overall conclusions has been called a "warranted assertion analysis" by some researchers (Onwuegbuzie & Combs, 2010). It was one of the five cross-over analytical techniques adopted in this study to strengthen conclusions from multiple mixed methods. See Chapter 3 (subsection 3.2.2) and Table 3.3 for more details on cross-over techniques. As a detailed validity plan can also be found in Chapter 3, subsection 3.2.3, Table 7.1 provides a summary of this plan in the two stages of the main study, listing the corresponding research questions, data collection, and analytical methods.

Table 7.1

Summary of validity plan in the two stages of the main study

Stage	Validation Source	Research Aim and Question	Method of Collecting Validity Evidence	Summary of Data Analysis and Cross-over Strategy (denoted with *)
1	Test content	What is the experts' professional judgement about the new items?	Informal review of new items by two experts	Qualitative review by experts
2	Test content Response processes Internal structure	Aim: Produce evidence that the PSRT is measuring a unidimensional construct What are the item characteristics in the PSRT? How does the PSRT function regarding the interaction of the person–gender, person–school, and person–booklet in each item? Aim: Examine if the underlying construct measured in the PSRT conforms to the hypothesised construct of scientific reasoning Do the item location and threshold estimates conform to the expectation predicted by theory?	Administer PSRT to 431 Singapore pupils in six participating primary schools	Conduct fit analysis using Rasch modelling Qualitative analysis of pupils' performance in the PSRT items *Warranted assertion analysis – draw meta-inferences or broader inferences from qualitative and quantitative methods

7.3 Validation Stage 1

The goal of this stage was to obtain test content evidence from a second expert panel review. This informal review collected evidence from three test features of content validity to evaluate the extent of alignment between the construct and the curriculum for the new items (Sireci, 1998a, 1998b). These test features were domain representation, domain relevance, and the appropriateness of the test development process. As mentioned, there are detailed accounts of these features and the methods for collecting validity evidence in Chapter 3 (subsection 3.2.3). Therefore, Table 7.2 briefly summarises the review activities used for collecting evidence related to the feature of content validity.

Table 7.2

Evidence of content validity collected from review activity by the expert panel

Test Feature	Review Focus	Review Activity by Expert
Domain representation and domain relevance	Construct–item alignment	Correspondence between the descriptors, and the actual science practice and knowledge types elicited by the items
	TOS–item alignment	Correspondence between TOS specifications, and the actual science practice, knowledge types, and cognitive demand elicited by the items
Appropriateness of the test-development process	Item writing	Design of items adheres to established practices and guidelines
	Content	Accuracy of science content and alignment to curriculum
	Editorial	Clarity and technical quality
	Sensitivity and fairness	Insensitive use of language and contexts
	Verification of correct answer	Accuracy of intended answers
	Answer justification	Rationale behind responses from perspective of test-takers

Note. Adapted from Sireci (1998b) and Haladyna and Rodriguez (2013).

7.3.1 Participants and Materials

Two members from the expert panel in the pilot study constituted the expert panel for this stage. Both had affiliations with the national examination board in Singapore and agreed to participate because they were available. The materials consisted of the 34 new items that had not undergone review and pilot testing in the previous two phases. Additional items were necessary for two reasons. First, as reported in Chapter 5 (subsection 5.7.3), about 27% of the items from the 87-item Draft PSRT were found to be unsuitable. Second, the pilot study showed that the average completion time of the Draft PSRT was about 45 minutes, which was well below the one hour allotted. These post-pilot items were drawn up based on design parameters outlined in Chapter 4 (subsection 4.4.3) for consideration as potential items for inclusion in the PSRT.

7.3.2 Procedure and Data Analysis

All unreviewed items were sent by email to the two-member expert panel. Unlike the first review. reported in Chapter 5, the reviewers did not use the proforma to rate the items. Instead, in this small-scale review, they were instructed to draw upon their experiences from the previous review to raise issues about the characterisation of the items, such as the item-construct alignment and match between intended and likely cognitive demand, as well as the quality of the items (see Table 7.2 for the specific activities). They sent their feedback separately to me about a week later. There were a few rounds of clarification via email with each expert. Item amendments subsequently took place based on their input.

7.3.3 Results and Discussion

Comments from both panel members flagged concerns with about one-third of the items. The predominant issues raised centred on the design, clarity, and technical quality of the items. Acting on the feedback from the reviewers, amendments were made to improve the item stimulus and the stems in these items. To illustrate the extent of the changes based on reviewers' inputs, Figure 7.2 shows an item in its original state and the amended item.

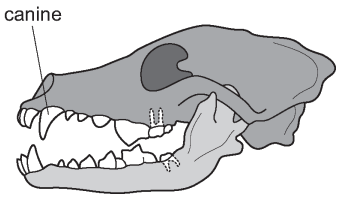
Original item	***BVIIQ1***
The shape of an animal's teeth can help it to feed. A lion has long, sharp canine teeth.  lion  canine skull of lion (a) How do its canine teeth help the lion eat its food?	The diagram below shows how a fork and a knife is used to eat meat. Study the diagram carefully.  using a fork and a knife when eating meat The lion has long, sharp canine teeth which it uses to eat the meat of other animals.  a lion  a skull of a lion (a) From the action of eating meat using a fork and knife, explain how the canine teeth help the lion eat its food.

Figure 7.2. The "before" and "after" changes to an item based on the two-member review.

As shown in Figure 7.2, the primary change made to the stem of the original item was the inclusion of additional information about eating meat using utensils. This modification was intended to draw pupil's attention to the action of feeding and resulted from concerns by reviewers that without appropriate scaffolding in the stem, pupils might not understand how they should respond.

Overall, the limited review provided some evidence that the new items met the requirements of domain representation and relevance. There was other evidence to suggest that some of the items reviewed, such as the "before" item shown in Figure 7.2, had construction and technical flaws such as ambiguity in the assessment objectives and stems with insufficient information for pupils to respond correctly. Had these issues not been flagged by the reviewers and addressed, these affected items could have impacted on the overall content validity of the PSRT.

The PSRT was assembled after the review, in preparation for the second validation stage of this phase. Accompanying this, the construct descriptors were updated, and a TOS drawn up, to ensure adequate and appropriate coverage of the learning outcomes. Other materials drawn up included a "table of revision history" to document the changes and historical development of each item in the PSRT. These two tables provided important content validity evidence for the interpretation of pupils' performance on the PSRT in the next stage. Therefore, their contents will be discussed in detail in a later subsection (subsection 7.4.2). The next section introduces the second validation stage of the main study.

7.4 Validation Stage 2

As summarised in Table 7.1, the goal of this stage was to obtain validity evidence from three validity sources – test content, internal structure and response processes – through the administration and analysis of the PSRT. To provide valuable validity evidence in this stage, analysis of the responses to the PSRT used similar psychometric and qualitative examinations to those adopted for the Draft PSRT. Inferences from the evidence gathered addressed two primary research aims.

The first research aim and its associated questions, which will be addressed in the present chapter, evaluated whether the PSRT measures a unidimensional construct. Establishing this claim meant performance could be confidently attributed and interpreted with respect to a singular construct. Evidence to support this claim came from the fit statistics produced from Rasch modelling, as addressed in section 7.5, as well as script and item examination.

The second research aim, which will be examined in the following chapter, was to investigate whether the construct measured was the postulated scientific reasoning construct. To this end, the analyses checked the degree of agreement between the observed difficulty estimates of items and their theoretically expected values according to five item features. Table 7.3 summarises these research aims, questions, and the methods to collect the evidence.

Table 7.3

Analytical methods to address two research aims and questions in validation stage 2

Chapter Focus: Research Aim	Research Questions	Analytical Methods
Chapter 7: Produce evidence that the PSRT is measuring a unidimensional construct	What are the item characteristics in the PSRT? How does the PSRT function regarding the interaction of the person–gender, person–school, and person–booklet in each item?	Examine Rasch fit indices; conduct within-person *t*-tests and binomial test from principal component analysis (PCA); qualitative analysis of pupils' performance in the PSRT items
Chapter 8: Examine whether the underlying construct measured in the PSRT conforms to the hypothesised construct of scientific reasoning	Do the item location and threshold estimates conform to expectation?	Analyse five key item features and qualitative responses to produce inferences about impact on location and threshold estimates; draw inferences from qualitative and quantitative methods

The following subsection provides an overview of this second validation stage by describing the participants, materials, and procedure. Thereafter, the data-analytical methods to address the first research aim will be discussed. This chapter concludes by reporting and discussing the findings from the analysis using the methods.

7.4.1 Schools and Participants

Six co-educational primary schools participated in the main study. Five of them – schools A to E – are designated "government" schools because they are fully funded by the Singapore government. The sixth (School F) is a "government-aided" school, as it receives additional funding from private sources such as its alumnus association. Unlike the other five schools, School F has greater autonomy to modify its curriculum and learning activities. Situated near residential areas across the island of Singapore, the six schools cater to the needs of their surrounding neighbourhoods. As such, the schools predominantly have pupils who are

of mixed abilities. Altogether, 431 eleven to twelve-year-old pupils from the schools participated. At the time of the study, all participating pupils had completed the mandatory end-of-primary-course national examinations, which included two science papers. As shown in Table 7.4, the number of participants from each school ranged from a low of 61 to a high of 91. Overall, the pupil sample consisted of slightly more boys than girls.

Table 7.4

Distribution of pupil sample in the main study by school and gender

Gender	School						
	A	B	C	D	E	F	Total
Girls	29	29	41	29	37	48	213
Boys	32	33	32	42	36	43	218
Total	61	62	73	71	73	91	431

7.4.2 Materials

PSRT: Like the Draft PSRT, the PSRT consisted of 11 blocks of questions. Similarly, the organisation of these blocks into five paper-and-pencil test booklets followed a balanced and incomplete sampling design. As shown in Table 7.5, there were 44–48 items per booklet.

Table 7.5

Arrangement of anchor block and ten question blocks across five test booklets, including the number of items in each block

Booklet	Blocks and No. of Items										Total
	AN	I	II	III	IV	V	VI	VII	VIII	IX	
1	11	6	8	10	9						44
2	11				9	9	8	10			47
3	11	6				9			13	8	47
4	11		8				8		13	8	48
5	11			10				10		8 8	47

Note. AN refers to anchor block, I to X are question blocks. Each block had two to four questions.

Sixty-three items of the 100-item PSRT required test-takers to produce constructed-responses that were subjectively scored (i.e., CRSS). See Table 7.6 for the exact distribution. In contrast, the total number of objectively scored items was considerably fewer. This combination of format types in the PSRT mirrored the Draft PSRT to capitalise on the relative advantages conferred by each format (Table 4.18 in Chapter 4). Appendices 16 and 17 show the items in the PSRT and the item format for each PSRT item respectively.

Table 7.6

Distribution of the 100-item PSRT by item format

Number of Items			Total
SROS	CROS	CRSS	
19	18	63	100

Tables of Specifications (TOS): There were two TOS to guide the distribution of marks in the PSRT. The first focused on the curriculum, while the second emphasised the knowledge types and science practices in the scientific reasoning construct.

Curriculum: Like the Draft PSRT, questions in the PSRT embedded the assessment of procedural and epistemic knowledge within the disciplinary content specified in Singapore's primary science curriculum. This approach ensured the assessment objectives drew on the taught curriculum (Nitko & Brookhart, 2007). Although there was no pre-specified topic coverage, this study nevertheless aimed to sample as many topics as possible from the curriculum. There were also no overlapping topics in each booklet. In the PSRT, topic selection came from each of the five curricular themes. Broad descriptions of these themes were given in the initial test design phase outlined in Chapter 4 (Table 4.3). As summarised in Table 7.7, about half of the marks available assessed outcomes from advanced topics found in the two themes of *Energy* and *Interactions*. Appendix 18 presents the specific mapping of PSRT items to each of five curricular themes.

Table 7.7

Mark distribution in the PSRT according to the disciplinary themes in the curriculum

	Disciplinary Themes in Curriculum					Total
	Diversity	Cycles	Systems	Energy	Interactions	
Number of Marks (%)	8 (7%)	29 (26%)	18 (16%)	18 (16%)	40 (35%)	113

Knowledge Types, Science Practices and Cognitive Demand: The 100-item PSRT carried a total of 113 marks. In accordance with the curriculum focus, about 80% of the marks were for the assessment of *content* and *procedural* knowledge. For a similar reason, items measuring the two science practices of *giving scientific explanation* and *interpreting and analysing data and evidence* were worth 75% of the available marks. The third practice of *designing and evaluating investigation* had a reduced focus in the test. About one-third of the marks went into items that tested the combination of *procedural knowledge* and the practice of *interpreting and analysing data and evidence*. Assessing learning objectives from these two areas within the same item permitted pupils to easily demonstrate understanding of experimental concepts as they reviewed data and evidence.

Few or no items assessed certain combinations of knowledge and practice (e.g., *content* and *designing and evaluating investigation*). These combinations required the coordination between advanced knowledge and abilities, which were beyond standard curriculum requirements and instructional exposure. See Table 7.8 for a summary of the assignment of marks across knowledge types, science practices, and cognitive demand.

Table 7.8

Mark distribution according to knowledge type, scientific practice, and cognitive demand

		Number of Marks in Scientific Practice			
Knowledge Type	Cognitive Level	Giving Scientific Explanation	Designing and Evaluating Investigation	Interpreting and Analysing Data and Evidence	Total (%)
Content	Low	21	0	2	23
	Medium	4	0	4	8
	High	4	0	0	4
				Total	35 (31%)
Procedural	Low	5	5	10	20
	Medium	0	8	13	21
	High	0	1	14	15
				Total	56 (50%)
Epistemic	Low	1	0	2	3
	Medium	0	7	1	8
	High	0	7	4	11
				Total	22 (19%)
Overall Total (%)		35 (31%)	28 (25%)	50 (44%)	113(100%)

Overall, the total marks assigned to each knowledge type and scientific practice fell within the targeted range specified in the test framework in Table 4.25, subsection 4.5.3 of Chapter 4. In keeping with the curriculum's emphasis on basic scientific understanding, about one-fifth of the marks belonged to low-demand items that assessed both *content knowledge* and the practice of *giving scientific explanation*. In contrast, the marks allocated to the combined assessment of *procedural knowledge* and *interpretation and analysing data and evidence* were mostly spread evenly among items from each of the three cognitive levels. Correspondingly, marks allocated to the assessment of epistemic knowledge belonged largely to medium- or high- demand items, as pupils had less instructional exposure to this knowledge type. The detailed TOS is shown in Appendix 19.

Table of Revision History: A table of revision history was drawn up to record the origins of, development of, and amendments made to the items in the PSRT. As explained in the first validation stage, feedback from two expert reviews and the findings of the pilot study guided the assembly of the 100 items in the PSRT. Approximately one-third of these items were newly designed after phase III (pilot study). As such, these items did not undergo the first expert review; nor were they psychometrically tested. Instead, they were reviewed by the two-member expert panel. The remaining 64 items originated from the Draft PSRT. About half of these were amended based on the findings of the first review and the pilot study. More details about the revision history of each item in the PSRT is provided in Appendix 20. Table 7.9 summarises the historical development of the items.

Table 7.9

Summary of the historical development of the 100 items in the PSRT

Development	Outcomes	Number of Items	
		Excluded from PSRT	Included in PSRT
87 items from Draft PSRT after an expert review (phase II) and a pilot study (phase III)	Unchanged		35
	Modified		29
	Rejected	23	
36 post-pilot items after expert review	Unchanged		22
	Modified		14
	Rejected	0	
Total number of items		23	100

7.4.3 Procedure

Together with an assistant, I visited every school to administer the five one-hour PSRT booklets, which were professionally printed. Before the visits, the schools collected signed parental or guardian consent forms on my behalf. Before the start of the test, participating pupils were seated individually in rows to reduce unwanted interaction. Following a similar

procedure in the pilot study, I then gave a pre-test briefing to the pupils. In the briefing, I provided a short introduction to the research and shared the testing objectives and purposes. I informed the pupils that their participation was voluntary and that they could withdraw from the test at any time without being penalised.[26] I also assured participants that individual results on the test would not impact their marks at school, as schools and parents would not be informed of the results. Finally, I highlighted the duration and other test instructions, such as checking that all booklets had printed pages.

I repeated the pre-test briefing in all schools to ensure that there was consistent messaging. During the briefing, my assistant assigned one of the five booklets of the PSRT (booklets 1 to 5) on a rotation basis to every pupil and provided the necessary stationery. As shown in Table 7.10, the total number of each booklet attempted was 86 or 87 across the six schools. The two rightmost columns in the table indicate that an almost equal number of boys and girls attempted every one of the five booklets.

Table 7.10

Distribution of five-booklet PSRT by school and gender

Booklet	School							Gender	
	A	B	C	D	E	F	Total	Girls	Boys
1	13	13	11	15	15	19	86	44	42
2	13	13	15	15	15	15	86	43	43
3	13	12	15	13	15	18	86	44	42
4	11	12	16	14	14	19	86	40	46
5	11	12	16	14	14	20	87	42	45
Total	61	62	73	71	73	91	431	213	218

From the start to the end of the one-hour test, my assistant and I walked around the classroom to ensure prompt attention to queries. After the administration of the PSRT, I conducted a preliminary round of marking. In a similar way to the pilot study, an intern

[26] Like the pilot study, there were no instances of pupil withdrawal in the main study.

teacher enrolled in the PGCE course at the University of Oxford reviewed the marking. She initially marked all scripts and flagged those cases where the marks awarded were not consistent with those given in the preliminary round. These cases involved 12 items from nine questions spread across the anchor block and six other question blocks. She grouped these cases by question and discussed them with me.

During the discussion, the scoring guide specific to every affected question was considered in turn to understand how the differences in rating could have occurred in these cases. The intern then used the resulting adjustments to the scoring guide to re-mark all flagged cases. Subsequently, I conducted a final check by sampling every tenth script from every school. Although I found no instances of inconsistent or incorrect application of the scoring guide, I did identify some mark-recording issues from data entry errors. The intern addressed these issues and conducted final checks on all scripts to ensure there were no further cases of inaccurate recording.

7.5 Data Analysis Strategy

This section begins with an overview of the analytical approach used in the second stage to address each of the three research aims. Thereafter, it discusses the methods used to produce evidence to achieve the first research aim. Details of the procedures used to answer the remaining aims, along with the relevant results, will be provided in the next two chapters. Overall, the data analysis mirrored and extended the pilot study, which used Rasch Measurement Theory (RMT) to examine pupils' performance in the test. Therefore, a similar set of test and fit indices was produced, using the software RUMM2030 (Andrich et al., 1997-2012b).

One extension between the pilot and current study was the attempts in the current study to remediate items with poor or unacceptable model fit. Such remediation attempts were possible because the larger sample size enabled greater confidence in the accuracy of the parameter estimates and fit statistics. These post-hoc improvement and remedial measures resulted in the sequential examination of items, as explained later in this section. Also

incorporated into the remediation were qualitative inferences drawn from item examination and pupils' responses in the scripts. The next subsection discusses the procedures to obtain evidence to answer the first research aim and questions. A later subsection uses item exemplars to demonstrate the use of mixed methods for item remediation.

7.5.1 Data Analysis

Initial Calibration: First, an initial estimation of parameters and fit statistics took place without constraints on the item location estimates. The item-level fit statistics produced after the initial estimation were like those described in the pilot study, including an additional procedure. This procedure involved an examination of the person–item residual correlations for violations of local independence.

There are two primary sources of local dependence between items: trait and response (Hagquist et al., 2009; Marais & Andrich, 2008). The former usually manifests in a test whose total score comes from the summed responses to items organised in sub-scales for measuring different aspects of a larger construct. It can also occur when subsets of items share common features, such as testing objectives or item anatomy (e.g., stimulus and/or content). Response dependence is statistically induced when an item influences the response to a later item, such as when an item gives a clue to the answer of another item.

Unlike trait dependence, response dependence is not a feature intentionally built into a test. Still, both forms of dependence spuriously inflate the PSI, which, as discussed in the previous chapter, is a measure of test reliability (Andrich, 2006a). Unfortunately, there is as yet no empirical approach to distinguish between the two sources of violations in affected items, and inferences must be made by examining the test and item format to understand the nature of the local dependence (Marais & Andrich, 2008). Table 7.11 provides a summary of these examined indices.

Table 7.11

Summary of the fit statistics examined in the PSRT

Focus of Analysis	Fit Statistics
Test	Mean and standard deviation of the fit residual of items in the test
	Mean and standard deviation of location estimates of persons
	Mean and standard deviation of the fit residual of persons
	Chi-square test
	Person Separation Index (PSI)
Item	Fit residual of each item and chi-square test
	Analysis of variance test
	Residual correlations
Person	Location estimates of sample
	Fit residual of each person

Item Diagnosis and Remediation: Initial fit statistics diagnosed issues such as disordered thresholds in polytomous items, differential item functioning (DIF), or local dependencies between items. At this juncture, person–fit residuals outside of the acceptable range were also identified. However, these pupils were not removed from the sample, as it was unclear how the item irregularities had impacted on these estimates.

The next step of the analysis aimed to resolve the issues in these flagged items and review whether items with acceptable fit statistics could be further improved. Besides the use of psychometric data, improvement strategies also drew on theoretical perspectives and pupils' written responses. These corrective efforts regarding items with statistical anomalies resulted in one or more of three possible actions: deletion, merging with other item(s) to form "testlets", or "DIF-split". Testlet formation is a common remedy to account for local dependency in the model estimation. It involves grouping items that display residual correlations greater than 0.3 into a higher-order polytomous item (see e.g., Christensen, Makransky, & Horton, 2016). "DIF-split" involves resolving items affected by DIF into separate items for each affected group (e.g., "splitting" an item by gender results in two items: one for boys only and one for girls only). As descriptions of other fit indices were

provided in the previous chapter on the pilot study (subsection 6.6.3), Table 7.12 provides a summary of all follow-up corrective measures that were undertaken.

Table 7.12

Summary of corrective measures for statistical anomalies in item-fit statistics or tests

Fit Statistics/ Statistical Test	Statistical Anomaly	Description of Corrective Measure
Analysis of variance test	Uniform and non-uniform DIF: statistically significant main effect and trait-group interaction	"Split" affected item into two separate items for the subgroups; check if resolving DIF-affected items improves model fit
Threshold order	Disordered thresholds present in polytomous items	Collapse affected adjacent response categories; check if resolution improves model fit
Residual correlations	Correlations between items above 0.3	Group affected items into a higher-order polytomous item; check if grouping improves model fit

Examination of other items not flagged as statistically problematic aimed to uncover potential areas for improvement, resulting in two possible outcomes. These items could be retained without changes or marked for future amendments. After every resolution, the remaining items underwent recalibration. Since misfitting items were present in almost all blocks, item examination began with the anchor block of items in the order presented to the pupils. This block was included in all booklets so its questions received responses from the entire sample. As the response number across all blocks was similar, subsequent blocks selected for analysis followed a numerical sequence, starting with block I.

This remediation–recalibration iterative process was repeated until resolution attempts on the last affected item in the last booklet were completed. Although not all attempts were successful, the revised version of the PSRT (Revised PSRT) contained the most updated fit statistics at this juncture, which ensured that more robust inferences could be made about

the item and person fit. Figure 7.3 presents a flow-chart summarising the steps in the iterative examination of item fit.

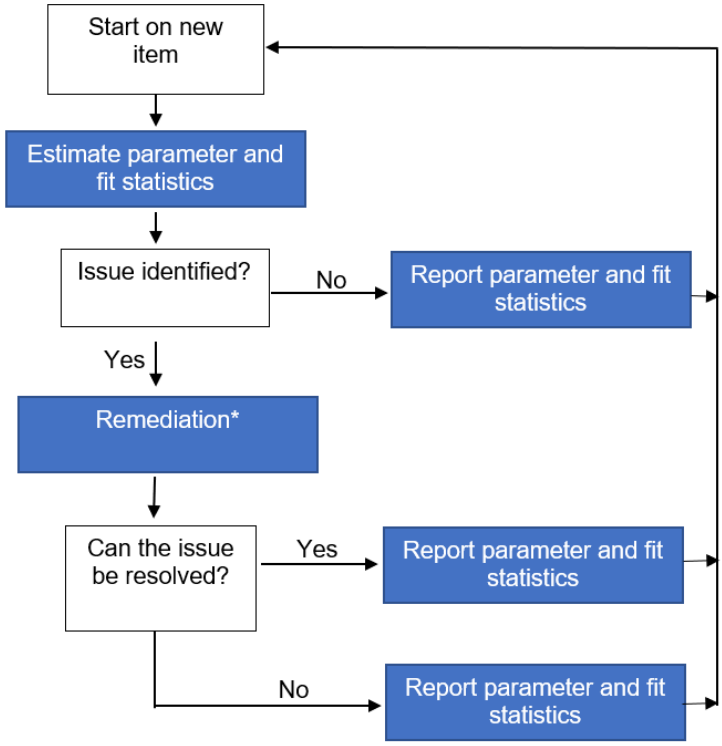


*Figure 7.3. Flow-chart of the item diagnostic process comprising an iterative calibration and remediation loop. *Depending on the item, there could be multiple remediation attempts.*

Other Procedures: The model estimates and fit statistics from the Revised PSRT underwent further procedures to check performance differences between genders and among schools. Also examined was the functioning of the five booklets. Two additional indices considered at the booklet level were Cronbach's α and the item loadings derived from a principal component analysis (PCA) of person–item residuals. Procedurally, the computation of these indices required complete data, which in this matrix-design study was available only for booklets. Like the PSI, Cronbach's α functions as a measure of score reliability. Larger differences between the two reliability indices indicate more mistargeting between item difficulties and the sample's abilities (Andrich, Sheridan, & Luo, 1997-2012a).

The item loadings from a PCA were used to produce evidence that the responses to the questions in the individual booklets did not breach the local independence assumption (Smith, 2002). To test the assumption, the person–item residuals first underwent a PCA

analysis to compute loadings for identifying the two most divergent item subsets. Residuals represent the unmodeled aspect of the data, and the presence of patterns among the person–item residual correlations between items signals potential multidimensionality (Smith, 2002). Statistically, evidence of systematic relationships in the residual variance is indicated by loadings of residuals on the first principal component in a PCA outside the range of ±0.3 (Linacre, 1998a, 1998b; Pallant & Tennant, 2007). Also, as loadings in the same direction (positive or negative) potentially indicate the measurement of similar competencies, items with loadings greater than ±0.3 in the positive and the negative directions were identified as two subsets that potentially assessed the most disparate competencies.

Person ability estimates were calculated for the two subsets for each pupil and these estimates were compared using within-person *t*-tests to evaluate whether there was a significant difference between the estimates for every pupil. Of interest was the proportion of *t*-tests found to indicate a statistically significant difference between the estimates since the responses to any subset of items from a unidimensional scale should produce similar person ability estimates. Supporting the inferences were outputs of binomial tests that produce 95% confidence intervals (C.I.) around the observed proportions of significant *t*-tests. There is a stronger claim for unidimensionality if the lower bound of the C.I. overlaps 5% (Hagell, 2014; Tennant & Pallant, 2006).

7.6 Results

This section begins by reporting the initial results of the Rasch analysis performed on the 100-item PSRT. These results consisted of the overall test-fit statistics and a description of the statistical anomalies detected in the items from the initial analysis. Next to be described are the results of item examination and corrective measures enacted on problematic items, which culminated in an 86-item revised version of the PSRT. Thereafter, the discussion reports the fit statistics of the revised test and contrasts these estimates with the original test. Additional results report and compare the performance between genders and among

schools, as well as the fit statistics for the booklets. This section concludes with a discussion of the results of further investigations to examine scale dimensionality.

7.6.1 Overall Fit Statistics of the PSRT

Overall Fit: The fit of the data produced a significant total-item chi-square (538.96, *df* = 400, *p* = <.001). This result signalled the presence of misfit between the data and the model. The mean fit residual of items and persons were -0.198 (*SD* = 1.339) and -0.190 (*SD* = 0.827) respectively. When compared to the expected value of 0, these fit residual mean values suggested the presence of some model-fit issues among the items and sample.

Reliability and Targeting: The PSI was 0.866, indicating a very high proportion of true variance in the person estimates for the sample relative to the observed variance of these estimates. By extension, this PSI value also suggested that the error variance was low, and thus by implication, the fit statistics were well powered to detect misfit. The mean person location estimate was 1.174 (*SD* of 1.019), and the mean item location estimate was zero (*SD* of 1.381). The higher mean location estimate of the sample relative to the items suggested that there was some mistargeting.

From the person–item threshold distribution in Figure 7.4, the mistargeting occurred mostly at the lower end of the scale. For instance, there were only two pupils located at location estimates below -2 logits, whereas there were seven items, all of which were dichotomously scored, with difficulty location estimates below this point on the scale.

7.6.2 Item and Person Fit Statistics of the PSRT

Item-Level Fit Statistics: Results from the initial Rasch analysis suggested that 60% of the items had acceptable item fit. Among these, 57 were dichotomous items and three were polytomous items. The remaining items demonstrated four basic forms of statistical misfit. The first was a uniform gender DIF effect in an item (see Figure 7.5), which was significant after Bonferroni adjustment for multiple comparisons; $F_{1, 4}$ = 15.316, *p* = <.001. All pupils in the sample attempted this item, as it was part of a multi-item question in the anchor block.

An inspection of its item characterisation curve (ICC), shown in Figure 7.6, indicated that the boys consistently performed better than the girls, with the same overall ability estimate across the full range of ability.

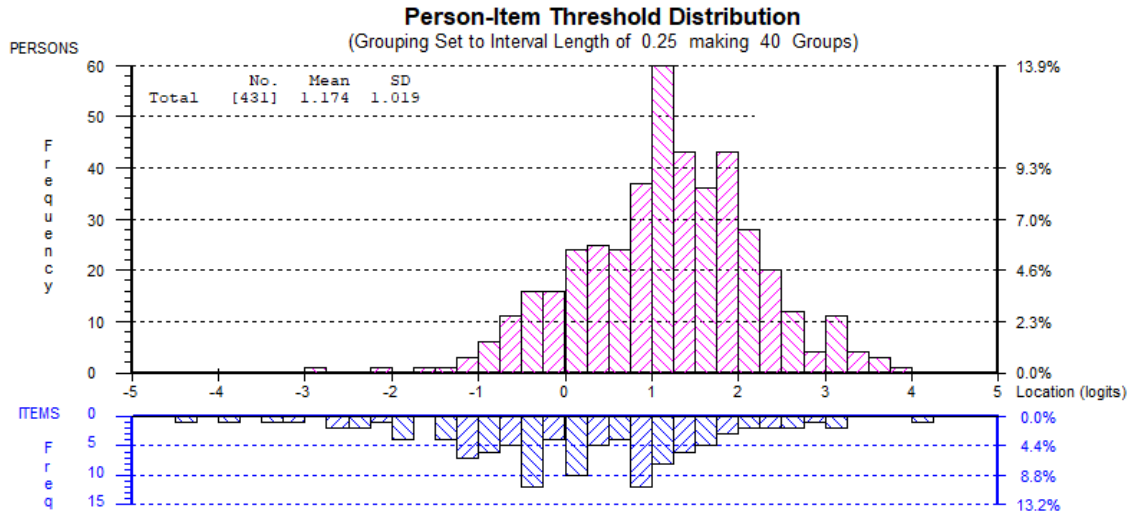


Figure 7.4. Person–item threshold distribution of the 100-item PSRT.

AQ3

	Experiment			
	A	B	C	D
Length of string (cm)	25	25	50	75
Mass of weight (g)	100	50	100	100
Time for 10 complete swings (s)	10.0	10.0	14.2	17.4

(e)(ii) Use Joe's table of results to predict the times for 10 complete swings in experiment F.
Write your answer in the table below.

	F
Length of string (cm)	100
Mass of weight (g)	100
Predicted time for 10 complete swings (s)	

[1 mark]

Figure 7.5. An item requiring pupils to extrapolate from existing data. Credit was awarded for any numerical answer from 18 to 25 seconds.

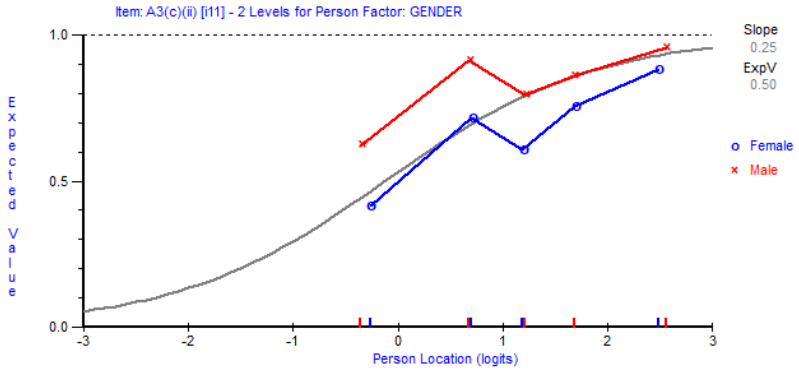


Figure 7.6. ICC of item AQ3(e)(ii) showing the systematic DIF between genders. The plots in red and blue are for the performance *of boys and girls respectively.*

The second irregularity was threshold disorder, which was seen in eight two-mark items and the only three-mark item in the PSRT. See Figure 7.7 for the category probability curve of the three-mark item.

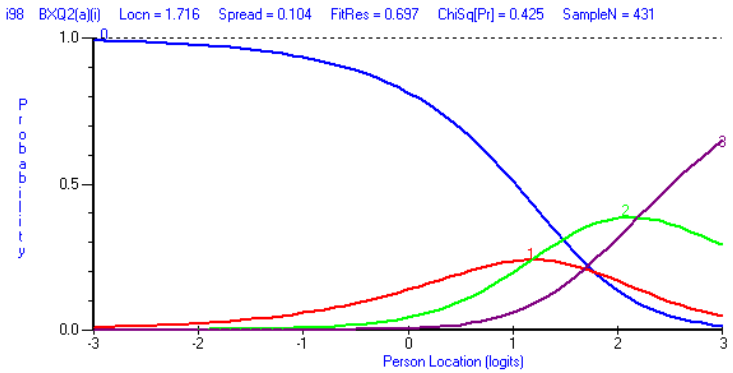


Figure 7.7. CPC of item BXQ2(a)(i) demonstrating the disordered thresholds. The scoring (response) category of one mark did not function as expected.

The third irregularity was observed in seven items with fit residuals outside of the ±2.5 range. Two among these items showed highly negative fit residuals and their ICC plots also indicated that these items were over-discriminating relative to all items. In contrast, the ICCs of the remaining five items indicated that they were under-discriminating relative to all items. Additionally, the responses to one of these five items also demonstrated a statistically significant chi square probability. See Figure 7.8 for the ICC of this item.

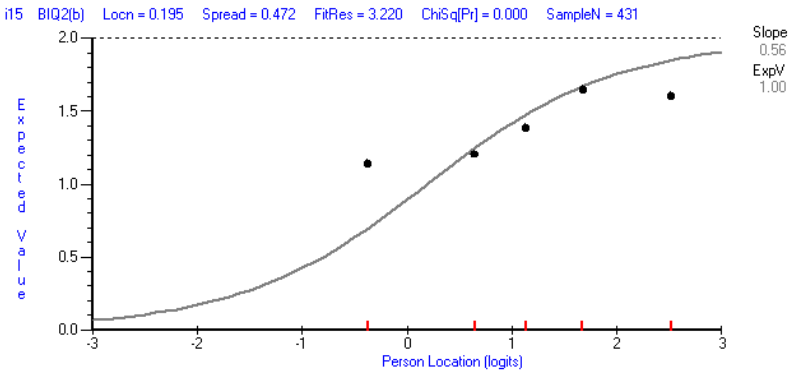


Figure 7.8. ICC of item BIQ2(b). The responses demonstrated an under-discriminatory fit to the model with a fit residual of 3.220 and a statistically significant chi-square probability.

The last irregularity – local dependency between questions – affected slightly more than one-fifth of the remaining test items. An example of an item (Figure 7.9) demonstrating this psychometric flaw will be shown and discussed in a later subsection. Altogether, the initial Rasch analysis flagged 40 problematic items in the PSRT. Table 7.13 summarises the number of affected items by the observed irregularities. Appendix 21 provides the fit indices of all items in the PSRT from this initial analysis.

Table 7.13

Summary of item irregularities in the PSRT from the initial Rasch analysis

No.	Item Irregularity	No. of Items
1	DIF	1
2	Residual correlation with another item in the same question greater than 0.3	23
3	Threshold disorder	9
4	Item fit residual not within ± 2.5	4
5	Item fit residual not within ± 2.5 and statistically significant chi-square probability	1
6	Item fit residual not within ± 2.5 as well as residual correlation with another item in the same question greater than 0.3	2
	Total	40

Person-level Fit Statistics and Tests: The person-fit statistics from the initial analysis showed four persons from the sample with positive fit residuals above the acceptable criterion of +2.5. As explained in the earlier data analysis section, this small group was kept in the sample for analysis as it was unclear how unresolved issues with a substantial proportion of items influenced the person measurement.

7.6.3 Review and Resolution Outcomes

This subsection describes the review and resolution outcomes of the 100 items in the PSRT. Table 7.14 summarises these outcomes, which produced an 86-item Revised PSRT.

Table 7.14

Formation of the Revised PSRT from the outcomes of review and corrective measures

Outcomes	Number in PSRT		Number in Revised PSRT	
	1-mark	2-mark	1-mark	2-mark
Retained without change	47	0	47	-
Recommended for change	30	3	30	3
Statistically corrected	4	9	1	5
Unsuccessful corrections as “testlet”	4	0	-	-
“DIF-split”	3	0	-	-
Total	88	12	78	8

Among the items in the revised test, 47 received no further recommendations for alterations while 33 items were identified for future improvements. These recommendations ranged from simple amendments in phrasing to changes in item format. Six other items resulted from the successful resolution of irregularities observed in 13 items from the original test. Although the initial analysis indicated the presence of uniform gender DIF in one item, DIF appeared in two more items during the iterative process of examining and diagnosing items. Overcoming these issues required a combination of corrective actions, including DIF-splitting the item by two groupings (i.e., school and gender) instead of one. Although these measures were successful in removing the DIF effects, their causes remained unknown.

Therefore, these three items were removed from the Revised PSRT. Besides the three DIF-affected items, four other items were deleted. This decision was taken because attempts to resolve dependencies in these items by forming testlets were unsuccessful. Moreover, there were other items with similar testing objectives to these deleted items that showed good model fit, and so minimal loss in construct coverage. Appendix 21 displays the final fit statistics of the resulting 78 one-mark and eight two-mark items in the Revised PSRT.

The following discussion focuses on four items from the third question of block II (see Figure 7.9) as exemplars to demonstrate how the combined deployment of quantitative and qualitative methods impacted on their appraisal and remediation. It begins by outlining the intended assessment objectives of the exemplar items. See Table 7.15 for the item characterisation scheme. Subsequently, drawing on fit statistics and plots from the Rasch analysis, inferences about the initial model fit of individual items are made. By investigating the efficacy of the various strategies to address identified fit issues, this discussion concludes with recommendations for item improvements.

Table 7.15

Item characterisation scheme of the third question of block II (BIIQ3)

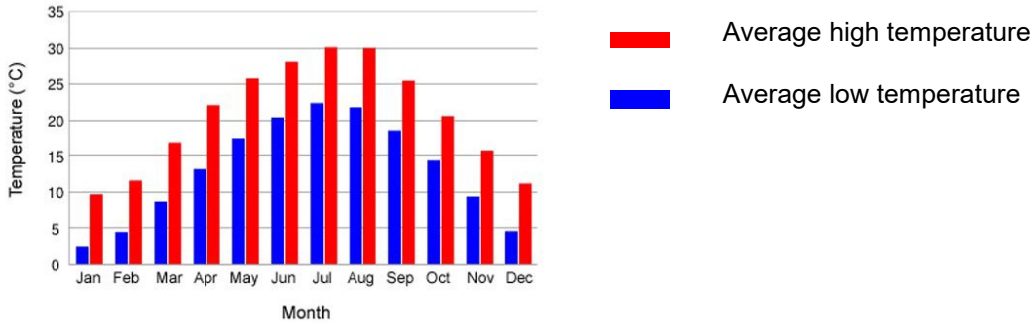
Framework Category	BIIQ3(a)	BIIQ3(b)(i)	BIIQ3(b)(ii)	BIIQ3(b)(iii)
Knowledge Type	Procedural	Epistemic	Epistemic	Epistemic
Practice	3	2	2	2
Sub-practice and levels*	1.2	1.4	1.4	1.4
Cognitive Demand	Medium	High	High	High
Total Mark	1	1	1	1
Item Format	SROS	SROS	SROS	SROS

*Note. Practice 2 – designing and evaluating investigative approaches, practice 3 – interpreting and analysing data and evidence. SROS stands for selected-response objective scoring. *sub-practice and levels are elaborated in the text.*

The animal P shown below is found only in forests of central China.



The graph shows the yearly average high and low temperatures of forests in which P is found.



Animal P is transported to other countries in specially equipped aeroplanes.

(a) From the graph above, what should the temperature range be inside the aeroplanes if the animals are transported in April? Tick (✓) the box with the correct answer.

25 – 30°C ☐ 20 – 25°C ☐ 15 – 20°C ☐

[1 mark]

Breeding a male P with a female P to get a baby is not easy.

(b) The questions below are about conditions affecting the breeding of animal P. Which of the questions could be answered with a scientific experiment? Give your answers by ticking (✓) either **Yes** or **No**.

Question	Yes	No
(i) Will the amount of food each animal P eats affect breeding?		
(ii) Will the playing of music to animal P affect breeding?		
(iii) Will the feelings of the male P for the female P affect breeding?		

[3 marks]

Figure 7.9. A four-item question exemplar from the PSRT.

Exemplars: The first item of the question in Figure 7.9 tested pupils' abilities to interpret and read from a bar chart. These procedural competencies came under the practice of *interpreting and analysing data and evidence* and targets the second level of sub-practice 1 (i.e., 1.2), which is to *identify less straightforward patterns in data presented in various formats.* The second to fourth items assessed the practice of *designing and evaluating investigative approaches* and required an understanding of the evidence needed to make

claims. This aspect of epistemic knowledge is the sub-practice of planning effective investigation and targets the competency of question formulation, i.e. 1.4 (see Appendix 5 for the complete description of both sub-practices).

Analysis of Exemplar Items: The initial Rasch analysis indicated that local dependence affected three of the four items as observed from their residual correlations. An examination of the items suggested the source of dependence could be trait-related, as they shared the same stimulus and testing objectives. Table 7.16 displays the correlation values and indicates the affected items.

Table 7.16

Residual correlations in the initial analysis between items

	BIIQ3(a)	BIIQ3(b)(i)	BIIQ3(b)(ii)
BIIQ3(b)(i)	-0.035		
BIIQ3(b)(ii)	0.008	0.021	
BIIQ3(b)(iii)	0.006	0.285	0.360*

*Note. *Residual correlations of above 0.3.*

The results from the initial analysis (run 1, summarised in Table 7.17), also showed that, apart from the potential local dependence, the four items had acceptable model fit. The location estimate of the first item – BIIQ3(a), which had a three-option selected-response format – was lower (i.e., 0.684) than the mean person location estimate, indicating that this would probably be an easy item for the average pupil ability in the sample (mean ability estimate = 1.184, *SD* = 1.080). Its ICC in run 1 (see Figure 7.10) showed reasonable fit, although the performance of pupils above 1.7 logits deviated from model expectations, as the item showed no discrimination between the two most able class interval groups.

Table 7.17

Results of the four items from an initial Rasch analysis (run 1) on the PSRT

Run Strategy	Item	PSI	Item δ	SE	Fit Residual	χ^2	*p*-value
1 Discrete	BIIQ3(a)	0.861	0.684	0.172	0.744	3.804	0.433
items	BIIQ3(b)(i)		-0.496	0.208	0.365	2.373	0.667
	BIIQ3(b)(ii)		0.931	0.17	-1.575	9.786	0.044
	BIIQ3(b)(iii)		0.737	0.172	0.309	6.739	0.150

Note. Item δ is the item location estimate, SE is standard error, χ^2 is chi-square value.

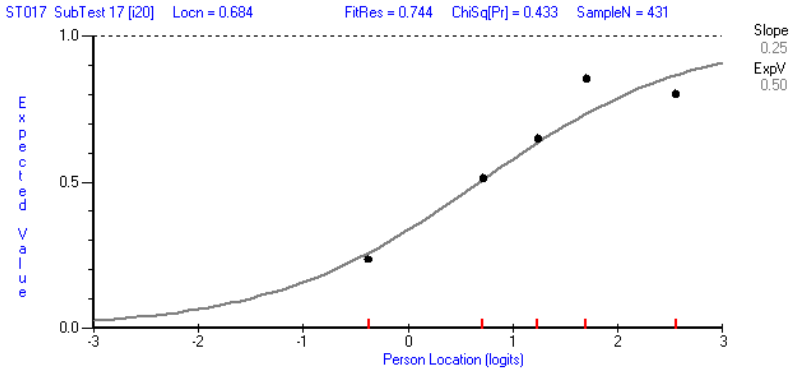


Figure 7.10. Item characteristic curve of item BIIQ3(a) in run 1.

From the results, possible improvements to the item included changing to a four-option selected-response format (e.g., by adding 5 – 10°C as a fourth option) or amending the item to demonstrate reasoning in constructed responses. Subsequent analyses from runs 2 to 4 investigated three combinations of testlet formation among the remaining items to remove the observed local dependence. As shown in Table 7.18 and Figure 7.11, a three-mark testlet in run 2 produced acceptable fit statistics but demonstrated disordered thresholds.

Table 7.18

Remediation outcomes of three run strategies (runs 2–4)

Run Strategy	Item	PSI	Item δ	SE	Fit Residual	χ^2	*p*-value
2 Form as one item	BIIQ3(b)(i) and BIIQ3(b)(ii) and BIIIQ3(b)(iii)	0.860	0.424	0.093	1.195	4.814	0.307
3 Form as one item	BIIQ3(b)(i) and BIIQ3(b)(ii)	0.861	0.22	0.131	-1.017	5.206	0.267
4 Form as one item and delete third item	BIIQ3(b)(i) and BIIQ3(b)(ii)	0.860	0.234	0.13	-0.591	3.603	0.462

Note. Item δ is the item location estimate, SE is standard error, χ^2 is chi-square value.

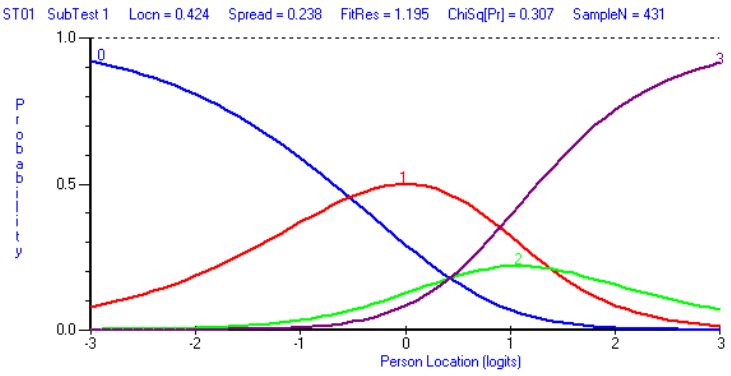


Figure 7.11. CPC with disordered thresholds obtained in run 2.

In run 3, combining the second and third items produced ordered thresholds. Further deletion of the last item in run 4 saw an improved fit, as the fit residual markedly dropped. Visually, the observed means of the class intervals in the resulting ICC in run 4 were closer to the model expectation than those in the ICC from run 3. Figures 7.12A to 7.12D show the CPCs and ICCs of the resulting items in runs 3 and 4.

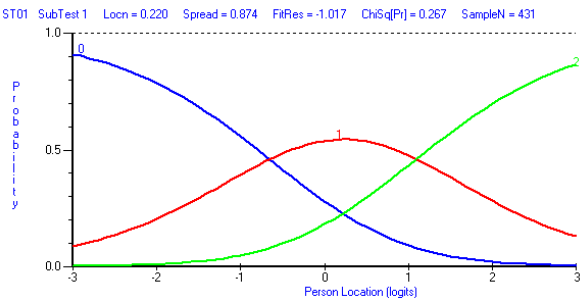


Figure 7.12A. CPC in run 3.

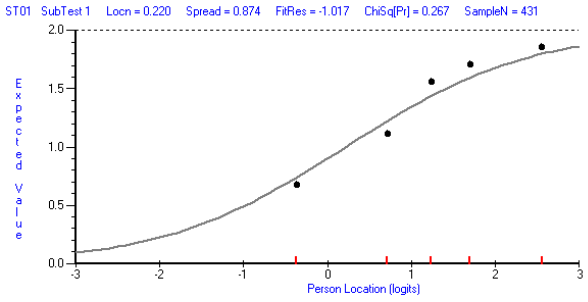


Figure 7.12B. ICC in run 3.

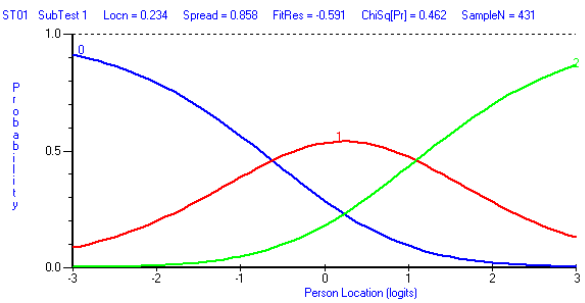


Figure 7.12C. CPC in run 4.

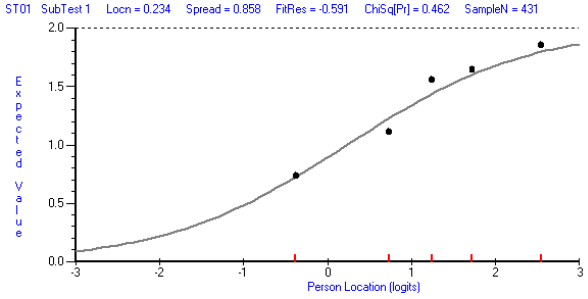


Figure 7.12D. ICC in run 4.

Runs 5 and 6 explored two other ways of combining the items. As shown in Table 7.19, combining the second and fourth items in run 5 produced ordered thresholds located at approximately 0.1 and 0.4 logits. Run 6 produced a testlet with disordered thresholds formed between the third and fourth items. Figures 7.13A and 7.13B show the CPCs of runs 5 and 6 respectively.

Table 7.19

Remediation outcomes of two run strategies (runs 5 and 6)

Run Strategy	Item	PSI	Item δ	SE	Fit Residual	χ^2	*p*-value
5 Form as one item	BIIQ3(b)(i) and BIIIQ3(b)(iii)	0.860	0.289	0.123	1.297	5.653	0.227
6 Form as one item	BIIQ3(b)(ii) and BIIIQ3(b)(iii)	0.860	0.818	0.11	0.775	6.084	0.20

Note. Item δ is the item location estimate, SE is standard error, χ^2 is chi-square value.

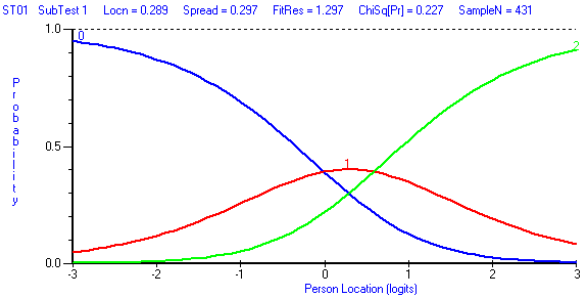


Figure 7.13A. CPC in run 5.

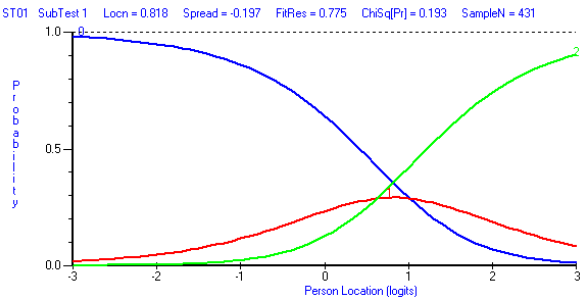


Figure 7.13B. CPC in run 6.

Jointly, the results of the six runs suggested that pupils who obtained credit in either the second or third item, in combination with the last item, might not have fully understood the concept of a research question that is amenable to investigation. However, it was possible that pupils misunderstood the concepts assessed in the last item, as the focus on the animal's emotional state was arguably unrealistic in an experimental context. Still, there was strong evidence from the results in runs 3 and 4 that pupils who obtained credit in *both* the second and third items understood the scientific concepts assessed. Significantly, removing the last item in run 4 did not detract from the value of the testing objective in the remaining testlet, but instead improved its model fit. Overall, the results of all runs supported the inference that appropriate remediation had been applied to address the local dependency obtained in the initial analysis. This remediation entailed the merger of the second and third items as a two-mark testlet and the deletion of the fourth item.

The preceding discussion in this subsection illustrates how decisions about item retention, removal, or modification were made based on the Rasch fit indices from iterative runs. Similar examination and review processes took place with the remaining items in the original 100-item PSRT to obtain a revised test with optimal item and person fit statistics. Given so many new items were added to the PSRT after the pilot, the fit statistics for this revised test provide stronger evidence to examine the unidimensional claim of the test, which was the research aim of this chapter. The next subsection describes the fit statistics and parameter estimates obtained from this Revised PSRT.

7.6.4 Revised PSRT

Overall Fit: In general, the overall fit statistics for the 86-item Revised PSRT changed only marginally from the original version. For instance, there was a slight drop in the PSI from 0.87 to 0.85, which was consistent with the resolution of local dependence between some items. The mean fit residual and *SD* for items were 0.162 and 1.200 respectively, which were closer to the expected values of 0 and 1 than the corresponding values in the PSRT. The same observation was true for the value of the mean fit residual of persons, which was -0.184 with an *SD* of 0.771. There was also an approximately similar total-item chi-square to degrees of freedom ratio (500.58/371 = 1.349) to that from the original test. Jointly, these results suggested there were smaller levels of item and person misfit in the Revised PSRT.

Reliability and Targeting: The mean location estimate of the pupils at 1.184 (*SD* = 1.080) increased slightly from that obtained in the original version. Its value showed that the sample would, on average, have found the test somewhat easy. As shown in Figure 7.14, there was minor mistargeting on both ends of the scale, there were some items below -3 logits and above 4 logits where there were no pupils estimated to be located.

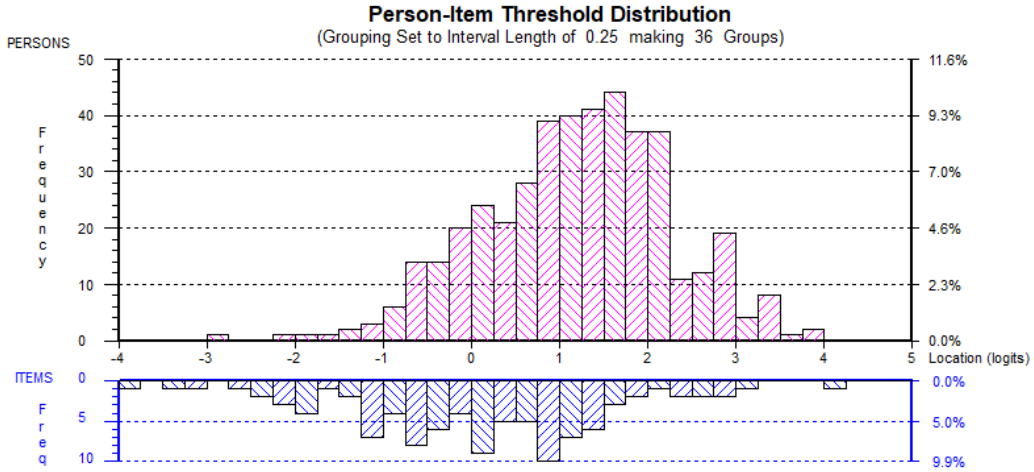


Figure 7.14. Person–item threshold distribution of the Revised PSRT.

Performance and Targeting by Gender: In the sample, the mean abilities of girls and boys were 1.210 (*SD* = 1.03) and 1.158 (*SD* = 1.13) respectively. However, the ANOVA, $F_{1, 429}$ = 0.245, p = .621, found no statistically significant differences in performance. As seen in Figure 7.15, there was a large degree of overlap in the distribution of performance by both male and female pupils.

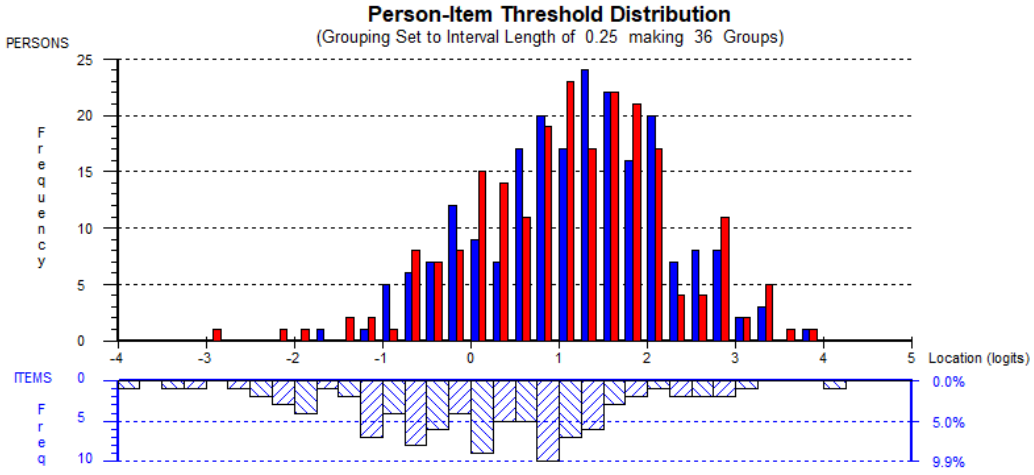


Figure 7.15. Person–item threshold distribution by gender of the Revised PSRT. *Blue and red columns represent the location estimate distribution of girls and boys respectively.*

Performance by School: There were statistically significant differences ($F_{5, 425}$ = 5.526, p < .001) in the performance of pupils among the six schools. The mean abilities of pupils in four of the schools were below the sample mean with the lowest recorded from school A at 0.837 (*SD* = 1.32). Pupils from school F had the highest mean performance at 1.503 (*SD* = 0.74). As Levene's test was significant for equality of variance (F = 6.965, p < .001), post-hoc analyses assuming unequal variances were applied, which showed significant differences between the means of three schools (A, B, and E) and school F. See Table 7.20 for the confidence intervals of these mean comparisons.

Table 7.20

The 95% confidence intervals of pairwise differences in mean ability estimates for the different schools

School	Ability Estimate		95% Confidence Intervals of Pairwise Differences				
	Mean	*SD*	A	B	C	D	E
A	0.837	1.32					
B	0.900	1.25	[-.74, .62]				
C	1.180	1.00	[-.95, .26]	[-0.86, 0.30]			
D	1.497	0.95	[-1.26, -.06*]	[-1.17, -0.02*]	[-.79, .16]		
E	1.016	1.09	[-.80, .44]	[-.71,.48]	[-.34, .67]	[-.02, .98]	
F	1.503	0.74	[-1.21, -.12*]	[-1.1, -.08*]	[-.73, .09]	[-.41, .40]	[-.92, -.05*]

*Note. * indicates difference in means was significant using Dunnett's C method at .05 significance level.*

These results suggested there was variable performance among the school sample, which was unsurprising since school and pupil selection was based on convenience sampling. As explained in Chapter 6, a sample with a broad range of abilities would have improved the accuracy of the parameter estimates of the test. Although it was not an aim of the research to account for between-school performance, it was nevertheless notable that pupils with the highest mean ability estimate belonged to the only government-aided school in the study.

Functioning by Booklet: The mean difficulty of the booklets ranged from a low of 1.051 in booklet 2 to a high of 1.344 in booklet 5 (see Table 7.21). Overall, the differences in difficulty among booklets were not statistically significant ($F_{4, 426}$ = 1.062, p = .375). Post-hoc tests to evaluate the pairwise differences among the mean difficulty assumed equal variances as there was an insignificant result from Levene's test for equality of variances ($F_{4, 426}$ = 1.566, p = .180).

Table 7.21 provides details of the pairwise comparisons of means. As all the confidence intervals of the mean differences contained zero, it can be reasonably concluded that the booklets had an approximately equivalent level of difficulty.

Table 7.21

The 95% confidence intervals of pairwise differences in mean difficulty estimates for the different booklets

Booklet	Difficulty Estimate		95% Confidence Intervals of Pairwise Differences			
	Mean	*SD*	1	2	3	4
1	1.126	1.19				
2	1.051	1.00	[-.38, .53]			
3	1.127	1.02	[-.45, .45]	[-.53, .38]		
4	1.269	0.97	[-.59, .31]	[-.67, .23]	[-.59, .31]	
5	1.344	1.19	[-.67, .23]	[-.74, .16]	[-.67, .23]	[-.52, .37]

Note. Comparison performed using Tukey's method at 0.05 significance level.

7.6.5 Item- and Person-Fit Statistics of the Revised PSRT

Item-level Fit Statistics and Tests: Unlike the original test, the fit statistics in the revised version revealed only two types of issues in items. The first anomaly concerned the item-model fit in two items that had fit residuals above +2.5, as reflected in the detailed fit statistics in Table 7.22. These items were part of the group identified with similar misfit from the initial analysis with the original test. The remaining 84 items in the revised test had acceptable fit statistics, as displayed in Appendix 21.

Table 7.22

Fit statistics of two items in Revised PSRT with anomalous fit residuals

Item	Location Estimate	SE	Fit Residual	*df*	χ^2	*df*	*p*-value
BVIQ2(a)	0.984	0.168	3.864	166.91	14.132	4	0.00689
BVIIIQ1(a)(i)	1.204	0.167	3.460	166.91	18.804	4	0.00086

Note. SE refers to standard error, df is the degrees of freedom, p-value is the probability, and χ^2 is chi-square value.

The second type of anomaly was identified in the high residual correlations between five item pairs found within questions. An examination of these items suggested that the source of local dependence was trait-related due to shared stimulus or assessment objectives. Recommendations for changes, such as testing alternative concepts, were made for the affected items. Table 7.23 identifies these items and the three affected question blocks. The residual correlations between the items ranged from .343 to .661.

Table 7.23

Five item pairs with anomalous residual correlations

Question Block	Item 1	Item 2	Residual Correlation
Block V	BVQ2(a)	BVQ2(b)	.427
Block VIII	BVIIIQ1(a)(i)	BVIIIQ1(a)(ii)	.532
Block VIII	BVIIIQ3(e)(i)	BVIIIQ3(e)(ii)	.562
Block X	BXQ1(b)(i)	BXQ1(b)(ii)	.661
Block X	BXQ2(a)(i)	BXQ2(ii)	.343

In sum, resolution efforts resulted in a reduction in both the types and instances of statistical misfit to the model in the revised test. The continued presence of misfit in some items suggested there were issues at the item-design level. After a further review that considered substantive theories and pupils' responses, specific recommendations for modifications were given for affected items. See Appendix 22 for the item-specific recommendations. The items had not been deleted because they assessed unique or valuable competencies and learning outcomes. Notably, the removal of the three DIF-affected items from the revised test meant there was now no further evidence of violations of invariance of item parameter estimates across key subgroups. Neither were there any threshold disorders in the eight remaining polytomous items.

Person-level Fit Statistics and Tests: Of the 431 pupils in the sample, only one had a person-fit statistic with a residual exceeding the criterion level. However, as the value (2.551) of the residual was just slightly above the criterion and removal of the person did not improve

the chi-square interaction statistics and PSI, a decision was made to retain the pupil as part of the sample.

7.6.6 Revised PSRT Booklets

This subsection reports the functioning of the five Revised PSRT booklets. The reliability indices and fit statistics of each booklet are shown in Table 7.24.

Table 7.24

Reliability indices and fit statistics of booklets 1 to 5

	Booklet 1	Booklet 2	Booklet 3	Booklet 4	Booklet 5
No. of items	34	39	43	40	42
PSI	0.844	0.836	0.857	0.834	0.874
Cronbach's α	0.872	0.849	0.862	0.845	0.885
Total-item chi-square/*DF* ratio (probability)	1.037 (*p*=.366)	1.043 (*p*=.341)	1.113 (*p*=.148)	1.324 (*p*=.004)	1.081 (*p*=.224)
Mean item-fit residual (*SD*)	-0.178 (0.757)	-0.092 (1.092)	-0.093 (0.802)	-0.013 (1.366)	-0.149 (0.730)
Mean person location estimate (*SD*)	1.633 (1.199)	1.032 (0.990)	1.019 (1.028)	1.250 (0.953)	1.316 (1.185)
Mean person-fit residual (*SD*)	-0.161 (0.634)	-0.170 (0.740)	-0.165 (0.650)	-0.191 (0.870)	-0.150 (0.712)

Note. DF and SD refer to degrees of freedom and standard deviation respectively.

In each booklet, the similarity between the PSI and the Cronbach's α coefficient values suggested an adequate degree of alignment between the person abilities and the item difficulties. Other than booklet 4, the remaining booklets had non-significant chi-square interaction statistics, providing evidence of good overall model fit.

PCA of Residuals and *t*-Tests: For each of the booklets, the PCA of person-item residuals showed the eigenvalue of the first principal component to be noticeably larger than the eigenvalues of the remaining principal components, and the difference between the first

and second principal components was more marked than the difference for the second and third principal components. See Table 7.25 for the eigenvalues of the first three principal components of every booklet.

Table 7.25

Eigenvalues of the first three principal components from the PCA analysis of person-item residuals for the five booklets

Principal Component	Eigenvalue				
	Booklet 1	Booklet 2	Booklet 3	Booklet 4	Booklet 5
1	2.527	2.716	2.936	3.149	3.007
2	2.268	2.432	2.457	2.764	2.402
3	2.020	2.341	2.387	2.649	2.377

On the basis that the results suggested the presence of a dominant first residual principal component, the selection of items to represent the two subsets with the most contrasting properties took place using the item loadings on only this first principal component (Hambleton et al., 1991). The results of the within-person *t*-tests and binomial tests using the person estimates from the two identified subsets of each booklet are presented in Table 7.26.

Table 7.26

Results of t-tests and binomial tests of person estimates from subsets of items identified from first principal component across all booklets

	Booklet 1	Booklet 2	Booklet 3	Booklet 4	Booklet 5
Total no. of *t*-test analyses	86	86	86	86	87
% of significant *t*-test	8.14%	8.14%	11.63%	10.47%	8.05%
Proportion exceeding 95% lower C.I.*	0.035	0.035	0.070	0.059	0.035

*Note. *Expected proportion set at 0.05.*

As shown in Table 7.26, no booklets registered the percentage of significant *t*-tests between individual estimates for the subsets of items identified by the PCA, under the acceptable threshold of 5%. Values for booklets 1, 2, and 5 were below 10%, while booklets 3 and 4 obtained values slightly over 10%. From the binomial tests, the observed proportions of significant *t*-tests in three of the booklets fell within the acceptable range. The proportions for the remaining two booklets marginally exceeded the expectation. The findings of this procedure suggested that responses to the booklets demonstrated limited and varying levels of multidimensionality, that were most evident in booklets 3 and 4.

7.7 Discussion

This chapter has presented evidence from activities conducted over two stages in the main study (phase IV) to evaluate whether the 100-item PSRT measured a unidimensional construct. The first stage produced content validity evidence that items in the assembled PSRT measured the scientific reasoning construct correctly and adequately. Part of this evidence came from the findings of the earlier two phases that gathered expert reviews and preliminary psychometric information about item functioning in the Draft PSRT. As explained in the last chapter (see subsection 6.8.1), 64 of the items in the PSRT originated from the draft test. Of the remaining 34 items that were newly designed, the findings of a two-member expert panel highlighted no concerns about the intended science practices and knowledge types measured but instead found editorial and item construction issues with one-third of them. Therefore, while there was evidence that the new items satisfied two of the features of content validity – domain representation and relevance – problematic items had to be rectified to produce stronger evidence about the appropriateness of the test-development process.

The first validation stage culminated in the drawing up of two TOS to produce evidence that the items in the PSRT sampled important learning outcomes adequately. These tables outlined the allocation of the assessment objectives according to the curriculum and the scientific reasoning construct. Additionally, there was evidence that issues in PSRT items

had been tracked and appropriately resolved from the record of item development in the table of revision history. In all, the collective evidence from activities in the first validation stage suggested that the 100-item PSRT had adequate content validity.

Validity evidence from the sources of response processes and internal structure came from the second validation stage. In this stage, procedures enacted during the administration and scoring of the PSRT ensured there would be minimal validity threats to score interpretation. On the assumption that the underlying construct responsible for the response patterns was unidimensional, the subsequent analyses adopted the Rasch model to assess the fit of every item in the PSRT. As explained in Chapter 6, claims of unidimensionality are dependent on the extent to which the data fitted the Rasch model. Therefore, violations of the model and its assumptions, as detected from a suite of fit statistics and graphical information, reduce the strength of such claims.

The initial Rasch analysis identified 40 items with DIF and model-fit issues. After implementing an iterative diagnostic procedure that attempted to remediate item fit issues, the fit statistics of the resulting 86-item Revised PSRT had markedly reduced levels of item and person misfit compared to the original test. For instance, the revised test had only seven items which indicated model-fit issues and violations of local independence. These items were recommended for further modifications after considering substantive theories and pupils' responses. Of the remaining 79 items that had acceptable fit statistics, 33 received further recommendations for improvement based on pupils' responses to the items. Also, none of the eight polytomous items in the revised test evidenced disordered thresholds. Still, an inspection of the person-item threshold distributions indicated some mistargeting that was localised to both ends of the scale.

Analyses showed that both genders had comparable overall performance on the Revised PSRT and that each of the five booklets had approximately equal difficulty. These results suggested that the items functioned similarly between genders at the test level, and that regardless of the assigned booklets, pupils generally encountered equivalently difficult

items. Four of the booklets demonstrated good overall model fit, as evidenced by the overall non-significant chi-square statistics. Finally, the results from the combined procedure of a PCA of residuals and *t*-tests on person estimates provided evidence that responses across the booklets demonstrated varying and limited levels of multidimensionality, but most individual response patterns were consistent with the model assumption of unidimensionality. Overall, these post-hoc results supported broad inferences that most of the items were appropriate for the sample, and that each of the booklets predominantly measured a single construct.

In summary, acceptable reliability and fit indices for the majority of items in the Revised PSRT, and limited evidence of violations to invariance, suggested that there was good data to model fit. Such evidence that pupils' performance on most of the test items was consistent with the requirements of the Rasch measurement model provided significant statistical support to conclude that a unidimensional construct accounted for responses to the PSRT. The observed misalignment between the distribution of pupils and item estimates across the range of abilities, which indicated that the PSRT was on average a little easy for the sample of pupils, will be discussed in the concluding chapter of the thesis – Chapter 9. This final chapter will also discuss the presence of multidimensionality in the booklets.

7.8 Conclusions

The results in this chapter addressed the first of two research aims in the study. This aim included the evaluation of whether the PSRT measured a unidimensional construct. The collection of evidence occurred over two stages. The first stage involved the systematic selection, design and review of items intended for the test. The next stage administered the test to the pupils, while ensuring fair testing conditions across the schools. Analysis of response patterns using an array of fit statistics provided evidence that most of the items in the 86-item Revised PSRT displayed good fit to the unidimensional Rasch model, substantiating the claim that the PSRT measures a unidimensional construct. The next

chapter reports the results of analyses on the nature of this construct by investigating whether pupils' performance on the items accorded with the model of scientific reasoning.

Chapter 8 Main Study – Identification of Construct

8.1 Introduction

As introduced in Chapter 6, this study adopted Rasch modelling as the main analytical approach. The approach was applied to the 100-item PSRT through an iterative calibration and remediation process to produce an 86-item Revised PSRT. During the process of identifying model-fit issues, evidence was produced to address the first aim of the research; that is, to examine the extent to which the revised test measured a singular construct. As reported in the previous chapter, there was adequate evidence of a unidimensional construct underlying the pupils' performance on the test.

This chapter reports the results of analysis using the location estimates of one-mark (dichotomous) items and the threshold estimates of two-mark (polytomous) items to address the second research aim. This aim evaluates whether the Revised PSRT measured the hypothesised construct of scientific reasoning, and specifically, whether the item location and threshold estimates conform to the expectations predicted by the framework of scientific reasoning presented in Chapters 2 and 4.

8.2 Item Difficulty and Item Features

The analytical methods reported in this chapter examine the five key features that characterised an item to understand how they impact on item difficulty (Chapter 4, subsection 4.5.1). See Figure 8.1 on the next page for a schematic representation of these item features. Other than the item's mark weighting, selection of the other four features was theoretically driven, as described in Chapter 4 (subsection 4.4.3). Two of these features were knowledge types and science practices (and their sub-practices) in the scientific reasoning framework. The remaining features were cognitive demand and the item format.

Each of the five features consisted of three categories. For instance, there were three types of knowledge and science practices, and there was a maximum of three marks assigned to

items. As explained in the test framework (Chapter 4, subsection 4.5.1), the categories in each of the five key features helped characterise an item. Theoretically, they also functioned *together* to determine the expected difficulty. As the analytical procedures reported in this chapter examined items separately according to their mark weighting, the discussion focuses on the predicted relationships between item difficulty and the remaining four features shown in Figure 8.1.

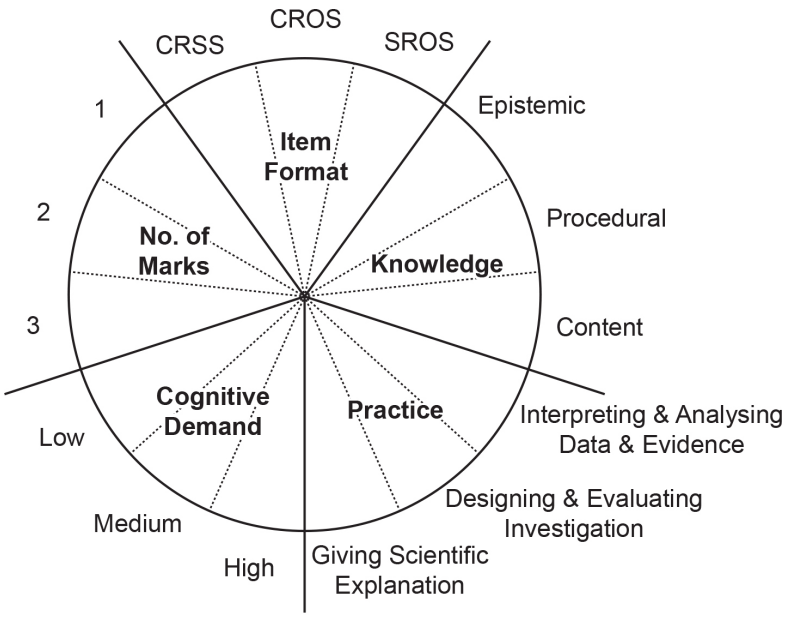


Figure 8.1. Schematic representation of the five key item features hypothesised to impact difficulty. SROS is selected-response objective scoring, CROS is constructed-response objective scoring, and CRSS is constructed-response subjective scoring.

Content, Procedural and Epistemic Knowledge: As described in subsection 4.3.1 of Chapter 4, the Singapore primary science curriculum emphasises the teaching of content knowledge in the early grades. Building on this foundation from the initial years, procedural knowledge is taught to older primary pupils. Compared to the two knowledge types, the pupils in the sample had least curriculum exposure to epistemic notions. Marks were thus deliberately allocated among the three knowledge types in the PSRT to reflect the curriculum focus and ensure pupils could respond adequately. For instance, about four-fifths of the marks were assigned to items assessing content and procedural knowledge. Pupils were therefore predicted to perform better on these items, i.e., items measuring content and epistemic knowledge should have been easiest and most difficult respectively.

Although age-appropriate epistemic knowledge was chosen for testing, these items varied in complexity. See details about this knowledge type in Chapter 4, subsection 4.3.1.

Three Science Practices: Among the competencies in the three science practices, pupils had greatest familiarity with those from the practices of *giving scientific explanation.* This familiarity corresponds to the curriculum's emphasis on the acquisition of basic science competencies for beginners. These competencies encompass scientific modelling and explanatory accounts of scientific phenomena. In contrast, pupils could only demonstrate simple and elementary forms of investigative work in the competencies of the science practice – *designing and evaluating investigation* – as they have yet to acquire sufficient scientific knowledge to show greater proficiency. In the PSRT, only about a quarter of the marks were allocated to items measuring the latter competencies. See details in the TOS (Chapter 4, subsection 4.5.3). Therefore, the model of scientific reasoning posits that the easiest items were those that tested the reasoning underlying pupils' modelling and explanatory competencies, while the most difficult items tested their investigative competencies.

Item Format: Generally, item formats that require the construction of reasoned arguments, explanations, or answers, have been shown to be more difficult (Haladyna & Rodriguez, 2013). In particular, responses that involve extended writing should pose higher difficulty than shorter answers. This study therefore expected items of the CRSS and SROS formats to be the most difficult and easiest respectively.

Cognitive Demand: The level of cognitive demand in an item, which represents the combined cognitive load and mental processing required of pupils, is expected to correspond directly to the difficulty of the item (Haladyna, 2004). Hence, an item requiring analysis or evaluation of complex phenomena or data should be more difficult than one that involves one or two simple steps such as direct reading of data. See Chapter 4, subsection 4.5.3, for details of the cognitive demand categorisation scheme in this study. The next

section discusses how the study's data-analytical procedures utilised these features to determine if the location or threshold estimates accorded with *a priori* expectations.

8.3 Data Analysis

The analysis began at test-level by comparing the average difficulty of items in the same category with those of other categories within a feature to evaluate whether the difficulty order aligned to theoretical expectations. For instance, the mean location estimates of items that had high cognitive demand were contrasted with those of medium- and low-demand. Thereafter, other analytical procedures examined dichotomous and polytomous items separately, as the procedures and analysed item indices were different.

Three procedures were conducted on the location estimates of the dichotomous items. The first examined the item map, which juxtaposed the approximate location estimates of pupils and items as two vertical histograms on the same scale. The left histogram distributes the test-takers from the strongest-ability at the top to the least able at the bottom. Likewise, the right histogram displays the item order in descending order of difficulty, from the most difficult to the least difficult. When a pupil and an item are found on the same point along the scale on the map, there is comparable item difficulty and pupil ability (Wright & Stone, 1979). Therefore, items that invoke correct responses from only the most able pupils should be found on the more difficult end of the scale, and vice versa. Using the item order on the map, the 78 one-mark dichotomous items were first compared to discern the influence of item features on their difficulty.

To enable stronger inferences about the bearing of item features on item difficulty, the next procedure examined the estimates of those dichotomous items in the revised test that had neither been recommended for change nor undergone remediation. Item analysis using the key features initially took place on these items *within* questions to reduce the possible effects of extraneous factors, such as item context and reading load (Haladyna & Downing, 2004). Of interest were the location estimates of items with features that were identical or differed by only one.

Items with identical features should have similar location estimates. Otherwise, the single different feature among items should have significant impact on the relative order in location estimates. This procedure was then repeated with items *across* questions. In the third and final procedure, there was a comparative examination of features belonging to the 10 most difficult and 10 easiest dichotomous items with no change recommendations or model-fit irregularities (such as those described in Chapter 7). This approach, of comparing items with the most disparate location estimates, aimed to further support inferences about the roles of item features in dichotomous items.

As mentioned in the previous chapter, there were eight two-mark (polytomous) items in the Revised PSRT. Six of these were polytomous items that had undergone re-marking to correct issues such as disordered thresholds or had been retained from the PSRT with change recommendations. The remaining two were testlets formed from the merger of dichotomous items; these were not examined further since there were no *a priori* expectations about how they would function as polytomous items. Therefore, the analysis only examined the first six polytomous items and focused on the information about the hierarchical levels of knowledge or competencies pupils demonstrated to evaluate whether the behaviours accorded with theoretical expectations.

The final analysis combined a comparison of key features in items that measured competencies from the same practices with a qualitative examination of pupils' raw responses. When such comparisons were supported by information about the actual location or threshold estimates, inferences could be made about whether the elicited performance and reasoning accorded with those intended by the model of scientific reasoning presented in this thesis. The following section reports the outcomes of the procedures conducted, first at the test level, and then on the two item types in turn, before reporting the results of the examination of pupils' responses.

8.4 Results

8.4.1 Item Difficulty Estimates in the Revised PSRT

Overall Item Location Estimates: As shown in Table 8.1, the difficulty estimates of the 86 items ranged from -3.865 to 2.362. Forty-nine items, representing 57% of the revised test, had location estimates above the mean location of 0 logits.

Table 8.1

Distribution of 86 items in location estimate intervals, in descending order of difficulty

Location Estimate Interval (logits)	Highest and Lowest Location Estimates	Number of Items (%)	
		Dichotomous	Polytomous
≥ 1.0	1.021 to 2.362	18 (20.9%)	5 (5.8%)
0.0 to < 1.0	0.011 to 0.984	23 (26.7%)	3 (3.5%)
-1.0 to < 0.0	-0.096 to -0.096	16 (18.6%)	0
-2.0 to < -1.0	-1.998 to -1.022	12 (14.0%)	0
< -2.0	-3.865 to -2.612	9 (10.5%)	0
	Total	78 (90.7%)	8 (9.3%)

Mean Item Difficulty Estimates by Categories of Item Features: The mean item location estimates of the three categories in each of the item features are shown in Table 8.2. A discussion of these results follows the table.

Table 8.2

Mean location estimates and 95% confidence intervals of 86 items across four item features

Item Feature	Category	No. of Items	Mean Item Location Estimates	[95% C.I.]
Cognitive Demand	Low	42	-0.904	[-1.268, -0.541]
	Medium	25	0.697	[0.248, 1.146]
	High	19	1.287	[0.828, 1.746]
Item Format	SROS	13	-0.260	[-0.976, 0.455]
	CROS	17	-1.675	[-2.186, -1.165]
	CRSS	56	0.639	[0.329, 0.968]
Knowledge Type	Content	29	-0.183	[-0.706, 0.341]
	Procedural	40	-0.029	[-0.485, 0.426]
	Epistemic	17	0.732	[-0.122, 1.342]
Science Practice	1	38	-0.136	[-0.551, 0.279]
	2	23	0.493	[-0.091, 1.077]
	3	25	-0.091	[-0.775, 0.593]

Note. C.I. = confidence intervals. SROS is selected-response objective scoring. CROS is constructed-response objective scoring, and CRSS is constructed-response subjective scoring. Practice 1 – giving scientific explanation, practice 2 – designing and evaluating investigation, practice 3 – interpreting and analysing data and evidence.

***(a) Cognitive demand*:**

The observed order of mean difficulty among the items with the three levels of cognitive demand supported theoretical expectations. Also, the lack of overlap in the C.I. for the mean location estimate of low-demand items versus the medium- and high-demand items indicated the mean value for the low-demand items was statistically lower than the corresponding values for medium- and high-demand items. There was substantial overlap in the C.I.'s of the mean location estimates of the medium- and high-demand items indicating a lack of statistical confidence in the difference between these means.

(b) Item format:

As expected, CRSS items recorded the highest mean location estimate among the three formats. This value significantly differed from the mean estimate of CROS items, but not from the mean estimate of the SROS items. The order of the mean difficulty estimate of

SROS items was not as anticipated. Instead of turning out to be the format associated with the easiest items, the mean location estimate of SROS items was significantly greater than that of CROS items.

(c) Knowledge type:

Overall, the observed difficulty order was as postulated with items measuring content and epistemic knowledge registering the lowest and highest average difficulty respectively. However, the large overlap between the C.I.'s of the mean location estimates for each knowledge type indicated the three values were not statistically different.

(d) Science practice:

The order of the mean difficulty among items testing competencies from the three science practices conformed to theoretical expectations. Items testing practice 2 competencies – *designing and evaluating investigation* – attained higher mean difficulties than items measuring competencies from the other two practices. The overlap in C.I.'s of the mean estimates of items measuring the three practices indicated the differences among the three estimates were not statistically significant.

8.4.2 Item Difficulty Estimates of Dichotomous Items

Features of 78 Dichotomous Items Across the Scale: For a concise snapshot of the effect of item features on the location estimates of these items, the dominant features of the three easiest and most difficult dichotomous items, as shown in the item map in Figure 8.2, are described first. Thereafter, the 78 dichotomous items along the scale will be systematically analysed according to four item features – knowledge type, science practice, cognitive demand, and item format.

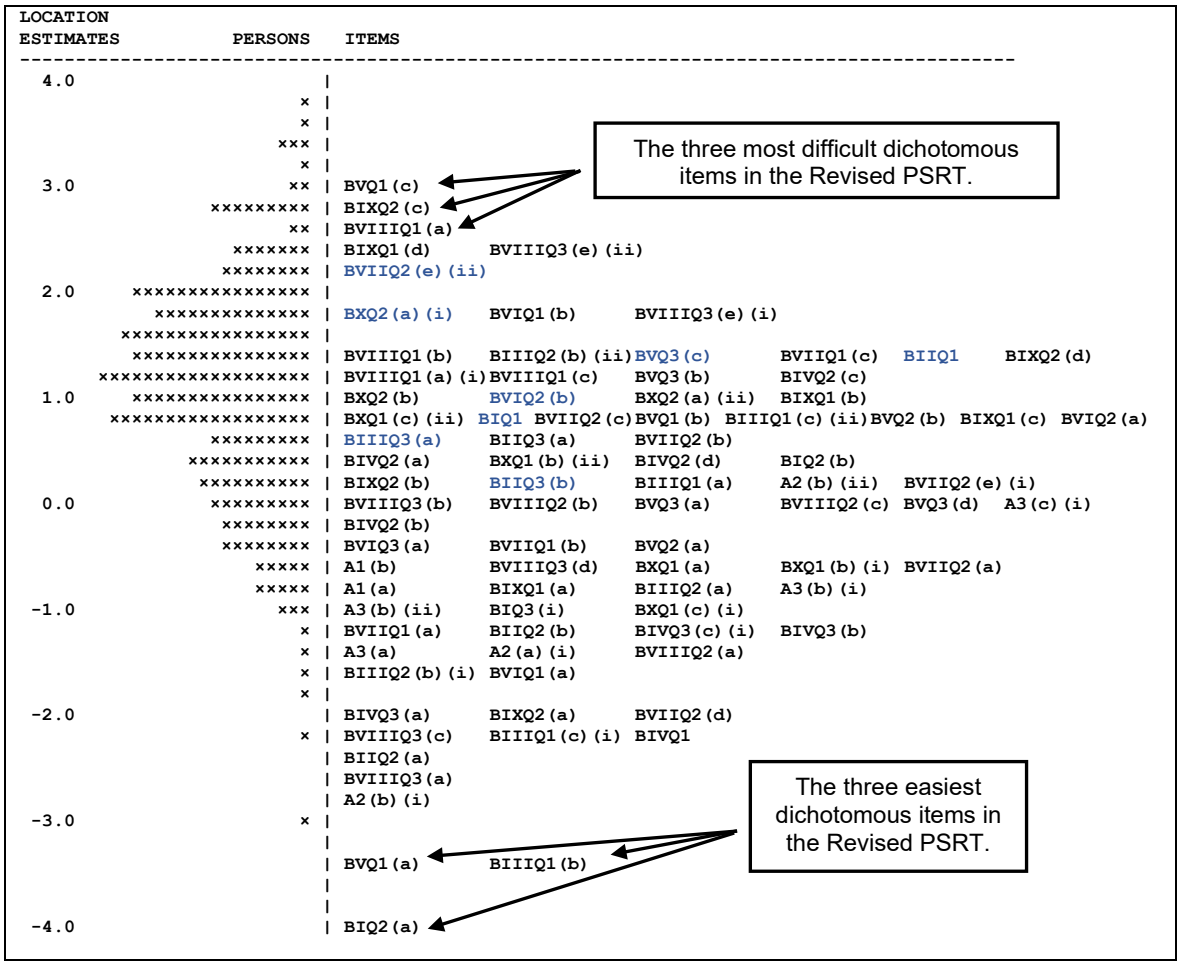


Figure 8.2. Item map from RUMM2030 of the 86-item Revised PSRT. Every "X" represents two pupils. Item label indicates the block and order of items; e.g., A1(b) is the second item of the first question in the anchor block, BIIQ3(b) is the second item of question 3 from block II. Labels in blue indicate the eight polytomous items. The seven DIF-resolved items were left out of this item map.

The three most difficult dichotomous items had a CRSS item format. Two of them – BVQ1(c) and BVIIIQ3(e)(ii) – measured epistemic knowledge using experimental competencies related to practice 2 (*designing and evaluating investigations*). They were of high- and medium-demand respectively. The third item – BIXQ2(c) – tested procedural knowledge during data processing (practice 3: *interpreting and analysing data and evidence*). It had a medium cognitive demand. These observations about the item features of the three most difficult items generally conformed to the expectations outlined in section 8.2. The items were of either high- or medium-demand and involved the production of responses. Also, two of the three items measured a combination of epistemic knowledge and investigative competencies on which the primary pupils had the least curriculum instruction.

In contrast, the three easiest items had a CROS format, and were designated of low cognitive demand. Two of them – BIQ2(a) and BVQ1(a) – assessed content knowledge in modelling and data-analytical competencies of practices 1 (*giving scientific explanations*) and 3 respectively. On the other hand, the third item – BIIIQ1(b) – measured procedural knowledge and competencies from practice 3. In a similar way to the most difficult items, these features of the easiest items were largely consistent with predictions. All required objective scoring of short answers and were of low demand. Also, two out of the three items tested knowledge and competencies of the two practices that received the greatest attention in the primary curriculum.

Scrutiny of the remaining items in the map revealed that those located towards the higher end of the scale tended to be of high- or medium-demand and possessed the CRSS item format. On the other end of the scale were more low-demand items with significantly reduced numbers of CRSS items. Unlike the findings from the examination of the three most difficult and easiest items, the findings were less conclusive about the relationship between item difficulty and knowledge types or science practices. See Appendix 23 for the features of the 86 dichotomous and polytomous items arranged in order of their difficulty. In the next analysis, the relative distribution of the three categories of each item feature within an interval on the scale was systematically investigated among the dichotomous items for further evidence that these conformed to theoretical expectations.

Distribution of Cognitive Demand Levels in Location Estimate Intervals: The observed distribution of items with the three demand levels within each interval, and across the intervals on the scale, was consistent with theoretical expectations. As shown in Figure 8.3, the proportions of low-demand and high-demand items among all the items in the interval with the highest location estimates (≥1.0) were the smallest and largest respectively. Conversely, approximately 80–90% of all the items in each of the three intervals on the lower end of the scale, were low-demand items. There were no high-demand items in the two intervals with the lowest locations.

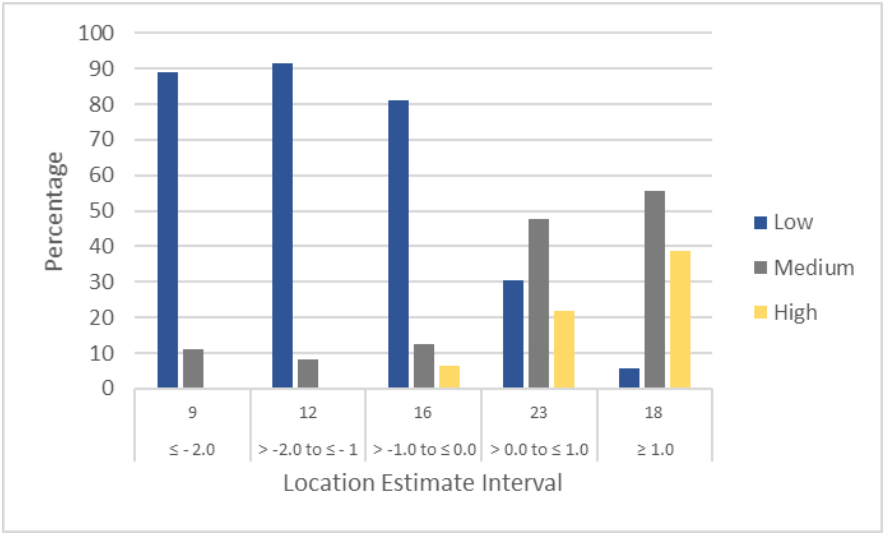


Figure 8.3. Items of each demand level, as a percentage of the total items (indicated by the number below the bars) in each location estimate interval.

Distribution of Item Format Types in Intervals Across the Scale: As seen in Figure 8.4, successive intervals towards the higher end of the scale contained more CRSS items, as a proportion of all the items, in each interval. CROS items only appeared in the intervals towards the lower end of the scale. These observed trends were consistent with the general prediction about constructed-response formats; that extended responses were likely more difficult over shorter answers. SROS items were present in every interval across the scale.

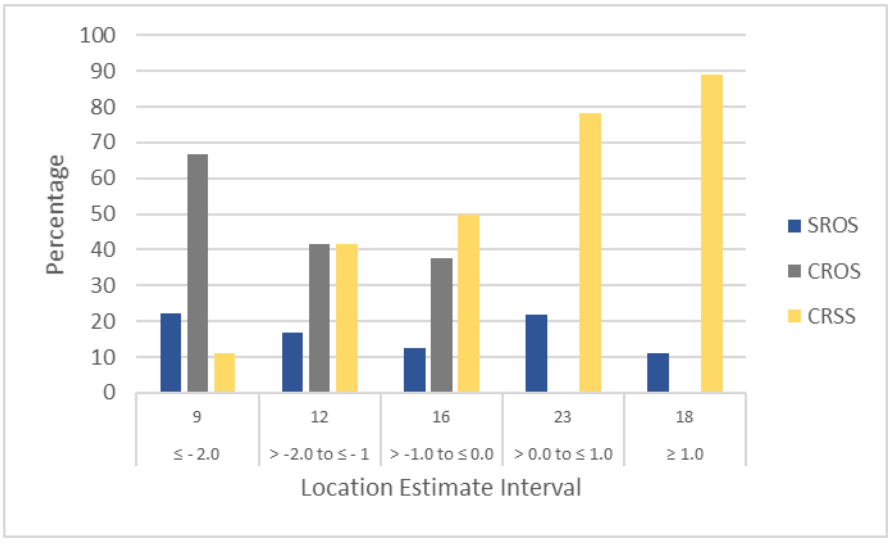


Figure 8.4. Items of each format, as a percentage of the total items (indicated by the number below the bars) in each location estimate interval. SROS is selected-response objective scoring, CROS is constructed-response objective scoring, CRSS is constructed response subjective scoring.

Distribution of Knowledge Types in Intervals Across the Scale: As evidenced in Figure 8.5, the items measuring epistemic knowledge, as a proportion of the total items in every interval, were the smallest towards the easiest end of the scale. This trend is consistent with theoretical expectations.

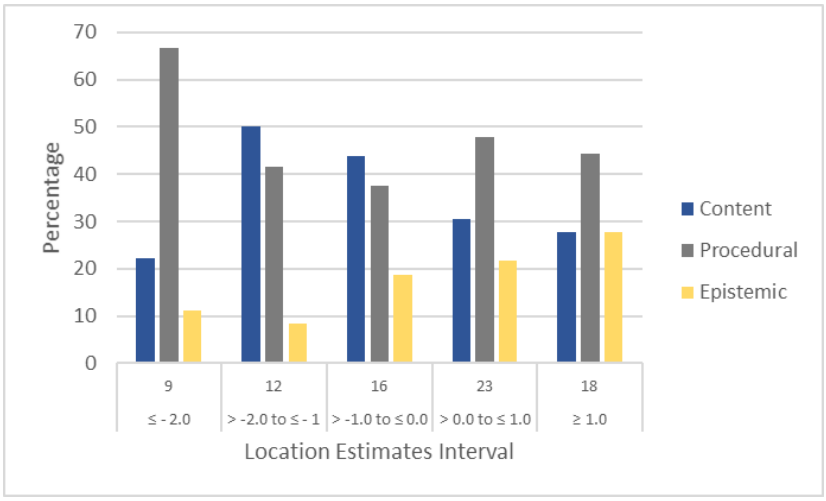


Figure 8.5. Items measuring each knowledge type, as a percentage of the total items (indicated by the number below the bars) in each location estimate interval.

Distribution of Science Practices in Intervals Across the Scale: Across the last three intervals towards the more difficult end of the scale, there was an increasing proportion of items, out of all items in each interval, assessing practice 2 (*designing and evaluating investigation*) competencies. This observation conforms to the predicted distribution.

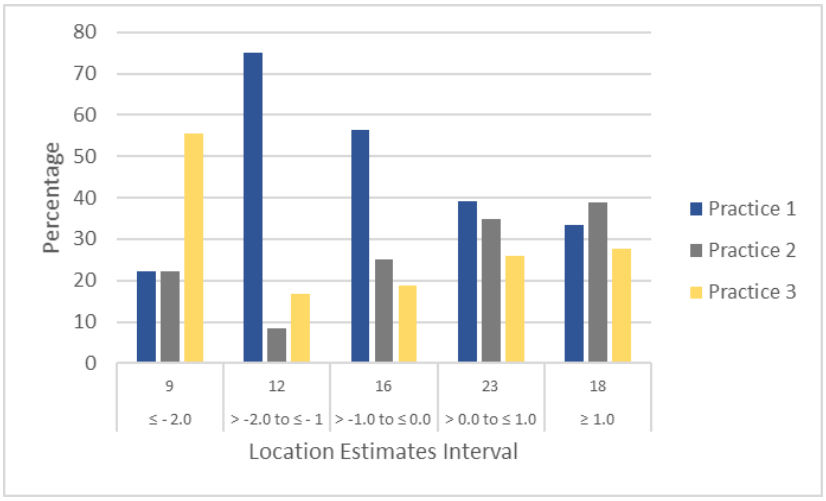


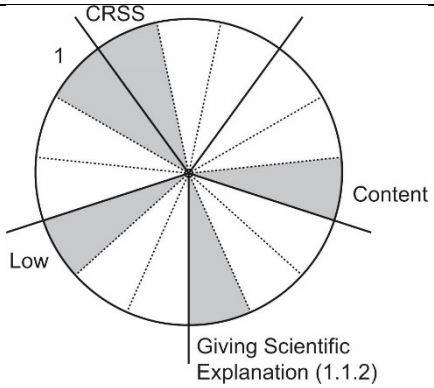
Figure 8.6. Items measuring each science practice, as a percentage of the total items (indicated by the number below the bars) within each location estimate interval. Practice 1 – giving scientific explanation, practice 2 – designing and evaluating investigation, practice 3 – interpreting and analysing data and evidence.

Subsequently, to gather more evidence and strengthen inferences about the alignment between the order of difficulty and the item features, the features of only those items that were not statistically remediated and underwent no recommended changes were examined. As mentioned in the data analysis section, this examination compared the location estimates among items that: (1) differed by only a single feature; and (2) shared the same features within and across questions. The results are discussed below.

Features of Dichotomous Items from the Same Question: The investigation involved an initial comparison of features among items found in each question. The examined items came from the group of 47 single-mark items that passed both qualitative and statistical checks (see Appendix 24). Of the six pairs of items that fulfilled these criteria, only one pair had identically assigned features (see Table 8.3).

Table 8.3

Details of the only pair of items within a question in block VIII that had identical features

Item Pair	Item	Characterisation Scheme	Location Estimates (SE) [95% C.I.]	Similarity in Location Estimates?
1	BVIIQ1(a)		-1.11 (0.244) [-1.588, -0.631]	×
	BVIIQ1(b)		-0.288 (0.198) [-0.676, 0.100]	

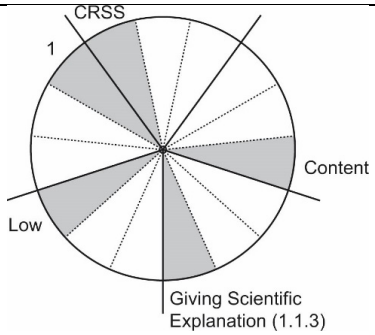
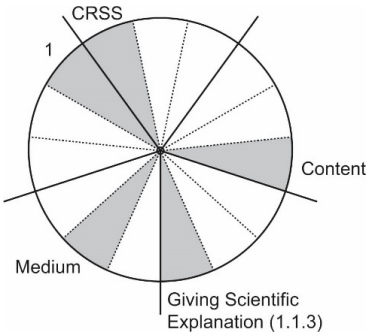
Note. × *means no. SE = standard error, C.I. = confidence interval.*

As shown in Table 8.3, this item-pair assessed the ability to recall and apply content knowledge. Both items were also of low demand and had a CRSS format, so their location values were expected to be similar. However, the observed location estimates were not close. An examination of the raw responses suggested that this difference arose because one item underwent adjustments to its scoring scheme. The easier item resulted from a

widened outcome space to accommodate the pupils' inability to express science processes using formal "concept" words (Wellington, 2000). Specifically, non-specialist words like "squish" and "smash" were alternatives accepted in this item to signify the breakdown or pulverisation of food. Given the very small overlap in the C.I. for the two location estimates and the qualitative evidence from the responses, there was greater support for the view that the items substantially differed in their difficulty. Of the remaining five pairs with a single difference in item features, only one pair differed by cognitive demand (see Table 8.4).

Table 8.4

Details of the item pair within a question in block III that differed by cognitive demand

Item Pair	Item	Characterisation Scheme	Difference	Location Estimates (SE) [95% C.I.]	Expected Order of Location Estimates?
2	BIIIQ2(b)(i)^		Low	-1.523 (0.278) [-2.068, -0.978]	✓
	BIIIQ2(b)(ii)^		Medium	1.481 (0.173) [1.136, 1.826]	

^Content knowledge: Evidence that one magnet is stronger than another from set-up

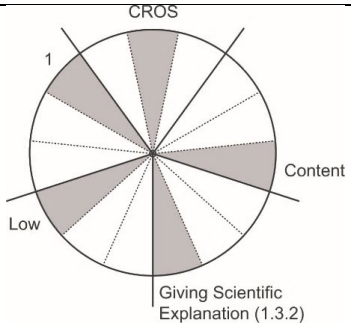
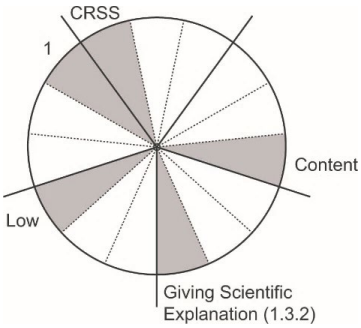
Note. ✓ means yes. SE = standard error, C.I. = confidence interval.

The location estimates between the items in this pair were statistically different. While the order of location values obtained in the items was as intended, the difference spanned almost 3 logits, which was half the scale between the easiest and the most difficult items.

This large difference was unexpected, as the items only differed by one level of cognitive demand. The item pair required pupils to apply their knowledge and experiences with magnets and magnetic materials to respond. However, an examination of the scripts revealed evidence that the larger-than-intended difficulty arose because pupils were unaccustomed to the required application depth of the scientific concepts. This point will be elaborated later in subsection 8.4.4. The items in each of the last four pairs differed by their item format (see Tables 8.5A, 8.5B, and 8.5C). While the location estimates in each pair were consistent with the expected order, only the values of the last pair (Table 8.5C) were significantly different. Given the large overlap in the C.I. of the estimates of the first three item-pairs, there was no evidence that the items' different item formats led to the predicted differences in difficulty.

Table 8.5A

Details of the first of the four item pairs found in questions that differed by their item format

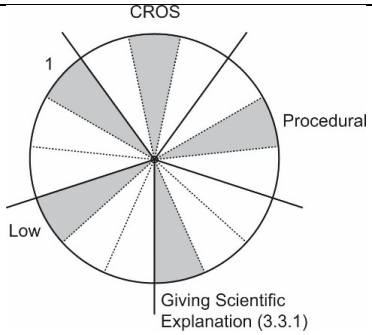
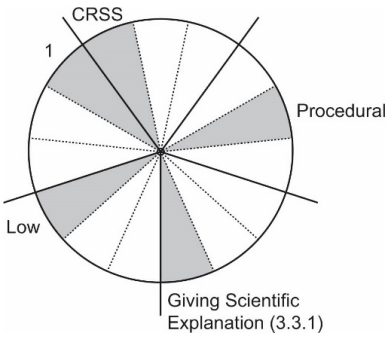
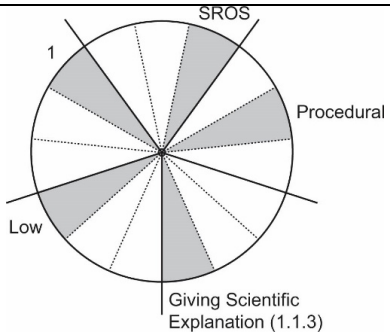
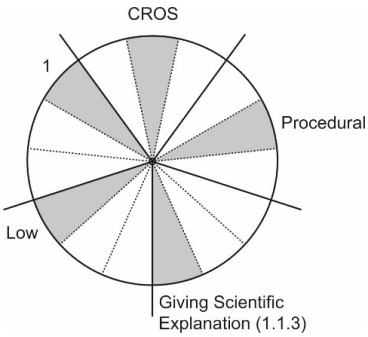
Item Pair	Item	Characterisation Scheme	Difference	Location Estimates (SE) [95% C.I.]	Expected Order of Location Estimates?
3	A1(a)*		CROS	-0.744 (0.140) [-1.018, -0.470]	×
	A1(b)*		CRSS	-0.543 (0.133) [-0.804, -0.483]	

*Content knowledge: Magnets attract magnetic materials

Note. × means no. SE = standard error, C.I. = confidence interval.

Table 8.5B

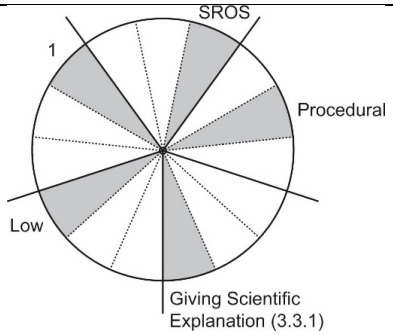
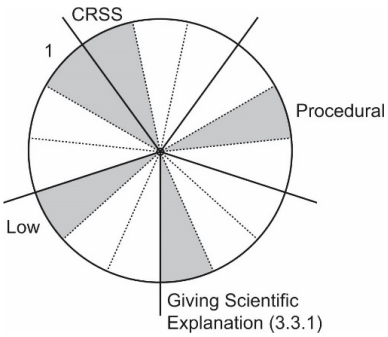
Details of the second and third item pairs found in questions that differed by their item format

Item Pair	Item	Characterisation Scheme	Difference	Location Estimates (SE) [95% C.I.]	Expected Order of Location Estimates?
4	BIIIQ1(b)*		CROS	-3.212 (0.528) [-4.247, -2.177]	×
	BIIIQ1(c)(i)*		CRSS	-2.052 (0.334) [-2.707, -1.397]	
	* Procedural knowledge: Identification of key variables in experimental set-ups				
5	BIVQ3(b)**		SROS	-1.022 (0.230) [-1.473, -0.571]	×
	BIVQ3(c)(i)**		CROS	-1.050 (0.231) [-1.503, - 0.597]	
	** Procedural knowledge: Basic features of circuit diagrams				

Note. × means no. SE = standard error, C.I. = confidence interval.

Table 8.5C

Details of the last of four item pairs found in questions that differed by their item format

Item Pair	Item	Characterisation Scheme	Difference	Location Estimates (SE) [95% C.I.]	Expected Order of Location Estimates?
6	BXQ1(b)(i)*		SROS	-0.408 (0.210) [-0.820, 0.004]	✓
	BXQ1(b)(ii)*		CRSS	0.447 (0.179) [0.096, 0.798]	

*Procedural knowledge: Identification of key variables in experimental set-ups

Note. ✓ means yes. SE = standard error, C.I. = confidence interval.

As mentioned, the items in the last pair, shown in Table 8.5C, evidenced significantly different location estimates. The items had formats that differed in both response format (selected versus constructed) *and* scoring requirement (objective versus subjective). In contrast, the format of items in each of the other three item-pairs differed in either response format *or* scoring. Still, very limited inferences could be drawn about the impact of the specific characteristics of item format on item difficulty using evidence solely from this single pair in the test.

Overall, the analysis presented in this subsection provided some support for earlier findings about the postulated effect of item format and cognitive demand on item difficulty. However, it was not possible to address the impact of knowledge types and science practices. The next analysis investigates how item features impacted on difficulty in items across questions.

Features of Dichotomous Items across Questions: Only one item pair in the test differed by just a single feature: cognitive demand. Analysis indicated that the order of the location estimates was as intended. Both items assessed the COV strategy, which is an important form of scientific reasoning. The performance of pupils on these two items will be discussed in detail in subsection 8.4.4. The analyses across questions did not uncover any items that shared identical item features.

For pragmatic reasons then, given the limited number of items for comparison, the wide disparity in features among the items, and the difficulty in attributing variation in location to a single item feature, the analysis proceeded to identify two groups of 10 dichotomous items each with the most disparate locations. These two groups represented the “hardest” and “easiest” items in the revised test. A point worth reiterating is that the selected items were free from change recommendations and statistical irregularities to enable the drawing of stronger inferences about the impact of item features on location estimates. See Table 8.6 for the item details.

“Easiest” Dichotomous Items: The range between the lowest and highest location estimates in this easiest group spanned a large range of approximately 2.2 logits. All were low-demand items. The three easiest items assessed basic declarative (content) and procedural knowledge, such as recall of elementary ideas and direct reading of data. These items had location estimates below -3 logits on the scale. An interval of approximately 0.7 logits separated these three items from the remaining items. In the group, seven items were of the CROS item format. The two items with the highest locations in the group had a CRSS item format. Significantly, these items also measured similar procedural notions for discerning patterns and relationships from experimentally orientated data or models.

Another notable observation about items in the “easiest” group was the focus on measuring procedural and content knowledge. None of the items tested epistemic knowledge, which is least prominent in the curriculum. Given the familiarity of pupils with the two science practices of *giving scientific explanations* and *interpreting and analysing data and evidence*,

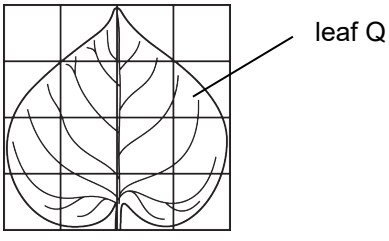
the presence of eight items evaluating these practices in the "easiest" group was unsurprising. The remaining two items gauged simple abilities such as recognising key variables in the third practice of *designing and evaluating investigations*.

"Hardest" Dichotomous Items: In comparison to the "easiest" group, locations of items in the "hardest" group were found above the mean location, ranging from 0.562 to 3.137. The interval between the two most difficult items, which had high cognitive demand, and the remaining items was approximately 1 logit, relative to a spread of 6 logits in difficulty estimates of items in the revised test. All items had a CRSS format and assessed a mixture of knowledge types and practices.

Interestingly, the only two items in the group that tested epistemic knowledge focused on identical and advanced competencies from the practice of *designing and evaluating investigation.* These competencies, which involved an understanding of valid scientific measurement, had been shown to be challenging, even among older science pupils (Sandoval & Çam, 2011). To illustrate the complexity involved, Figure 8.7 shows one of the two items. In this item, pupils struggled to explain why a grid with smaller squares produced a better estimate of leaf area than one with larger squares.

BVQ1

(b) Nurul said grid A with smaller squares will measure the area of leaf Q more reliably and accurately than grid B.



grid A

Give a reason why Nurul is correct. [1 mark]

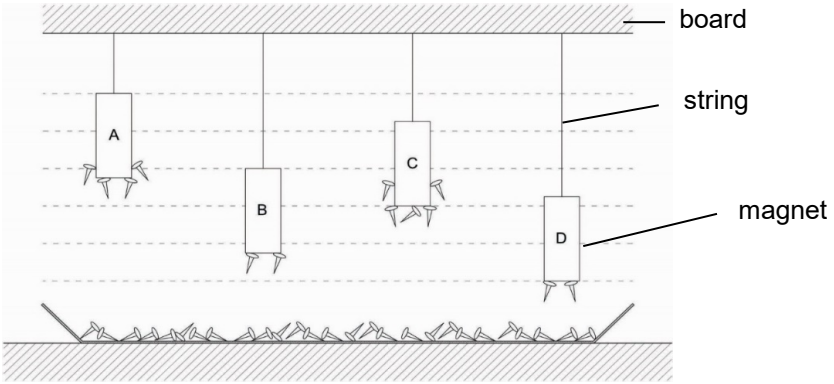
Figure 8.7. The figure above is part of BVQ1. The item's location estimate of 3.137 was the highest in the Revised PSRT.

A final comment on the "hardest" group was the presence of item BIIIQ2(b)(ii), which was highlighted earlier as part of a pair of items that differed from each other by the level of cognitive demand (see Table 8.4). Significantly, the other item of the pair, BIIIQ2(b)(i), was found in the "easiest" group. Both items, as shown in Figure 8.8, targeted pupils' proficiency in providing evidence to support a hypothetical outcome (claim) of an investigation. As indicated in the caption below the figure, the location difference between the item-pair was almost three logits, which was half the interval spanned by the item difficulties in the *entire* revised test. As explained earlier, the large difference in magnitude was unexpected. Unlike the first item, (i), which required a direct application of the scientific concept (i.e., magnet will attract more pins), pupils had difficulty producing the second piece of evidence (i.e., pins attracted to the magnet will attract other pins) in response to item (ii).

BIIIQ2

Ruby wanted to compare the magnetic force of four magnets A, B, C and D which were hung from a board on strings. The magnets were of the same size.

He set up the experiment below and observed the number of steel pins that were attracted by each magnet.



(b) Using the same set-up as shown in the diagram above, Ruby replaced magnet A with another magnet X of the same size.

State two possible observations that could be made if magnet X was a much stronger magnet than magnet A.

(i) Observation 1: _____

(ii) Observation 2: _____

[1 mark each, 2 marks in total]

Figure 8.8. Items (i) and (ii), of BIIIQ2(b), had locations of -1.523 and 1.481, respectively.

Comparison between "Easiest" and "Hardest" Items: Jointly, there was preliminary evidence that the observed proportion of low, medium, and high cognitive demand items in the two item groups accorded with expectations. Also, the results supported the postulated effect of item format. The combination of knowledge and practice assessed by items in both groups showed broad alignment with theoretical expectations. For instance, items testing epistemic ideas related to reliability and validity of data in the practice of *designing and evaluating investigation* were found to be the most difficult among the 20 items. Similarly, recall and simple application of basic content knowledge proved to be the easiest. Table 8.6 displays the items in both groups, as well as their location estimates and test features.

Table 8.6

Features of 20 items in the "easiest" and "hardest" groups in ascending location estimates

No.	Item	Location	SE	Knowledge	Practice	CD	Format
"Easiest" Items							
1	BIQ2(a)	-3.865	0.698	Content	1	Low	CROS
2	BVQ1(a)	-3.387	0.575	Content	1	Low	CROS
3	BIIIQ1(b)	-3.212	-0.356	Procedural	3	Low	CROS
4	BVIIIQ3(a)	-2.49	-0.999	Procedural	2	Low	CROS
5	BIIQ2(a)	-2.327	-0.59	Procedural	1	Low	SROS
6	BIIIQ1(c)(i)	-2.052	-0.049	Procedural	3	Low	CRSS
7	BIVQ1	-2.015	-1.054	Procedural	2	Low	CROS
8	BIXQ2(a)	-1.882	-0.473	Content	3	Low	CROS
9	BVIIQ2(d)	-1.822	-0.328	Procedural	3	Low	CRSS
10	BIIIQ2(b)(i)	-1.523	-0.058	Procedural	1	Low	CRSS
"Hardest" Items							
1	BIVQ2(d)	0.562	0.174	Content	1	Medium	CRSS
2	BVIIQ2(b)	0.737	0.172	Epistemic	2	Medium	CRSS
3	BVQ1(b)	0.842	0.169	Procedural	2	Medium	CRSS
4	BXQ2(b)	1.021	0.17	Procedural	1	Medium	CRSS
5	BIXQ1(b)	1.177	0.17	Content	1	Medium	CRSS
6	BIVQ2(c)	1.321	0.171	Content	3	Medium	CRSS
7	BIIIQ2(b)(ii)	1.481	0.173	Content	1	Medium	CRSS
8	BIXQ2(d)	1.557	0.17	Content	3	Medium	CRSS
9	BVIIIQ3(e)(ii)	2.566	0.19	Procedural	3	High	CRSS
10	BVQ1(c)	3.137	0.225	Epistemic	2	High	CRSS

Note. SE = standard error. CD = cognitive demand. Practice 1 – giving scientific explanation, practice 2 – designing and evaluating investigation, practice 3 – interpreting and analysing data and evidence. SROS is selected-response objective scoring, CROS is constructed-response objective scoring, CRSS is constructed response subjective scoring.

To further substantiate the general conclusions from the comparison, post-hoc analyses of the remaining 27 items with no recommended changes underwent a similar analysis. Overall, the results showed broad agreement. Items located near the “easiest” group tested less complex concepts or required fewer steps in processing using the three kinds of knowledge. Likewise, items with higher locations typically involved more than one level of cognitive processing or tested fundamental knowledge for which children hold common misconceptions. For example, in the selected-response item BIIIQ1(a) shown in Figure 8.9, pupils’ performance suggested that some were unsure whether gravity was a push or pull force. Appendix 24 shows the item features of all 47 items used in the analyses.

BIIIQ1(a)

Luke drops a piece of paper and measures the time for it to reach the ground.



Tick one box to show how the force of gravity acts on the paper.

The force of gravity acts on the paper by

pushing downwards	☐	pulling downwards	☐
pushing upwards	☐	pulling upwards	☐

[1 mark]

Figure 8.9. An item that tested understanding of gravitational force with a location of 0.248.

8.4.3 Threshold Estimates of Polytomous Items

A check of the location estimates of the eight two-mark items in the 86-item revised test, as displayed in Table 8.7, showed all values to lie between 0.227 and 2.362, a range that overlapped the mean location estimate of the pupils at 1.184. This result indicated that an average-ability pupil in the sample found the eight items to be a mix of easy and difficult. As mentioned in the data analysis section, the analysis first grouped items according to their development history from the iterative Rasch remediation and review procedure (see Chapter 7, subsection 7.5.1).

Subsequently, the items with revised scoring and recommendation for change were examined jointly using their observed threshold estimates to evaluate whether the hierarchical levels of knowledge or competencies exhibited in the responses of pupils accorded with theoretical understanding. The testlets were not part of the analysis, since there were no theoretical expectations before their formation from dichotomous items.

Table 8.7

Development history of the eight two-mark items in ascending order of location estimates

No.	Original Item	History of Development	Location Estimates (SE)
1	BIIQ3(b)(i) and BIIQ3(b)(ii)	Testlet formation	0.227 (0.132)
2	BIIIQ3(a)	Revised scoring	0.631 (0.123)
3	BIQ1(i) and BIQ1(ii)	Testlet formation	0.824 (0.144)
4	BVIQ2(b)	Change recommendation	1.113 (0.111)
5	BVQ3(c)	Revised scoring	1.496 (0.118)
6	BIIQ1	Change recommendation	1.53 (0.113)
7	BXQ2(a)(i)	Revised scoring	1.886 (0.126)
8	BVIIQ2(e)(ii)	Change recommendation	2.362 (0.152)

Note. SE = standard error.

Group 1: Features of Three Two-Mark Items that Underwent Re-scoring:

The three items shown in Table 8.8 were polytomous items that underwent re-scoring because the Rasch analysis indicated aberrations in them, such as disordered thresholds or unacceptable fit statistics. After re-scoring, the items had acceptable fit statistics and their CPCs evidenced ordered thresholds. The first two items in the table measured the competencies of practice 3: *interpreting and analysing data and evidence.* Specifically, the two sub-practices (and levels) targeted were *identify less straightforward patterns in data presented in various formats* (1.2) and *identify and explain how the evidence has been used to make the required conclusions and claims* (2.2). The third item assessed the sub-practice

of e*valuating abstract ideas or models used to explain processes or phenomena* (1.4) in the practice *of giving scientific explanations.*

Table 8.8

Features and item estimates of three two-mark items that underwent re-scoring

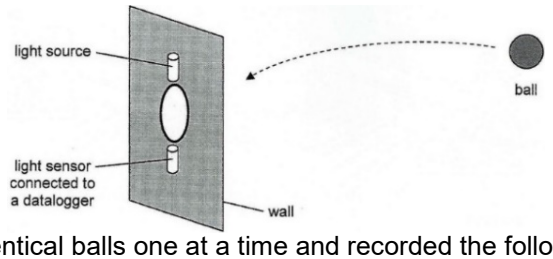
No. and Item	Threshold Estimates (SE)		Knowledge	Practice	Sub-practice and Level*	Cognitive Demand	Format
	First	Second					
1. BIIIQ3(a)	0.119 (0.104)	1.144 (0.112)	Procedural	3	2.2	High	CRSS
2. BVQ3(c)	0.938 (0.102)	2.055 (0.126)	Procedural	3	1.2	High	CRSS
3. BXQ2(a)(i)	0.977 (0.100)	2.795 (0.147)	Procedural	1	1.4	High	CRSS

*Note. Practice 1 – giving scientific explanation, practice 3 – interpreting and analysing data and evidence. *see text for sub-practice and level competencies. CRSS is constructed-response subjective scoring.*

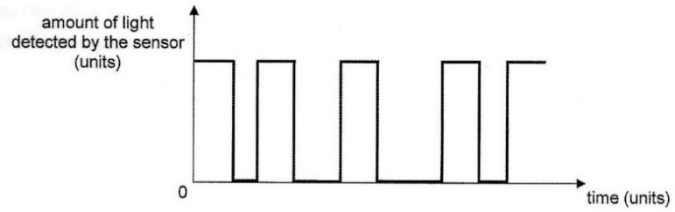
The three items in this group shared the same cognitive demand and item format, and all measured procedural knowledge (Table 8.8). Consequently, a post-hoc examination of the differences in threshold estimates considered only the differences among the science practices involved. As mentioned, the first item (shown in Figure 8.10) assessed the ability to interpret visual and graphical information (evidence) used to support claims. The threshold estimates of this item (indicated in Table 8.8) were lower than the corresponding estimates in the other two items. This result meant that both marks were easier to attain in this item than the other two polytomous items. Essentially, pupils had to relate the binary results in the graph (i.e., 0 units = no light, ≠ 0 units = light) to the set-up (i.e., ball through hole, ball not through hole).

BIIIQ3(a)

Muhammed set up a light source and a light sensor to count the number of balls going through a hole as shown.



Muhammed threw a few identical balls one at a time and recorded the following results.



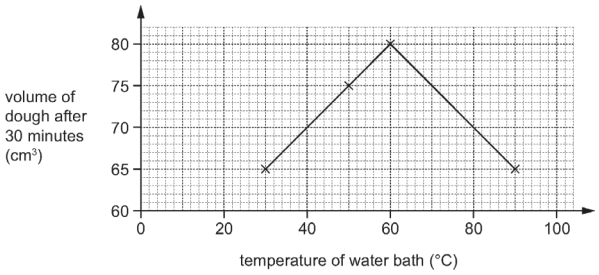
Explain how Muhammed used the results in the graph to find the number of balls going through the hole and if all went through the hole at the same speed. [2 marks]

Figure 8.10. An item which required processing of graphical information.

While the second item (Figure 8.11) measured a competency from the same practice as the first item, the specific competency tested pupils' ability to interpret complex data patterns.

BVQ3(c)

Her results are shown in the graph below.



Describe the relationship between the variables on the graph from 30°C to 90°C. [2 marks]

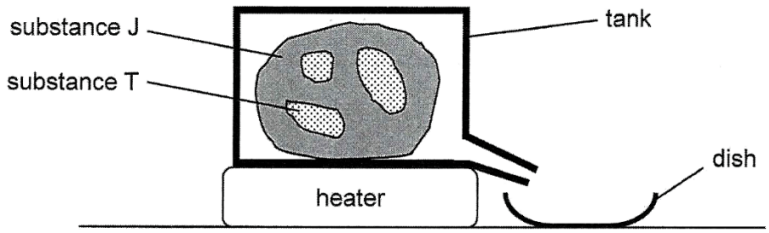
Figure 8.11. An item with a graphical stimulus composed of two linear relationship.

Like the first item in Figure 8.10, the second item (Figure 8.11) incorporated graphical data with two distinctive trends. However, the key variables involved multiple values, which may account for the relatively higher threshold estimates for the second item. Compared to the two threshold estimates in the first two items in the group, the corresponding values of the

third item were the highest, indicating this was the most difficult item. As mentioned, this item measured modelling proficiencies. See Figure 8.12 for the item.

BXQ2(a)(i)

John placed a solid made of substances J and T in the set-up as shown below.
Substance J had a lower melting point than substance T.



He wanted to obtain a solid made of substance J only. Explain how he could use the set-up to do so. [2 marks]

Figure 8.12. An item testing *the ability to draw out scientific ideas from an abstract model.*

Unlike the earlier two items, which measured competencies in practice 3, this third item measured advanced abilities in practice 1 relating to the use of abstract models with more than one step to describe processes. Like most advanced scientific competencies, proficiency in modelling requires a consistent emphasis in science instruction (Gilbert, 2004). On the other hand, graphical competency is aligned to the numeracy focus in the primary curriculum and thus receives early,[27] regular, and concurrent attention in primary mathematics education.

Moreover, limited studies (see, e.g., Lenton & Stevens, 2000) have suggested that prior exposure and practice in graph work facilitates transfer of learning to other, similar contexts. In sum, the combined understandings from the literature suggest that the evaluation of abstract and unfamiliar models might be more difficult for primary-aged pupils. This reason could have accounted for the higher threshold estimates of the third item.

[27] In Singapore, mathematics is taught in primary one, which is the first year of primary education. In contrast, science education begins two years later, in primary three.

Group 2: Features of Three Two-Mark Items with Change Recommendations:

This group originally consisted of polytomous items with acceptable fit statistics and ordered thresholds (see Table 8.9). The first two items in the table measured a common competency in practice 1. This shared competency (sub-practice 2.3) is to "*identify and explain how the evidence has been used to make the required conclusions or claims"*. The third item assessed a competency (sub-practice 3.3) in practice 2 that requires pupils to explain how the collection of data took into consideration potential sources of errors. The response patterns to the three items and other model-fit information, such as response frequencies (i.e., number of pupils achieving each mark category), suggested that further improvements could be made. A brief presentation of the reasons why each item warranted change recommendations follows the table as necessary background for the subsequent analysis.

Table 8.9

Features and item estimates of three two-mark items with change recommendations

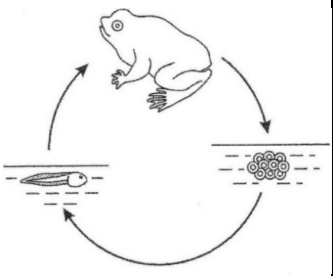
No. and Item	Threshold Estimates (SE)		Knowledge	Practice	Sub-practice and Level*	Cognitive Demand	Format
	First	Second					
1. BVIQ2(b)	0.914 (0.107)	1.311 (0.117)	Content	1	2.3	Low	CRSS
2. BIIQ1	1.387 (0.109)	1.673 (0.125)	Content	1	2.3	High	CRSS
3. BVIIQ2(e)(ii)	0.510 (0.104)	4.214 (0.218)	Epistemic	2	3.3	High	CRSS

*Note. Practice 1 – Giving scientific explanation, practice 2 – designing and evaluating investigative approaches. *See text for sub-practice and level competencies. CRSS is constructed-response subjective scoring.*

An examination of the first item – BVIQ2(b), shown in Figure 8.13, indicated that non-credited or partially credited responses to the second item contained repeated ideas or out-of-context answers. To address this concern, the item may be modified to focus on the assessment of only a single advantage.

BVIQ2(b)

The diagram shows the life cycle of a frog.



The young of the frog lives in water while the adult lives on land. This helps the frog to survive better.

Suggest two advantages for the young and the adult to live in different surroundings. [2 marks]

Figure 8.13. A polytomous item of low cognitive demand with change recommendations.

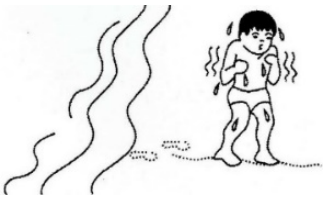
The observed responses on the second item: BIIQ1 (see Figure 8.14) indicated that many pupils had not understood the central assessment concept of evaporation (that heat is gained by the water on the body) but could attain the next mark point by producing responses that referenced “higher rates of evaporation” caused by the strong wind.

BIIQ1

Eric went for a swim. When he came out of the water, he felt cold.

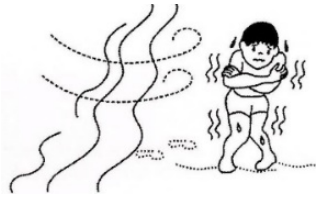


his body was dry before swimming



his body was wet after swimming

At that moment, a strong wind blew, and he felt even colder.



a strong wind blew over his wet body

Explain why he felt even colder when a strong wind blew. [2 marks]

Figure 8.14. A polytomous item of high cognitive demand with change recommendations.

An additional concern with the second item was the observed deviation between the actual means of the class intervals and the model expectations. More able pupils beyond 1.4 logits did better than expected, while the reverse was true for less able pupils with an average ability around 0.8 logits. The over-discriminatory fit of the responses in the ICC suggested that the assessment of the two scientific ideas in a complex item context

hampered the elicitation of ideas from weaker pupils. Recommended changes to the item could include partitioning the context in its stem or revising the mark scheme.

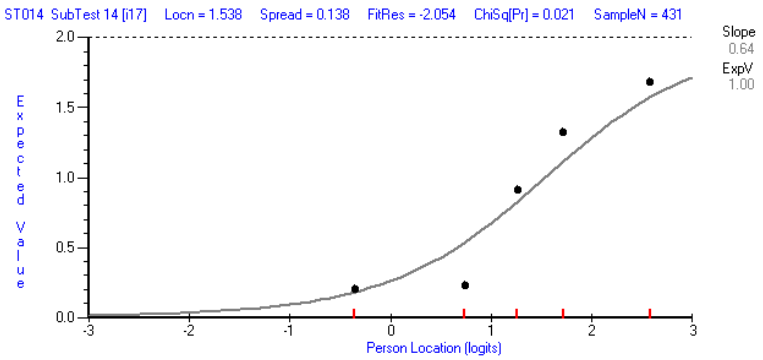


Figure 8.15. ICC of BIIQ1.

Unlike the concern identified in the first two items of this group, the third item had a significantly larger second threshold estimate (see the CPC in Figure 8.16). As mentioned in earlier analyses of dichotomous items, the span of location estimates between the most difficult and easiest items in the test was 6 logits. Hence, the observed interval between the threshold estimates of this item was more than half the span of the scale. Unsurprisingly, given the large second threshold estimate, only 10 out of 173 pupils who attempted the item attained the second mark and hence obtained full credit.

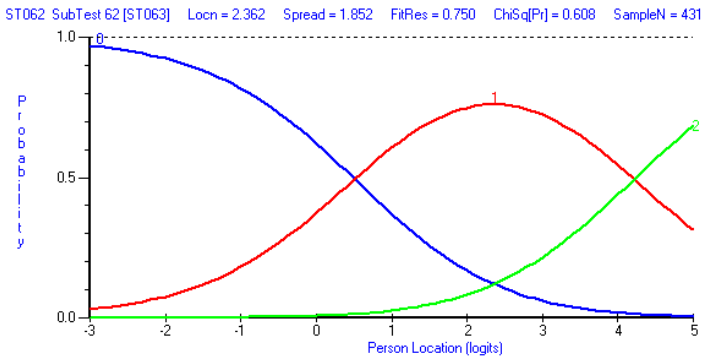


Figure 8.16. CPC of BVIIQ2(e)(ii).

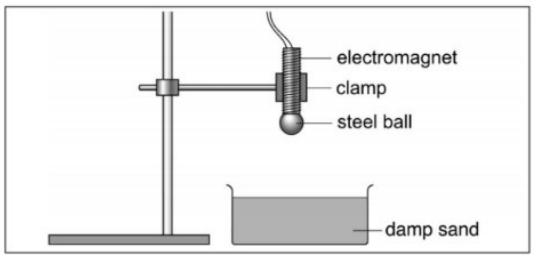
Given the issues identified in the three items, any interpretation of the impact of their item features on threshold estimates could be biased. Therefore, limited analysis was performed on the three items. For instance, a comparison of the features of the first two items in Table

8.9 showed they differed only by their cognitive demand. It was thus expected that the first item (of low demand) would attain lower threshold estimates than the second item (of high demand). Although this predicted pattern was observed, the highlighted concerns with the responses and other fit indicators suggested the results might not have stemmed from differences in cognitive demand alone; it was also possible that pupils interpreted the question incorrectly, or that there was a faulty mark scheme.

Similarly, an examination of the threshold estimates of the third item (see Figure 8.17) in the group, and the noted performance of the pupils across the three response categories suggested that the order could be attributed to reasons not considered by the theory.

BVIIQ2(e)(ii)

Jack said that the ball could be dropped using an electromagnet instead of dropping it by hand.



Explain why this would make the results more accurate. [2 marks]

Figure 8.17. A polytomous item with change recommendations.

Like the other two items, this item had a CRSS format. However, the item measured epistemic knowledge in the practice of *designing and evaluating investigations*. While the value of its first threshold was the lowest among the three items, its second threshold was significantly greater. The ease of obtaining the first mark point was consistent with the intended assessment objective, as the answer was obvious from the stem and the visual stimulus. However, although the second epistemic notion assessed (second mark point) was intentionally challenging, the exceptionally difficult second threshold estimate of the item meant its actual difficulty had been theoretically under-estimated. Possible reasons could be the lack of relevant instruction, or an overly stringent marking requirement.

Collectively, the evidence suggested that its features, such as knowledge type, science practice, and cognitive demand, were inadequate to fully account for its threshold estimates.

8.4.4 Analysis of Pupils' Responses

This final analysis focused on the actual performance and responses of pupils in the three science practices using content, procedural, and epistemic knowledge. Specifically, there was a comparison of pupils' performances on the items that purported to test similar competencies outlined in the practices or sub-practices. Responses to specific items will be highlighted as illustrations of key observations and results. Unless otherwise indicated, all the items identified for discussion in this subsection had no relevant issues (e.g., statistical anomalies) that would impact on drawing inferences from this analysis. Each of the practices and their sub-practices will be discussed in turn. Details of the anticipated types of reasoning engendered by the practices and sub-practices are provided in the literature review (subsection 2.5.2), and further elaborated in Chapter 4 (subsection 4.3.3).

Practice 1: Giving Scientific Explanations

Sub-practice 1.1: Identify, use, and evaluate models and representations

As expected, performance on items that measured competencies in this sub-practice showed an increase in difficulty as the level of abstraction in schematics, unfamiliarity with the context, and the assessment demand (e.g., create versus describe) increased. The items shown in Figure 8.18 illustrate these general conclusions. Both measured procedural knowledge and competencies in this sub-practice. The location estimates and 95% C.I.'s of BIVQ3(a) and BIVQ3(b) were -1.998 [-2.613, -1.383] and -1.022 [-1.473, -0.571] respectively. The results meant that the first item was likely easier than the second for the pupils, although there was some small overlap in their C.I.'s.

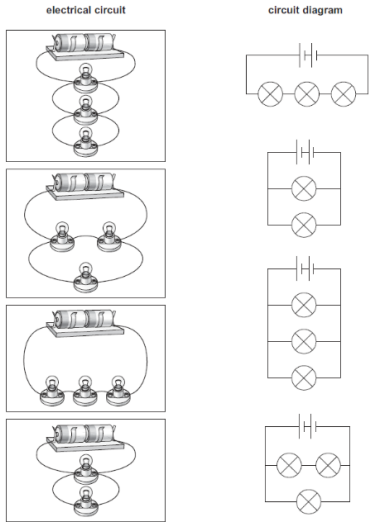
The easier item assessed the ability to identify simple models, corresponding to level 1.1.2 in this sub-practice, while the other item tested understanding of an abstract model, corresponding to another level (1.1.3) in the same sub-practice. The design of the first item,

BIVQ3(a), juxtaposed the electric circuit and circuit diagrams to facilitate the drawing of analogical similarities and formation of understanding, or “rules”, to guide pupils in their responses. The realism of the electric circuit diagrams was also intended to help the identification process.

In contrast, the circuit diagram in the second item, BIVQ3(b), consisted of a complex arrangement of symbols depicting electrical component, which required pupils to retrieve and apply rules inductively from knowledge. Overall, the results provided some tentative evidence that the order of difficulty between the two items was as expected. This expectation was premised on the model intricacy and the targeted proficiency of the items, which impacted on the cognitive load and reasoning exhibited.

BIVQ3(a)

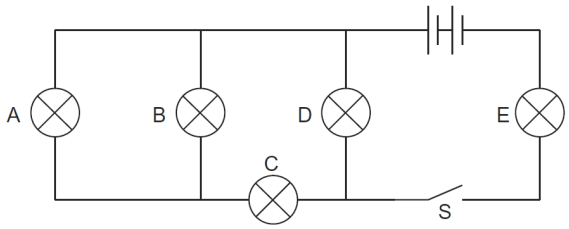
Draw a line from each electrical circuit to the correct circuit diagram. Draw only **four** lines.



[1 mark]

BIVQ3(b)

(a) Max built circuit 1 as shown below.



circuit 1

He closed the switch, S, and all the bulbs turned on. One of the bulbs then broke, and all the bulbs went off.

Which bulb must have broken?
Circle the letter below.
A / B / C / D / E

[1 mark]

Figure 8.18. Two items testing competencies in the first sub-practice of practice 1 – giving scientific explanation.

Practice 1: Giving Scientific Explanations

Sub-practice 1.2: Recall and apply appropriate scientific knowledge

As expected, analysis of items testing competencies in this sub-practice showed that the construction of explanations crucially depended on pupils’ ability to deduce the appropriate

conclusions from the clues, or “facts”, in the item context. Therefore, the presence of useful information such as pictorials or data providing “mental scaffolds” made it easier for pupils to respond correctly. Similarly, familiarity with the item context facilitated the activation of the relevant mental schemas and understanding.

As mentioned in Chapter 2 (subsection 2.5.2), studies show that deductive reasoning becomes ineffectual when pupils have a poor grasp of the knowledge tested. Evidence from the responses on these items, which predominantly assessed deductive thinking, generally conformed to this finding. The question in Figure 8.19 illustrates these observations.

BVQ2

Alieen brushed her teeth in front of a mirror as shown in Diagram 1. After she took a hot shower, the mirror became fogged as shown in Diagram 2.

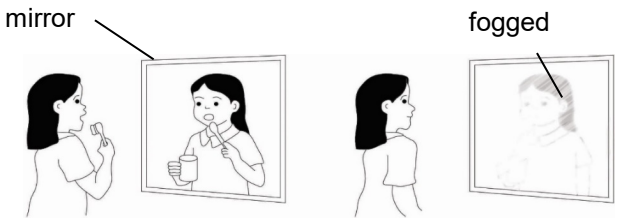


Diagram 1 Diagram 2

(a) Explain how the mirror became fogged. [1 mark]

(b) Give a reason why the mirror became clear again after some time. [1 mark]

Figure 8.19. Two items in the same question testing the recall and application of content knowledge in the changes of state of water.

The two items in Figure 8.19 tested physical science concepts related to condensation and evaporation. The intentional inclusion of two diagrams that directly demonstrated the condensation process (i.e., appearance of fog, which is water droplets) was expected to assist pupils’ reasoning. In contrast, the answer to the second item required pupils to infer that the droplets had evaporated. The revised location estimates and 95% C.I.’s for BVQ2(a) and for BVQ2(b) were -0.222 [-0.598, 0.154], and 0.872 [0.541, 1.203] respectively. These results showed that the observed order of item difficulty between the two items was as intended, even after accounting for the uncertainty of the estimates. Further, underlying the importance of having a strong grasp of the tested content was the general observation that pupils who could not answer the first item also did not respond correctly to the second.

Practice 2: Designing and Evaluating Investigative Approaches

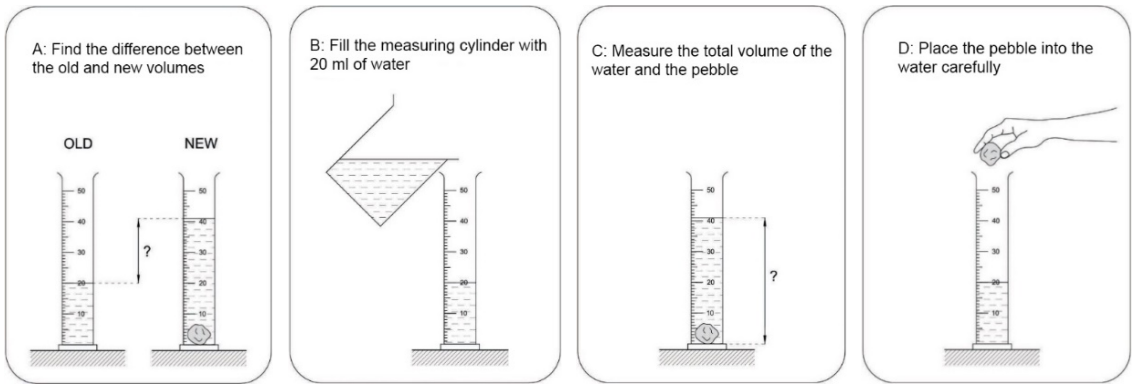
Sub-practice 2.1: Plan for effective scientific investigations or evaluate approaches

In this first of three sub-practices in practice 2, the types of reasoning elicited depended on the assessment objectives, item context, and anatomy. The item in Figure 8.20 measured competencies from this sub-practice and instantiated two fundamental forms of reasoning linked to normal intellectual development: conservation and seriation. The revised location estimate and 95% C.I. of this item was -2.015 [-2.634, -1.396]. These results suggested that this was an easy item for an average-ability pupil with a mean estimate. This was an unsurprising finding; given their age, the pupils would be expected to be in the pre-formal or formal-operational stages, in which these forms of fundamental reasoning have been mastered (Piaget, 1972). Proficiency on this item served as a milestone to indicate pupils' readiness for more complex types of reasoning.

BIVQ1

Sally was shown four cards, A, B, C and D. Each of the cards describes one step of an experiment to find the volume of a pebble (stone).

Card A Card B Card C Card D



Help Sally arrange the cards in the correct order to find the volume of the pebble.

Step 1: Card _____ Step 2: Card _____

Step 3: Card _____ Step 4: Card _____ [1 mark]

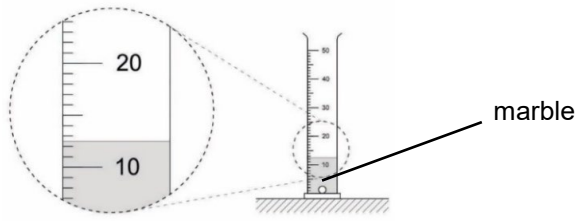
Figure 8.20. An item demonstrating the assessment of conservation and seriation reasoning.

The results of two other items, shown in Figure 8.21, supported this inference regarding pupils' readiness. Building on basic reasoning, both items – BIXQ1(c) and BIXQ1(d) –

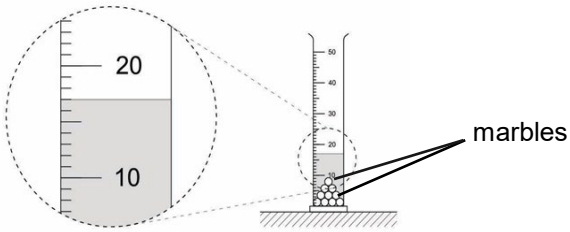
tested different, more sophisticated competencies in similar contexts, involving volume measurement using measuring cylinders. However, the second item was anticipated to be more difficult, as the competencies involved epistemic notions of instrumental precision. The revised location estimates and 95% C.I.s of both items were 0.912 [0.577, 1.247] and 2.414 [2.049, 2.779] respectively. Therefore, not only was the order of difficulty among these two items and the item in Figure 8.20 as expected, but the estimates were also significantly different. An average-ability pupil in the current sample would probably be able to respond correctly to the first item but would find the second difficult.

BIXQ1

Betty and Paul have a bag of similar marbles. They wanted to find the volume of a marble.
Betty then puts a marble into the measuring cylinder to find its volume.
Her result is shown below.



Paul used ten marbles in the set-up below to find the volume of one marble. His result is shown below.



(c) Describe how the result will be used to find the volume of one marble. [1 mark]

(d) Give a reason why Paul's method is more accurate than Betty's. [1 mark]

Figure 8.21. BIXQ1(c) and BIXQ1(d) assessed procedural and epistemic knowledge using competencies outlined in practices 3 and 2 respectively. The first two items are not shown.

Practice 2: Designing and Evaluating Investigative Approaches

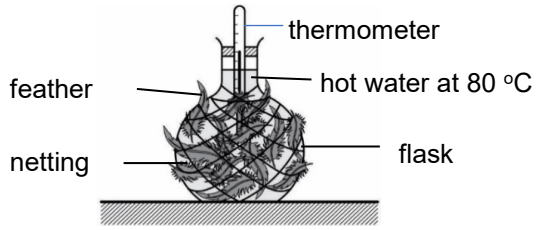
Sub-practice 2.2: Identify and manipulate variables within an investigation

The main reasoning tested in this sub-practice was the COV strategy. It was predicted that the types of variable in the items would influence mental processing and cognitive load, thereby affecting item difficulties. Generally, the results of the analysis supported the

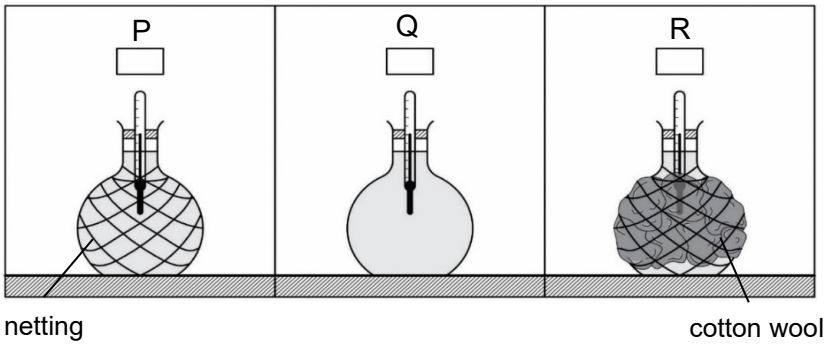
prediction. A comparison of two such items that shared the same item format and tested closely related science concepts illustrates this argument. See Figure 8.22 for the two items.

BIVQ2

Ah Yong wanted to study how feathers affect the way hot water at 80°C loses heat. The diagram below shows one set-up of his experiment.

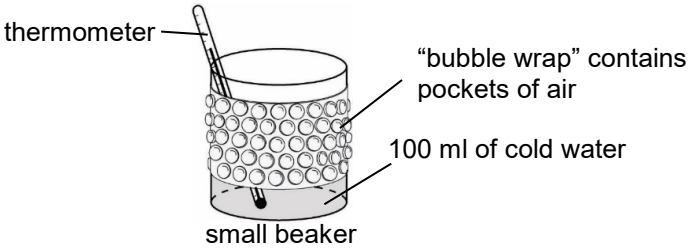


(a) Which other set-up without feathers, P, Q or R, as shown below must he use in his experiment? Choose your answer by ticking (✓) the box.



BVIIIQ2

John wants to study whether the material, 'bubble wrap' affects how cold water gains heat. The diagram below shows one set-up of his experiment.



(a) Which other set-up, P, Q or R, must he use in his experiment?
Choose your answer by ticking (✓) the box.

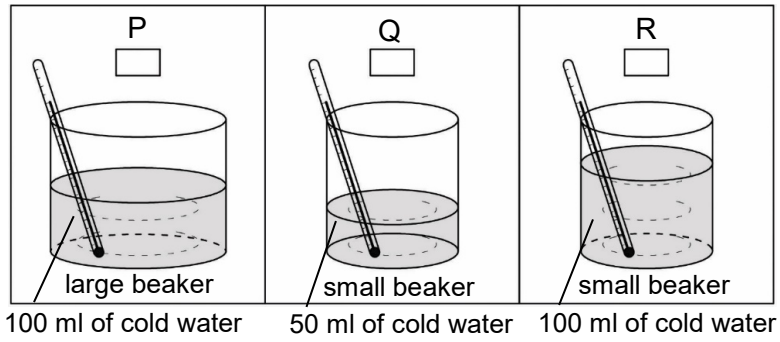


Figure 8.22. Two single-mark items testing the control-of-variable (COV) strategy.

The two items had a three-option, selected-response format, and tested procedural understanding of a fair test through the appropriate comparison of variables in experimental set-ups. Specifically, the items focused on the effect of a material (i.e., feathers, bubble wrap) on the loss or gain of heat within water in a container. However, a major difference between the two items was the choice of variables in the three options. Although all options in the items did not include the material for comparison, two of the options in item BIVQ2(a) had alternatives – netting, or cotton wool with netting. These alternatives (options P and Q) were intended as plausible distractors for pupils who were unclear about the function and role of the netting as a variable.

In contrast, there were no alternatives for the bubble wrap among the three options of BVIIIQ2(a). Instead, the volume of water and size of the beakers were the manipulated variables. Though the key (correct option) was without a material, it had an identical set-up as the stimulus diagram. The stem provided the information that the material was the change variable; this similarity was expected to facilitate COV reasoning. Therefore, the first item should have been more difficult than the second, because of the combination of variables presented in the options.

The results showed that the first item (revised location estimate = 0.419, 95% C.I. from 0.074 to 0.764) had a statistically significant higher difficulty estimate than the second item (revised location estimate = -1.224, 95% C.I. from -1.724 to -0.724), thereby supporting the predicted order of difficulty. Not surprisingly, a further examination of the two distractors in the second item showed them to be less effectual, even among less able pupils.

However, in other, similar items that investigated pupils' facility and reasoning with variables in the Revised PSRT, robust inferences about the impact of variables on item difficulties could not be drawn. Some of the reasons are explained in the following discussion using the two items in Figure 8.23.

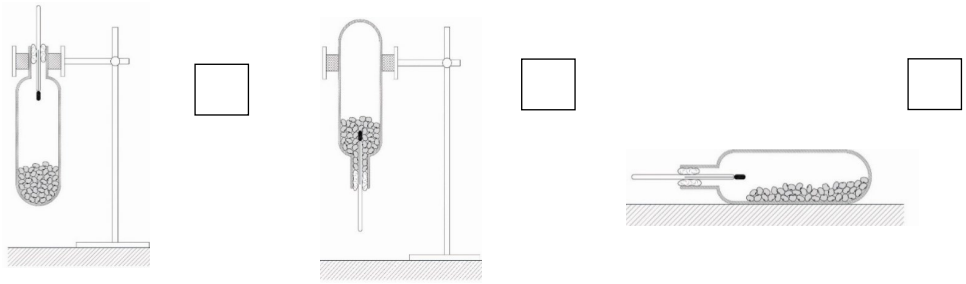
BVIIIQ1

John wants to study the change in temperature from heat given off by germinating seeds. He used the following things.

Thick-walled bottle thermometer germinating seeds cotton wool



(a) (i) Which set-up should John use to measure the change in temperature most accurately? Tick (✓) the box and explain your answer.



[1 mark]

(ii) Explanation: _____ [2 marks]

Figure 8.23. Two items that measured pupils' reasoning with experimental variables.

The items in Figure 8.23 were expected to have higher location estimates than the two examples shown in Figure 8.22 due to the lack of a control set-up for comparison. Contributing to the difficulty in the second item of Figure 8.23 was the requirement that pupils supply a reasoned explanation for their choice of answer in the first. The location estimates and 95% C.I.'s of the first and second items were 1.181 [0.856, 1.506] and 1.803 [1.593, 2.013] respectively. The results thus showed that the statistically different estimates conformed to the predicted order of difficulty. Unfortunately, inferences from the results were hampered by anomalous item-fit statistics observed in the two items, as briefly described next.

Responses to the first item produced an under-discriminating model fit in its ICC, as shown in Figure 8.24A. The resulting fit residual was 2.612, which was out of the acceptable range of ±2.5. Pupils' reasoning in the second item produced an ICC with a fit residual of 3.783 (see Figure 8.24B). Its CPC, shown in Figure 8.24C, also indicated disordered thresholds.

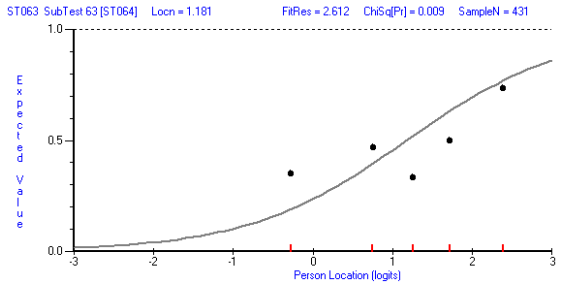


Figure 8.24A. ICC of BVIIIQ1(a)(i).

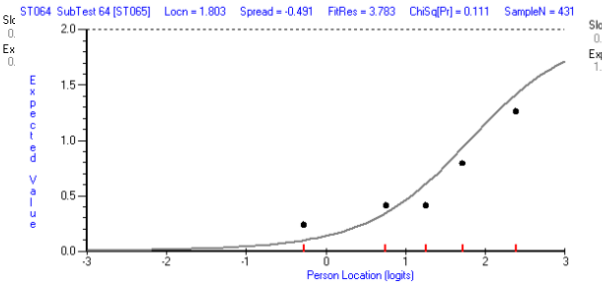


Figure 8.24B. ICC of BVIIIQ1(a)(ii).

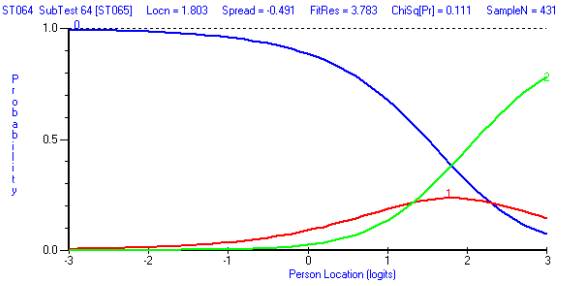


Figure 8.24C. CPC of BVIIIQ1(a)(ii).

An examination of the written responses to the second item indicated that the mark scheme, which credited answers that referred to or implied contact between the thermometer and the seeds for accurate heat measurements, may not have functioned as intended. Jointly, the performance of pupils on both items indicated that pupils who responded incorrectly did not focus on the function of the thermometer, but instead noted the positioning of the bottle or the seeds. There was also evidence of guessing in the first item.

In effect, the evidence suggested that the item context was not adequately salient to activate the relevant schemas involving scientific knowledge and competencies. Instead, incorrect answers made reference to the visual placement of objects without invoking the necessary scientific concepts. This example highlighted the unintended impact of poor item design on the functioning of items. At a broader level, it underscored the importance of considering the fit statistics of the items when comparing the theoretical and actual effect of item features on item difficulties. For instance, inferences drawn from the two items shown in Figure 8.22 were stronger, because these items had no detected item flaws from the statistical analyses and subsequent qualitative examination of the items and raw responses.

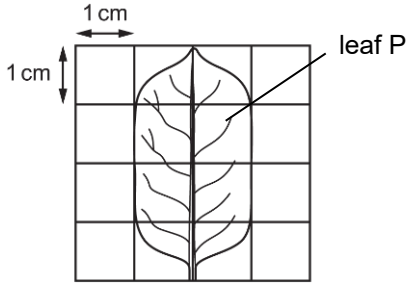
Practice 2: Designing and Evaluating Investigative Approaches

Sub-practice 2.3: Collect precise and reliable data during the gathering of evidence

The competencies in this sub-practice included making whole-number measurements, identifying appropriate measurement intervals, and explaining how sources of errors could be reduced in data collection methods. Like competencies from the other sub-practices and depending on the assessment objectives of the items, a few reasoning types could be involved. For instance, the predominant types tested in the two items shown in Figures 8.25 and 8.26 were quantitative and qualitative reasoning respectively.

BVQ1

Nurul used square grids drawn on clear plastic to measure the area of leaf P.



grid A

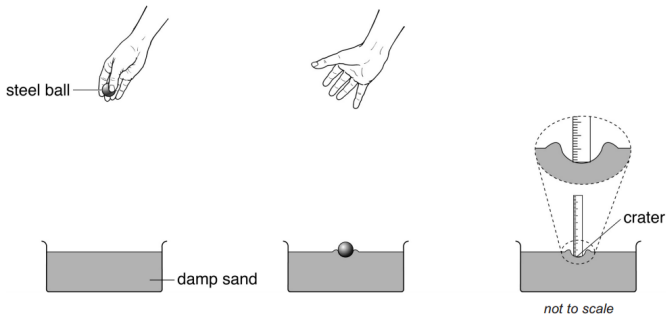
(b) Each square on grid A represents 1 cm^2. Estimate the area of the leaf P using grid A.

Estimated area of leaf P = _____ cm^2 [1 mark]

Figure 8.25. An item testing the ability to estimate the area of a leaf using square grids.

BVIIQ2(e)

Jack and Aneesa took turns to drop a steel ball into trays of damp (wet) sand. They measured the depth of the craters (holes) made by the steel ball using a ruler.



(i) Jack removed the steel balls using his fingers. Then he measured the depth of the crater. Aneesa said he should use a magnet instead of his fingers.

Explain why using a magnet to remove the ball would make the results more accurate. [1 mark]

Figure 8.26. An item testing procedural understanding related to accuracy in measurement.

The first item in Figure 8.25 was expected to be easier than the second in Figure 8.26 because of pupils' relative familiarity with numerical operations in estimation and the measurement of area. As mentioned earlier, these mathematical competencies were concurrently honed from consistent and regular instruction in the primary curriculum. On the other hand, the second item tested pupils' ability to justify the validity of collected data. Successful management of errors in measurement is an important and advanced epistemic competency for primary pupils. Unexpectedly, the results showed that the first item (revised location estimate = 0.842, 95% C.I. from 0.511 to 1.173) was harder than the second (revised location estimate = 0.342, 95% C. I. from -0.009 to 0.693), although there was no evidence that the location estimates were significantly different.

Inspection of the responses to the first item revealed a large group of pupils had inaccurately estimated the leaf area to be 8 $cm^{2.}$ This observation suggested that cross-disciplinary transfer and application of mathematical concepts might be difficult for many in the pupil sample, depending on the context of the science item. Findings from limited studies support this conclusion (see, e.g., Lenton & Stevens, 2000). Regardless of the observed order of difficulty between the two items, many pupils were challenged by the knowledge and competency tested in the second item, as evidenced by the noticeably longer responses and pronounced usage of layperson language seen in the raw responses. Unnecessarily lengthy responses were similarly noted in other items assessing similar knowledge and competencies (see, e.g., Figure 8.7), which signalled pupils' attempts to articulate weak or poorly understood knowledge. Jointly, the responses to both items in Figures 8.25 and 8.26, highlighted how item contexts, and mathematical and language competencies mediated scientific reasoning in unexpected ways, affecting the intended order of difficulties.

Practice 3: Interpreting and Analysing Data and Evidence

Sub-practice 3.1: Process, transform and analyse data

This first sub-practice described competencies involved in drawing conclusions from the processing and analysis of data, which included the interpretation of data from different formats (e.g., pie charts, pictograms) and the use of patterns in data for predictions.

Underlying these complex competencies were various advanced reasoning types. Two of these – pattern identification and extrapolation – were seen in the responses to the two items (in Figure 8.27) to be discussed next. These items were part of a multi-item question in the anchor block. Therefore, all pupils in the sample attempted this item.

AQ3

	Experiment			
	A	B	C	D
Length of string (cm)	25	25	50	75
Mass of weight (g)	100	50	100	100
Time for 10 complete swings (s)	10.0	10.0	14.2	17.4

(b) Look at the results table.
Describe how the length of the string affects the time for ten complete swings. [1 mark]

(e)(i) Use Joe's table of results to predict the times for 10 complete swings in experiment F.
Write your answer in the table below.

	E
Length of string (cm)	25
Mass of weight (g)	25
Predicted time for 10 complete swings (s)	

[1 mark]

Figure 8.27. Two items that assessed pattern identification and extrapolation competencies. The key in (b) is that as the length gets longer, the time for 10 complete swings becomes longer. The key in (e)(i) is 10.0 seconds.

Competencies to discern the implicit rules underlying pattern identification and the extrapolation of results have similarities with analogical and inductive reasoning. The first item shown in Figure 8.28 required pupils to describe the observed pattern in the data, while the second assessed the competency to make a valid prediction by selectively drawing information from the data. Therefore, as the assessment requirements were less demanding in the first item, it was expected to be easier than the second item. The location estimates and 95% C.I.'s of the first and second item were -0.690 [-0.960, -0.420] and 0.186 [-0.041, 0.413] respectively. These statistically different values supported the postulated order of difficulty.

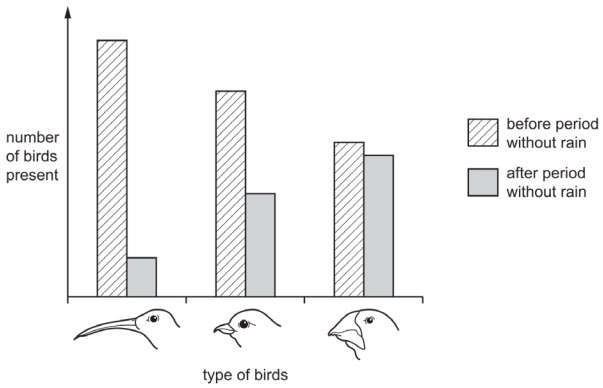
Practice 3: Interpreting and Analysing Data and Evidence

Sub-practice 3.2: Evaluate evidence to construct and critique conclusions, hypotheses and arguments

The second and last sub-practice of practice 3 described competencies for the critical evaluation of evidence. These competencies involved a combination of various types of reasoning, depending on the item context. For instance, even though both items to be discussed assessed similar competencies, the first item (Figure 8.28) required the demonstration of proportional reasoning, while the second (Figure 8.29) emphasised reasoning associated with variable separation and control.

A2(b)

The island went through a period without rain. The graph shows the number of three types of birds present on the island before and after the period.



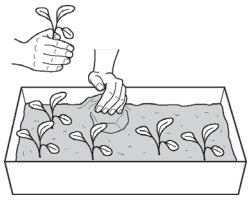
(ii) Explain how the graph helps you identify the correct bird. [1 mark]

Figure 8.28. An item assessing the justification of an epistemic claim from evidence.

The first item assessed the justification of an epistemic claim using evidence from a graphical stimulus. In contrast, the claim justified in the second item was procedural. Specifically, the evidence was drawn from a table of experimental data. Since pupils should be more familiar with procedural knowledge, the first item was anticipated to be more difficult than the second. The location estimates and 95% CI's of both items were 0.011 [-0.359, 0.381] and 0.223 [0.104, 0.550] respectively. Although the results conformed to the expected order of location estimates, there was no statistical evidence that these estimates were significantly different.

BVIIIQ3

A farmer wanted to find out the best conditions for growing a type of plant.



He took four trays and planted eight plants in each. The results of his experiments are shown below.

	Variables			
Tray	Light level	Air temperature (°C)	Soil moisture	Number of plants alive after seven days
A	medium	25	moist	8
B	medium	25	dry	6
C	medium	45	moist	2
D	medium	45	dry	0

(a) The farmer concluded:



Is the farmer able to make this conclusion from his experiment? Explain your answer. [1 mark]

Figure 8.29. An item assessing the justification of a procedural claim from evidence.

8.5 Discussion

Overall, there was broad agreement between the predicted and actual order of the average difficulty estimates of items with respect to the item features of cognitive demand, item format, knowledge type, and science practice. Also, the mean location estimates between items with high and low cognitive demand, as well as between CRSS and CROS items, were statistically different, further supporting the theory. However, on average, SROS items did not turn out to be easier than CROS items, which was not consistent with theoretical expectations.

As predicted by the postulated construct of scientific reasoning in this study, items measuring epistemic knowledge and competencies related to the practice of *designing and*

evaluating investigation produced the highest mean location estimates. Although the mean estimates of items measuring the remaining two knowledge types were of the intended order, they were not statistically different. Similarly, after accounting for uncertainties in the estimates, the average difficulty of items testing the other two science practices was likely to be similar. Taken together, the unexpected findings suggest that efforts to ensure a balance of familiar knowledge and competencies in the PSRT to measure pupils of mixed abilities could have been successful. For instance, as indicated in the TOS of the PSRT (Chapter 7, subsection 7.4.2), the numbers of low-demand items testing either content knowledge or procedural knowledge were approximately the same.

The analysis of the two item types in the revised test used separate sets of procedures. The first procedure on dichotomous items analysed the item difficulties of all such items from the test. While this examination of the 78 items produced some general inferences that supported the associations postulated for item format and cognitive demand on item difficulty, there were no consistent trends to conclude what the impact of science practices and knowledge types might be. A subsequent systematic analysis of the distribution of individual item features within the five intervals of the scale produced stronger evidence supporting the expected effect on item location estimates. As proportions of the total items within an interval, medium- and high-demand items were found in increasing proportions across intervals towards the end of the scale with higher location estimates. The reverse pattern was seen for items of low cognitive demand.

The distribution of items with CRSS and CROS formats, as proportions of the total items in an interval, across the five intervals on the scale also followed the expected trend: CRSS item proportions progressively became larger in intervals towards the higher end of the scale. Also, CROS items were only observed in the lower intervals. However, unlike items of the other two formats, SROS items appeared in all intervals. This unexpected result suggests that pupils encountered unintended difficulties with items of this format, supporting similar inferences from the earlier analysis about the average item difficulty. Finally, as predicted by the scientific reasoning model, the proportions of those items, out of all items

in an interval, assessing *epistemic knowledge* and competencies from the practice of *designing and evaluating investigations,* became progressively larger in successive intervals towards the more difficult end of the scale.

The next two analyses collected further evidence and strengthened inferences from earlier procedures by analysing a subset of 47 dichotomous items that passed qualitative checks and had acceptable fit statistics. In one of these two procedures, which involved the identification of items that shared similar features or whose features differed by only one, some evidence surfaced to support the postulated effect of item format and cognitive demand on item difficulty.

Similarly, the evidence from the second and final procedure, which compared the features of the 10 most difficult items with another 10 of the easiest items in this subset, produced preliminary evidence that reinforced some of the earlier conclusions about the influence of the item features on the difficulty of dichotomous items. For instance, all the easiest items were of low cognitive demand, while the hardest items had a mixture of medium and high demand. Similarly, all 10 of the most difficult items had a CRSS format, whereas a mix of formats was present among the easiest items. Notably, items testing a combination of content knowledge and recall competencies were only evident in the “easiest” group. In comparison, the hardest group had items assessing epistemic knowledge using investigative competencies.

Analysis of the two-mark items took place in groups depending on their developmental history. The results from the polytomous items that were re-scored produced some tentative evidence that the general order of threshold estimates aligned to theoretical expectations about the role of science practices. This was an encouraging result given that polytomous items awarded marks according to the hierarchy of knowledge or competencies evidenced in the responses. Unfortunately, limited conclusions could be drawn to support the postulated effect of the other three features as the three items in the group tested the same knowledge type and were of the same format and level of cognitive demand. Given

concerns that any inferences drawn might be biased or incorrect, polytomous items that had change recommendations or were formed from dichotomous items received limited or no analysis. Due to the small number of available polytomous items for analysis, there was thus only some evidence to support the hypothesised role of science practices.

The last analysis examined the responses of pupils to the items in the test for evidence of the expected reasoning underlying each of the sub-practices in the three science practices. This examination involved a comparison of items that measured the same sub-practice to study the impact that differences among items had on the elicited reasoning and hence the item difficulties. Some of the differences included abstraction levels in science models, availability of salient scaffolds (e.g., visuals, data) to guide schema activation, and the selective combination of science variables. Relevant psychometric data and information from the science education, mathematics, and literacy literature were also used to support the inferences made from this qualitative examination.

The comparative analysis among items measuring the same competencies produced evidence that the item difficulties generally accorded with theoretical expectations. In turn, this evidence supported the claim that the elicited reasoning matched those intended by the competencies of the three science practices and their sub-practices. Still, this analysis was constrained, as limited inferences could be drawn from items that had unresolved misfit to the model. As explained in the discussion with some items in the results section (see Figures 8.8 and 8.23 respectively), analysis of the misfit provided possible reasons to account for findings that did not conform to the theoretical model. These reasons included flaws in item design, unfamiliar item contexts, and the unintended influence of mathematical and language proficiencies on the mediation of pupils' performance on science items.

This study produced both quantitative and qualitative evidence to evaluate whether the test items measured the theorised construct of scientific reasoning. Analytical procedures at the test and item levels investigated the alignment between the observed and hypothesised effects of five item features on the item difficulty. There were stronger and less ambiguous

inferences about the influence of four of the item features from responses to dichotomous items than polytomous items due to a significantly larger pool of items. The lack of statistically sound polytomous items that did not require further amendments hampered the drawing of wider inferences about the effect of cognitive demand on the threshold estimates. Also, as the majority of the polytomous items were of a single item format (i.e., CRSS), possibilities to discern if this item feature functioned as postulated were limited.

8.6 Conclusions

This chapter has addressed the second research aim, which was to investigate if the items in the PSRT had elicited responses consistent with the theory of scientific reasoning used to develop the instrument. A variety of procedures analysed the performance data according to the theoretically postulated impact five item features had on item difficulty estimates. Overall, the collective evidence provided some confirmation that the measured construct broadly conformed to the proposed theory. Most of the evidence for this claim came from the analyses on the location estimates of dichotomous items, which constituted most of the items of the revised test. Further analysis of findings that were not consistent with the theory suggested the unintended influence of flawed item design and unfamiliar item contexts. There was also evidence that pupils' mathematical and language competencies affected their performance on the items in unforeseen ways. The next and final chapter of this thesis will draw concluding arguments from the evidence gathered from this and earlier phases.

Chapter 9 Validity of the PSRT

9.1 Introduction

This research aimed to develop a valid and internally reliable test of scientific reasoning. First developed was a novel conceptualisation of scientific reasoning. A four-phase validation framework then guided the collection and analysis of validity evidence during the construction and validation of a new instrument with items designed using the postulated construct. The first phase (phase I) operationalised the theoretical construct into definitions and behavioural descriptors to steer the design of draft items and test assembly. The draft items then underwent an expert review (phase II) and a pilot study (phase). In the fourth phase (phase IV), Year 6 pupils attempted items in the finalised test – the PSRT.

This chapter will first review the validity evidence and validation results produced across the four study phases to draw broader inferences concerning the extent the PSRT validly measures scientific reasoning. A discussion of the impact and contributions of the study to theory and practice then follows these conclusions. Key to this discussion are the implications of the study's findings on current understanding of scientific reasoning and test design practices, as well as the possible relevance for curriculum and instruction. The thesis closes by considering the limitations of the findings and the possibilities for future research.

9.2 Construct Validity of the PSRT

As mentioned in the validation framework (Chapter 3, subsections 3.2.1 and 3.2.2), the definition of validity from the *Standards* as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" guided the construct validation of the PSRT (AERA, APA, & NCME, 2014, p. 11). Accordingly, the study drew on the relevant theory and collected evidence during the construction and evaluation of the PSRT. This section thematically describes how the theoretical and methodological frameworks, development, review and results of piloting, as well as the validation outcomes, jointly contributed to an understanding of the construct validity of the PSRT (this chapter). See

Table 9.1 for a summary of the organising themes across the chapters of the thesis and the evidence collected.

Table 9.1

Summary of organising themes in thesis and validity evidence collected across the phases

Organising Theme in Thesis	Chapter (Phase)	Validity Evidence Collected
Theoretical and methodological frameworks	2 to 3	Robustness and utility of the substantive and validation frameworks adopted in study
Development, review and pilot of the Draft PSRT	4 to 6 (I to III)	Content, response processes, internal structure, relations to other variables
Validation of the PSRT	7 to 8 (IV)	Content, response processes, internal structure, relations to other variables
Construct validity of the PSRT	9	Integration of overall evidence for meta-inferences (broader conclusions)

9.2.1 Evidence from Theoretical and Methodological Frameworks

The development of the items and the assembly of the draft test in phase I drew from the initial theoretical and practical groundwork delineated in the literature review and the methodology chapters.

Scientific Reasoning Construct: Post-logic and post-process developments in science education drew attention to three substantive and important forms of science knowledge: *content, procedural*, and *epistemic*. They also foregrounded three wide-ranging and related science practices found in a scientific investigation cycle. The first of these practices relates to hypothesis generation or theory development. The second and third emphasise the gathering of data from investigations and the evaluation of evidence respectively. Further supporting the notion of the scientific enterprise as a collection of inter-dependent practices were emerging insights from cognitive research in philosophy and psychology. Together,

these ideas led to a novel articulation of scientific reasoning as a coordination between three distinct forms of science knowledge and scientific evidence using three general science practices.[28]

Underpinning this conceptualisation are 12 reasoning types commonly expected of primary-aged pupils. These included basic mental operations such as conservation and seriation, where evidence of attainment served as crucial developmental signposts. More complex reasoning such as analogical, proportional, quantitative, and probabilistic reasoning follow later in cognitive development. Identification of the common reasoning types for primary science pupils facilitated the subsequent development of detailed behavioural descriptors of the competencies in the three science practices.

Validation Framework: Modern validity theories and guidelines from the *Standards* steered the collection and analysis of validity evidence for test construction and validation in this study. There are five *Standards*-endorsed validity sources, all of which originated from historical notions of validity. Following the plan drawn up in a validation framework, this thesis gathered evidence from three of these sources. A mixture of qualitative and quantitative methods was used to gather and analyse evidence from the selected sources. Deploying this methodology ensured stronger inferences about the validity of the PSRT as evidence from multiple investigations could be compared.

9.2.2 Evidence from Development, Review and Pilot of the PSRT

The development, review and subsequent pilot of the Draft PSRT formed the second thematic focus of the thesis. Validity evidence gathered under this theme came from the first three phases of the study.

Development of the Draft PSRT: An examination of the relevant nomological networks, which are principles, rules, guidelines, models, or concepts linked to broader theories,

[28] This study adjusted the terminology of the practices to reflect the targeted group of primary-aged pupils. The three practices are *giving scientific explanations*, *designing and evaluating investigations*, and *interpreting and analysing data and evidence*.

produced the definitions and behavioural descriptors underlying the scientific reasoning construct. Substantive theories from child developmental psychology, philosophy of science, science education and NOS literature, as well as local (Singapore) and international curricular documents formed the theoretical basis for articulating the types of knowledge and science practices relevant for primary-aged children learning science.

These sources also facilitated the explication of the competencies espoused under each science practice into sub-practices, which are expected progressions of increasingly developed levels of proficiency. Altogether, seven sub-practices appropriate for primary science pupils were identified. Briefly, these sub-practices delineate competencies in: (1) modelling; (2) recall and application of scientific knowledge; (3) planning of investigations; (4) manipulation of variables; (5) collection of precise and reliable data; (6) analysis of data, and (7) evidence evaluation to critique conclusions, hypotheses and arguments.

The construction of items to measure the construct followed four design parameters – test content, item format, cognitive demand, and item features. Cognitive learning theories shaped these parameters by specifying the expected cognitions, depth of reasoning, and performance. Measurement theories (e.g., Rasch measurement theory for psychometric analysis) and test design practices (e.g., matrix sampling principles) guided other test design and assembly considerations. These considerations included adopting an appropriate approach to characterise the items using the definitions and descriptors and articulating a sound rationale for sampling important learning outcomes across five booklets in the test framework and the TOS. Jointly, the decisions undertaken in the development of the PSRT and its items provided evidence for the suitability of the test development process.

Expert Review of the Draft PSRT: An expert panel consisting of 18 members strongly endorsed the relevance of the construct to science learners and the primary science curriculum. The members also agreed that the domain description of the knowledge types and science practices were detailed, useful, and suitable for primary pupils. However, the panel highlighted that some of the descriptors lacked clarity and required refinement.

Generally, the members in the panel agreed that most of the draft items measured the intended construct and met the specifications in the TOS. There was also broad agreement that items were soundly designed and aligned to the curriculum.

Following the panel's findings, items were identified for deletion, and recommendations were made to retain or amend the remaining items. Items that did not provide useful evidence (e.g., items with inappropriate item format) or could contribute possible construct-irrelevant variance (e.g., construct-irrelevant easiness/difficulty) were either deleted or amended. In total, 23 items were deleted, representing about a quarter of all items in the 87-item Draft PSRT. Of the remaining items, approximately half had recommendations for changes; the rest were retained without changes. Final decisions about item amendment and retention also depended on evidence gathered from the next phase (phase III) to ensure robust inferences.

Pilot of the Draft PSRT: In the pilot study, the Draft PSRT was subjected to a psychometric analysis using Rasch modelling on responses from a sample of Year 7 pupils. The trial of the administration procedures surfaced mixed evidence. On the one hand, there was evidence that the number of draft items was inadequate to draw out maximal inferences about pupils' performance during the designated test duration. However, more positive evidence from the test administration (e.g., briefing, booklet design) indicated that pupils experienced conducive testing conditions and gave their best efforts. An examination of the responses during scoring also suggested that most of the items drew out the required performance and reasoning.

The value of the index of reliability (i.e., the PSI) from the Rasch analysis indicated the reliable discrimination of the pilot sample, which consisted of three approximately equal-sized ability groups. This result also implied the fit statistics had adequate power to detect misfit. Only two of the 78 dichotomous draft items evidenced statistical irregularities: one showed model misfit and the other showed DIF. Responses to six of the 11 two-mark items

produced disordered thresholds. Limited by the pilot sample size, no inferences about the dimensionality and the construct measured in the Draft PSRT could be made.

Results from the psychometric analysis were used to finalise the outcomes to items that had received preliminary recommendations from the expert review. Except for a single item, all other items that obtained recommendations for retention evidenced acceptable item-fit statistics. Thus, this latter group of items was retained without changes for subsequent inclusion in the final test – the PSRT. The single item with misfit, along with the other items that received the second type of recommendation for amendments, were included in the PSRT only after effecting the required changes. The nature and extent of changes depended on both the amendments that the panel recommended and the items' psychometric results. For example, amendments took place to items that showed acceptable model-fit according to the panel's feedback. However, items with detected statistical anomalies had their mark schemes amended, on top of other changes from the panel's feedback.

9.2.3 Evidence from the Validation of the PSRT

Validation of the 100-item PSRT in the main study (phase IV) was the third and main thematic focus of the thesis. The first aim of this phase examined the claim that the PSRT measured a unidimensional construct. Having robust evidence to support the claim meant that pupils' scores could be interpreted confidently with respect to the measured construct. The second aim investigated whether the construct measured was the postulated theory of scientific reasoning. The gathering of validity evidence, as reported in Chapter 7, took place over two validation stages. Stage 1 was a review of new items by a two-member expert panel and stage 2 administered and analysed the performance data.

Two-Member Expert Review: In stage 1, the 2-member panel conducted an editorial, item-writing, and content review of 34 newly constructed test items. These items were intended as possible replacement for items deleted from the Draft PSRT in phase II because of possible threats to construct validity. The reviewers agreed that the new items satisfied the

TOS requirements and measured the intended knowledge types and science practices. However, they also found that one-third of these items had construction flaws such as “unfocused” stems that did not pose specific questions or “missing” information in stems that are required for correct responses. Revisions were made to these items based on the reviewers’ feedback and suggestions.

Other subsequent activities in this stage finalised and assembled the items into the PSRT. These activities included organising or revising items found suitable from earlier review and pilot phases, and drawing up a TOS and a table of revision history based on the assembled items to reduce threats from construct underrepresentation and construct-irrelevant variance, such as testing out-of-curriculum materials (Messick, 1995b). The table of revision history also documented when validity threats (e.g., unclear phrasing) were detected and resolved (Haladyna & Downing, 2004).

Administration, Scoring and Analysis of Responses to the PSRT: Evidence gathered from the administration of the PSRT showed that pupils had adequate information about the study and understood the testing purposes and score usage. These assurance measures resulted in pupils’ full participation and cooperation. Pupils’ performance took place under optimal and uniform testing conditions.

The initial Rasch analysis of the scores from the PSRT produced a high PSI value, which meant that fit statistics could detect item misfit reliably. The analysis also showed that 60% of the items in the test had acceptable fit statistics. After subjecting the items through a methodical and iterative cycle of statistical examination and remediation, the resulting 86-item Revised PSRT had fewer types and occurrences of model-fit irregularities. These irregularities were limited to two items that demonstrated unacceptable levels of fit residual, and five item-pairs with high residual correlations. There were no items with disordered thresholds and DIF. On average, the test was a little easy for the pupils. Also, there were minor and varying levels of multidimensionality across booklets. However, a consideration of all the evidence suggested there was reasonable measurement invariance and a good

fit of the data to the model. In turn, there was support for the claim that most of the items in the PSRT measured a unidimensional construct.

Detailed scrutiny of the correct responses provided substantial evidence that the elicited reasoning in most of the items corresponded to the intended knowledge types and competencies of the three science practices. Pupils' partial explanations and errors, which surfaced misconceptions and learning issues, also provided valuable clues. These shortcomings were found to be commonly associated with primary science pupils. For example, in some items, the manner pupils ignored, reinterpreted or rejected anomalous data predictably followed research accounts. Therefore, the evidence from an examination of the written responses supported preliminary inferences that there was reasonable alignment between the actual and intended performance to be measured.

Further analysis that compared items measuring the same competencies indicated their relative order of item difficulties generally accorded with theoretical expectations. Although this analysis also surfaced items that did not function as expected, the evidence from fit statistics and substantive theories suggested these issues arose because of item design flaws and person-related factors. These factors mediated pupils' reasoning in unanticipated ways and included literacy and mathematical competencies, as well as proficiency in transferring understanding across unfamiliar item contexts.

Summing up, the collective empirical results provide general support for inferences that the Revised PSRT measures a unidimensional construct. There is also credible evidence to suggest that the construct measured by the test broadly conforms to the scientific reasoning model proposed in this thesis. Reinforcing this claim are current perspectives from education research in science, mathematics, and language, which account for items that did not functioned as expected. The next subsection critically reviews the evidence across the themes to draw concluding arguments about the construct validity of the PSRT.

9.2.4 Construct Validity of the PSRT

Three validity sources – test content, internal structure, and response processes – formed the cornerstones of the validation framework in this study. As reviewed in the preceding subsections, various methods gathered the evidence from each source. See Figure 9.1 for a summary. This subsection evaluates the relevant key evidence to answer the overarching research question posed at the beginning of the thesis, "to what extent does the PSRT validly measure scientific reasoning?"

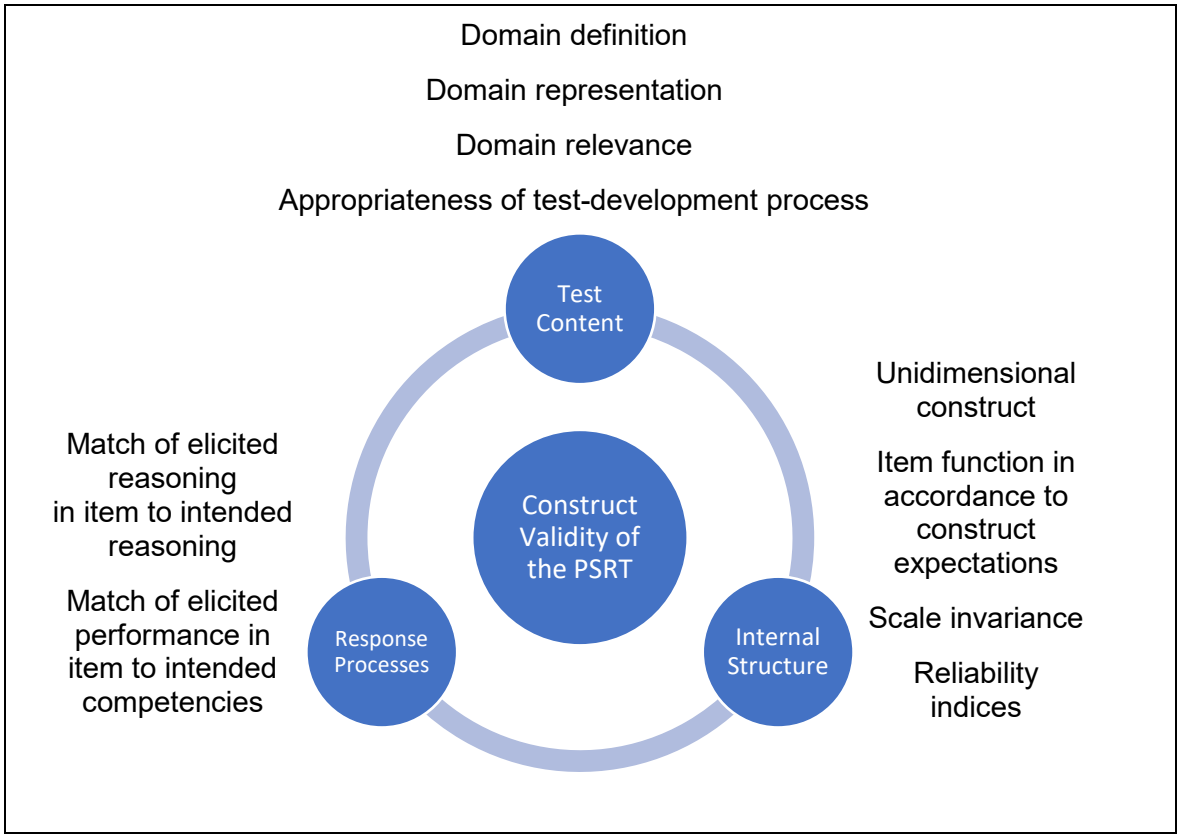


Figure 9.1. Key evidence gathered under each of the three validity sources in current study.

Key Test Content Evidence: Using Sireci's (1998a, 1998b) conceptualisation of content validity, evidence came from four test features. A novel construct of scientific reasoning and detailed descriptors of the knowledge types and science practices appropriate for primary-aged pupils established a comprehensive domain definition. Evidence of adequate and appropriate domain representation and domain relevance came from the item characterisation scheme of the drafted items and a clearly articulated TOS. An experienced panel of experts uncovered further evidence that endorsed the conceptual and content

validity of the Draft PSRT. In preparation for the main study, the pilot produced information that further strengthened the content validity of the items in the final test, improved the test administration and showed that Rasch modelling was a feasible psychometric approach.

Key Internal Structure and Response Processes Evidence: Evidence based on internal structure and response processes came largely from the main study. Psychometric analysis indicated good test reliability and scale invariance across subgroups. Although there were minor levels of multidimensionality and the test was a little easy, the overall results supported the claim that the PSRT measured a unidimensional construct. Evidence based on response processes compared whether test-takers' actual behaviours and mental engagement resembled those intended by the items. The results from a comparison of item difficulties among items measuring the same competencies and a qualitative examination of the responses suggested that explanatory concepts in the construct of scientific reasoning generally accounted for the PSRT scores.

Construct Validity of the PSRT: In conclusion, there are adequately strong inferences that the construct measured by the PSRT is unidimensional and conforms to the postulated theory of scientific reasoning. Other evidence attests to the robustness of the defined theory, construct descriptors, and the general soundness of the item design, test review, assembly, test administration, and data analytical procedures. Drawing on the strength and scope of the combined validity evidence from multiple methods across different phases, this thesis concludes with the argument that the PSRT achieves reasonably strong construct validity. In other words, the PSRT and its items are found to be valid measures of scientific reasoning in primary-aged pupils. The next few subsections discuss theoretical and practical implications of this conclusion, as well as highlight pertinent limitations of the study. Possible avenues of future research are also discussed.

9.3 Significance of the Study for Theory and Practice

This section presents the significance of the study for current conceptions of scientific reasoning and the construction and validation of items and tests designed to measure

scientific reasoning. It will also highlight the relevance of the study's construct and results to curriculum and instruction, and describe how the adopted data analytical approach contributes insights about item modification and item construction.

9.3.1 Defining the Construct of Scientific Reasoning

The findings of this study provide significant evidence to substantiate the theoretical basis of the scientific reasoning construct underlying the PSRT. As mentioned in Chapter 1 (subsection 1.3.2), the conceptual aspects of scientific reasoning are widely contested, leading to the proposal of disparate frameworks underpinning current classroom-based reasoning instruments, many of which have not been validated. These frameworks promoted various conceptions of scientific reasoning, such as evidence-based argumentation, problem-solving and critical thinking, or broad-based cognitive skills based on nature of science or metacognition literature (Erduran & Jiménez-Aleixandre, 2012; Fischer et al., 2014; Lederman et al., 2002; Schraw, 2011). Without clarity about the expected performance when reasoning with and about science, tests cannot be designed to validly measure this trait.

Moreover, as mentioned in the same subsection, the conceptions of scientific reasoning underlying many tests do not consider the role of grade-appropriate scientific knowledge. As shown in Table 1.1 of Chapter 1, most of the validated tests consist of "knowledge-lean" tasks that require only minimal prior scientific knowledge to respond. Thus, these tests conceive scientific reasoning to be general abilities that are transferable across distinct domains of knowledge ("domain general"). Critics argue that this perspective of reasoning neglects the mediating effects of context, instruction, and domain content. Unfortunately, as the number of validated tests using "knowledge-rich" items for primary-aged pupils are rare, there is limited understanding about how these pupils reason with basic science knowledge. In other words, there exists a gap in the literature as the scientific reasoning of novice science learners in primary schooling is not well studied.

In response to these theoretical gaps, this study formulated and validated a novel construct of scientific reasoning that explicitly defines the expected performance of primary children when reasoning with learnt scientific knowledge. Essentially, the construct of scientific reasoning formulated in this study articulates a workable approach to "translate" a complex cognitive proficiency into observable behaviours that have been contextualised for young learners of science. As will be explained in the following subsections, the theoretically informed and methodically defined construct in this study has practical significance for science assessment using paper-and-pencil tasks, instruction, and validation.

9.3.2 Measuring Scientific Reasoning

The validity of an assessment increases when the intended cognition and abilities match the maximal performance of pupils (Lane, 2011). To this end, well-constructed assessments are critical to enable robust score interpretations about the construct of interest (Pellegrino & Wilson, 2015). Although guidelines (see, e.g., Brookhart & Nitko, 2011) for constructing items and tests to measure complex thinking skills exist, piloted and validated item exemplars for assessing primary-level science pupils are rarely available or studied in the research literature. The theoretical basis of these test development guidelines have also been questioned. Sijtsma (2006, p. 453), for example, claimed that much of the relevant theory underpinning item and test construction, "is still in its puberty, infancy, or even at the foetal stage". A decade later, Baird et al. (2017) argued that the current state of test development and research has not contributed significantly to an understanding of how children learn, and vice versa. This position is shared by many in the assessment and educational communities (see, e.g., Borsboom, 2006; Pellegrino, 2017).

This study contributes to the literature with respect to many of the gaps outlined in the preceding discussion. As described in Chapter 4 (section 4.4), the design parameters that guided the construction and assembly of items in the PSRT were heavily informed by cognitive learning theories. In turn, using modern test theory procedures, the paper-and-pencil PSRT was found to be a valid measure of scientific reasoning. This result means

there is support for the interpretation and usage of the PSRT scores to provide insights about pupils' understanding. In effect, pupils' progress or learning gaps and the effectiveness of the instruction can be reliably monitored from their performance on the PSRT. Therefore, I argue that this study demonstrates how learning and test theories can be drawn together to support learning and instruction.

Furthermore, the study's extensive and theoretically-derived documentation of the methods and processes to characterise items, classify cognitive demand, and control item difficulty may serve as useful guidelines and exemplars for the construction of science items in classroom tests. Essentially, the knowledge-rich, paper-based PSRT could function as a "test template" to guide the drawing up of new classroom tests that focus on assessing other specific areas of understanding (e.g., science topics) or performance (e.g., competencies).

Disseminating these assessment tools and strategies would thus help the teaching community, who have consistently found the design of valid science items for assessing higher-order or complex reasoning a challenge (Collins, Reiss, & Stobart, 2009; Harrison & Howard, 2010; Kim et al., 2013). Guidelines for test modification to meet instructional needs are not evident in the documentation of almost all current tests of scientific reasoning, such as those shown in Table 1.1 of Chapter 1. Most of these tests are designed as "stand-alone" assessments, which means that the items in them cannot be changed or modified to draw and compare inferences across the performance of different groups of pupils.

9.3.3 Validating the PSRT

As compared to large-scale assessments, little attention has been paid to the validation of classroom assessments (Pellegrino et al., 2016). Opitz et al. (2017) reported discrepancies and aberrances in the validity studies of many current classroom tests of scientific reasoning, as described in subsection 1.3.2 of Chapter 1. Elsewhere, validation studies in some research either have no theoretical or validity perspectives (as reported by Zumbo & Chan, 2014), or are conceptually too complex for standard classroom implementation (see some examples in Pellegrino & Wilson, 2015).

In contrast, this study illustrates how a mixed method methodology guided by modern validity theory may be used to validate a classroom-based scientific reasoning test (see Chapter 3, section 3.2). To recap, the validation entailed the systematic collection and strategic integration of evidence from three selected sources to make a supported and defensible validity argument. The study indicated that critical elements in this validity argument are clear and appropriate domain definition, meticulous test construction, and strong evidence that response patterns conform to the construct of scientific reasoning. Significantly, factors that prevent the fulfilment of these requirements in the validity argument became threats in the validation of the PSRT. Some of the factors uncovered in this study are construct-irrelevant issues such as item construction flaws and multidimensionality in items.

Finally, the study applied some novel approaches to gather and manage evidence (Pellegrino et al., 2016). For instance, this study adopted Sireci's more encompassing and helpful definition to maximise content validity in a classroom-based assessment of scientific reasoning. Similarly, this study drew on cross-over analyses to integrate quantitative and qualitative evidence for stronger inferences; these emergent techniques had yet to be widely applied in mixed-method research (Onwuegbuzie & Combs, 2010). Another notable approach, as detailed in Chapters 5 and 6 (subsections 5.7.3 and 6.8.1), was to consolidate evidence from different methods (i.e., expert review and psychometric analysis in pilot study) in early phases of the research to strengthen claims for a later phase (i.e., that items finalised for the PSRT have strong content validity).

9.3.4 Implications for Curriculum and Instruction

The study has two major curricular and instructional implications for primary science education. These implications are drawn from the results of the study, evidence from the 18-member expert panel, substantive theories and recent science education developments. As mentioned, the theoretical basis underlying many classroom tests of scientific reasoning had not been validated

Use of the Construct Definitions and Descriptors: According to the members of the expert panel, the definitions and descriptors were relevant, comprehensive, and appropriate for learning science in Singapore and England. The experts were also in broad agreement that the proposed progression of the sub-practices was age-appropriate, logical and reasonable. Therefore, the construct definitions and descriptors of the competencies underlying the science practices could potentially be refined for use as a resource to support or complement teaching and learning. For example, teachers can use the detailed definitions and descriptors (see Appendix 5) to customise learning.

Also, arising from concerns expressed by some panel members that pupils received uneven instruction across Singapore schools because of variability in teachers' experience and knowledge, wide dissemination of the definitions and descriptors could address some deep-seated and wide-spread misunderstanding and beliefs among pupils about important aspects of science. For example, findings from NOS literature have consistently reported (see, e.g., Sandoval, Sodian, Koerber, & Wong, 2014) that young pupils are at risk of developing erroneous epistemic notions (e.g., science is dependable and therefore data has no error). Instructional exposure to the competencies underlying the science practices could prevent or correct mistaken notions (e.g., errors could occur during data collection).

Gender Effects, Language and Mathematical Competencies on Science Learning: Although both genders were found to have comparable overall performance on the Revised PSRT, there was some evidence that they responded differently on certain items. For example, as reported in Chapter 7, the item shown in Figure 7.5 demonstrated uniform gender DIF; that is, boys performed better than the girls of the same overall ability estimate across the scale. This item required pupils to extrapolate from given data to predict an answer. Interestingly, many girls responded incorrectly with "28.4s", presumably by doubling the value "14.2" from the table. Together, the appearance of DIF and the response patterns of girls in this item suggests that boys demonstrated more proficiency at predicting and extrapolating a value beyond the measured range.

Findings from other studies have provided further insights about gendered learning styles. For example, Elwood (2005) found that the preferred response style of boys, which Murphy (2000, p. 171) described as “factual” and “clinical”, to be more valued in subjects such as physics. In contrast, studies have described girls to be more inclined towards producing “extended” and “reflective” writing (Murphy, 2000). Although the evidence about gender effects on science achievement from this study was not strong, it nevertheless implies that educators and item writers should look for such possible effects in teaching (e.g., written and oral communication) and testing (e.g., physical science and life science topics).

Similarly, indications that literacy and mathematical competencies impacted pupils’ performance on science items in this study are consistent with broad findings from research (Lenton & Stevens, 2000; Yore et al., 2004). For example, items in this study that assessed advanced scientific argumentation such as justification and critique of claims, were found to be more difficult than those testing simpler proficiencies (e.g., state a claim). Evidence from items testing quantitative reasoning demonstrated that some pupils responded wrongly not because of their weak scientific understanding. Rather these pupils had applied the mathematical concepts and heuristics (e.g., area and estimation) incorrectly.

9.3.5 Analysing Performance Data from the PSRT

Rasch modelling is widely used in many fields to analyse the validity of test constructs (Smith, 2001). However, there are relatively fewer reports about the use of this psychometric approach to diagnose flaws in items with respect to the tested substantive theories (Andrich, 2002; van Wyke & Andrich, 2006). This study adds to the literature by demonstrating how an iterative analysis using Rasch procedures, as outlined in Chapter 7 (subsection 7.5.1), could identify the substantive issues underlying some model-fit issues in science items, especially those testing multi-step cognitive processes such as the polytomous items in the PSRT.

For instance, in the only three-mark item from the PSRT shown in Figure 9.2, the observed disordered thresholds suggested that the single mark scoring category was not functioning as intended.

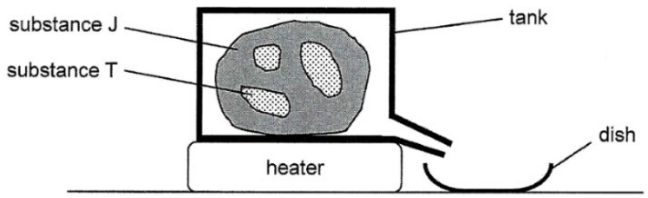
BXQ2(a)(i)	***Scoring Guide for BXQ2(a)(i)***
John placed a solid made of substances J and T in the set-up as shown below. Substance J had a lower melting point than substance T.  (a) He wanted to obtain a solid made of substance J only. (i) Explain how he could use the set-up to do so. [3 marks]	Instruction: Award one mark for each appropriate reference below. 1. Melting - Heater (heats up to) melt J (completely); 2. Separation - The liquid J flows out into dish; 3 Freezing - Allow J to cool/freeze/allow J to lose heat (so that J becomes solid again);

Figure 9.2. An example to show how results from Rasch analysis informed adjustments to the scoring guide of an item for a fairer assessment.

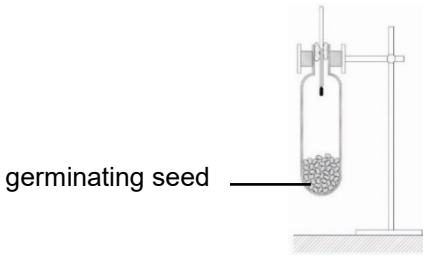
Consequently, pupils had higher probabilities of achieving either of the adjacent scoring categories (i.e., no marks or two marks) to the middle scoring category (i.e., one mark). Checks on pupils' scripts showed that the missing mark point referred to the idea of separation due to a differential in melting points between substances J and T. Further empirical investigation and remediation suggested that a two-mark version assessing *only* the scientific concepts of melting and freezing was more viable. This result indicated that the item context triggered mental schemas associated with physical changes of state. A more valid and fair assessment of pupils would thus entail revising the scoring guide to match the item's elicited competencies.

Besides understanding how the scoring could be correlated with elicited cognitions, the diagnostic process on the response patterns also presented ideas for the design of novel items by identifying the sources of pupils' difficulties and misconceptions. For instance, with the findings of two items with multiple statistical flaws shown in Figure 8.23 (Chapter 8), analyses highlighted the importance of crediting only responses that noted the contact

between the thermometer and the seeds. Taking cue from this finding, the new item shown in Figure 9.3 enhanced the saliency of the diagrams to focus attention on the different medium in contact with the two thermometers.

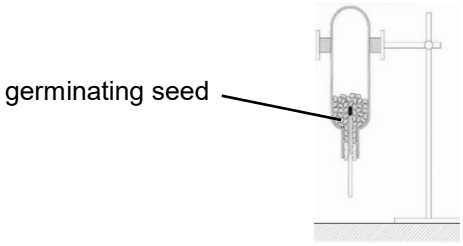
Two students, John and Lucy, experimented to know if germinating seeds give out heat.
John used the set-up shown below.

Set-up by John



Lucy disagreed with John's set-up. The diagram below shows her set up.

Set-up by Lucy



Give a reason why the set-up by Lucy was better than the set-up by John.

Figure 9.3. A novel item designed based on statistical findings from flawed items. This item assesses pupils' proficiency in justifying a claim.

In summary, this study contributes to the theoretical literature about the nature, learning and assessment of scientific reasoning in primary-aged school children. The basis of these theoretical contributions rests on the conceptualisation of a novel construct of scientific reasoning, and the subsequent development of age-appropriate construct definitions and descriptions. Other implications from the study include practical insights about appropriate test design, feasibility of validation techniques, and opportunities for enhancing instruction. Finally, the study demonstrates how Rasch modelling techniques can be extended to diagnose substantive issues and provide ideas for new items.

9.4 Limitations of Study

Although the study achieves its two research aims, and determines that the PSRT displays good construct validity, there are also limitations to this study. This section discusses two broad groups of shortcomings – theoretical and methodological.

9.4.1 Theoretical Limitations

Scientific Reasoning Construct in Study: As recounted in Chapter 4 (section 4.3), developments in science education and emergent sources in the literature guided the development of the postulated knowledge types and science practices in the construct. Other than this research and another small-scale study (i.e., Kind, 2013), the implications of these two aspects of the construct for science assessment have never been investigated systematically. In the other study, the author developed a test of 32 items for 139 secondary pupils based on two of the three science practices.[29] His overall results were mixed and indicated the presence of unexplained multidimensionality. Moreover, the published results did not explicate fully the specific competencies underlying the measured practices. Therefore, there was a lack of empirical support for the conceptualised construct in the current study.

There is also limited empirical evidence in the literature to support the proposed progression of proficiency levels outlined in the descriptors of the seven sub-practices in this study. As described in section 4.3, the development of these level descriptors followed a blend of four approaches: theory-based, empirically-guided, policy-driven and expert-devised (Baird et al., 2017). Substantive sources that were consulted were drawn largely from historical and current curricular initiatives and reforms. Unfortunately, the systematic validation of similar progressions from these sources had never been formally conducted (e.g. National Strategies, a now-defunct large-scale professional development programme in the UK), or

[29] The two science practices, as termed in the Kind's (2013) study, were "evidence evaluation" and "experimenting". These practices were analogous to the current study's practices *of interpreting and analysing data and evidence* and *designing and evaluating investigation*.

had yet to be implemented as the initiatives (e.g., NGSS) are new (Department for Education, 2011; Pruitt, 2014). Overall, the evidence (e.g., panel's feedback) suggested that the level descriptors of proficiency required further refinement.

Item Format and Cognitive Demand: As reported in Chapter 8, the observed relationships between item difficulty in the Revised PSRT, and the two test features of item format and cognitive demand, broadly conformed to expectations laid out in test theory. However, SROS items did not turn out to be the easiest among the three item formats. Inspection of these items suggested that the number of options could have unduly increased difficulties in some items (four items had four or more options). Other evidence indicated that some of the SROS items had options that tested well-entrenched misconceptions (see Figure 8.9 in Chapter 8 for an example).

Second, the observed order of item difficulty between items of medium and high cognitive demand indicated that the categorisation scheme (drawn up in Table 4.17, Chapter 4) had not been successfully used to distinguish items of the two demand levels. As addressed in Chapter 4 (subsection 4.4.3), this result was unsurprising as 60 years of research in taxonomies of cognitive demand (e.g., Revised Bloom's Taxonomy) has largely shown them to be ineffectual in categorising and identifying the sources of cognitive difficulties (Gorin & Embretson, 2013; Haladyna & Rodriguez, 2013). Given the evidence from the results, it was possible that only two levels – low and high – were pragmatically viable to categorise the cognitive demand of items in the study's context of primary science testing. This implied that the study's cognitive demand categorisation scheme needs to be revised.

In summary, the lack of empirical evidence in the science education and testing literature has limited the claims that can be made about the construct definitions and descriptors devised in this study. For example, the progression of competencies outlined in the level descriptors of the sub-practices could not be validated in the current study. Disconfirmatory evidence also indicated that the actual order of item difficulty estimates for some items did not conform to the theorised impact of item format and cognitive demand.

9.4.2 Methodological Limitations

Sources of methodological limitations relate to the functioning of the PSRT and the lack of validity evidence from incomplete analysis of some of the data in this study.

Test Instrument and Items: The results suggest that the paper-and-pencil PSRT was overall slightly easy for the 400-odd primary school children in the main study. As there was convenient sampling of the six schools, it is uncertain if the pupil profile in the sample had an impact on the overall test difficulty. The slight "easiness" of the test could also be because many of the items measured the lower levels of the sub-practices. Unfortunately, evidence about the abilities of this sample was not available as the data had not been fully analysed (this point will be elaborated later in this discussion). Without further evidence, it remains unclear what accounted for the test's observed difficulty level.

Also, some items across the five booklets displayed varying and minor levels of statistical multidimensionality. Other than substantive reasons (items measuring non-intended competencies), multidimensionality could also arise from factors such as the use of multiple item formats, and the presence of items testing similar competencies within a test (Andrich, 2006b; Linacre, 1998a). Limited by the scope of the research, the underlying reasons for the presence of multidimensionality were not investigated using post-hoc procedures. The reasons for the observed multidimensionality remain uncertain.

Validation Methods: As mentioned in Chapter 3, there was incomplete analysis of data collected in the main study from cognitive testing and psychological instruments. The validation framework therefore excluded the data from these methods. See Appendix 13 for the original validation plan using these methods. Complete analyses of data from these methods would have produced evidence based on response processes and relations to other variables. Jointly, the evidence could reinforce or lessen the strength of inferences from the other gathered evidence about the construct validity of the PSRT. The next section discusses how possible future research could address this limitation.

9.5 Future Research

Based on the preceding discussion, there are two immediate areas suggested for post-thesis research. The first area is to complete the analysis of discriminant evidence obtained from the administration of the Raven's Standard Progressive Matrices-Plus (SPM+) and the Mill Hill Vocabulary Scales (MHV), which are brief non-verbal and verbal measures of general intelligence (Raven, 2008). Both instruments have been subjected to extensive research for over 70 years and are regarded as quality tests (Gignac, 2015). Correlating the pupils' ability estimates from the performances in the PSRT and psychological tests would provide evidence for evaluating whether the PSRT's underlying construct is distinct from general reasoning (AERA, APA, & NCME, 2014).

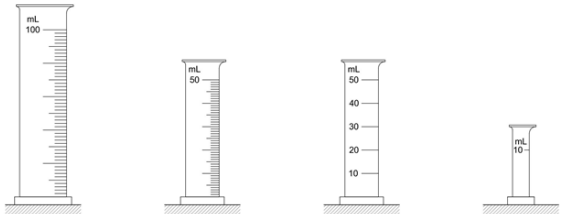
The second area, like the first, is to analyse data gathered from cognitive interviews, which are commonly deployed to find validity evidence based on response processes (Castillo-Díaz & Padilla, 2012). In research with children, cognitive interviewing focuses on investigating the degree of congruence between how pupils understand, think, and respond to test items and the intended cognitive processes (Ruiz-Primo, Shavelson, Li, & Schultz, 2001; Thelk & Hoole, 2006). Evidence from this method could evaluate the degree of match between the actual and intended cognitions. Overall, the evidence from both methods could provide information about the sample's abilities that might explain the easiness of the test.

A third possible area for future research is to examine the progression of the postulated levels in the sub-practices using some of the analytical approaches in this study. For instance, Rasch modelling procedures could check if the difficulties of items assigned to various levels within the same sub-practice and across sub-practices conform to theory. This research could also be furthered by conducting similar validation studies with different sub-populations of primary school pupils (other Singapore schools or English schools) to examine the construct validity of the PSRT.

9.6 Final Thoughts

How do primary school pupils reason scientifically? This question was considered at the start of the first chapter using the following item exemplars shown in Figure 9.4.

The diagram below shows four types of a piece of equipment.



(a) Identify the equipment shown above. _____

(b) Which type of equipment will measure 15 cm^3 of water most accurately?

Circle your answer below.

A / B / C / D

Figure 9.4. Question made up of two items adapted from a Singapore primary school paper.

The results of this study provide more definitive indications of how a primary-aged pupil reasons through these items scientifically. Drawing on the construct defined in this study, the pupil makes claims by weighing the evidence using his/her store of procedural knowledge and science practices. Specifically, the knowledge is about the relationship between the scales of the measuring cylinders and the precision of readings. The science practices involved are that of knowledge application and instrument selection.

By deconstructing how the pupil is reasoning scientifically into observable and measurable behaviours, the successful validation of the PSRT in this thesis offers insights into how targeted instruction could take place and valid assessments could be designed. The results of the PSRT could be used confidently to diagnose pupils' reasoning and inform instruction. The test development guidelines could also help teachers design new tests and items for classroom use. In addition, the theoretical basis of the PSRT and the study's validation and psychometric approaches would interest researchers and test developers. However, the ultimate beneficiaries of this thesis are the young learners of science in schools.

The Construction and Validation of the Primary Scientific Reasoning Test



Volume II

Appendices and References

References

Abd-El-Khalick, F., BouJaoude, S., Duschl, R. A., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., . . . Tuan, H.-L. (2004). Inquiry in science education: International perspectives. *Science Education, 88*, 397-419.

Achieve. (2010). *Taking the lead in science education: Forging Next-Generation Science Standards*. Retrieved from https://www.achieve.org/files/InternationalScienceBenchmarkingReport.pdf

Adey, P., & Csapó, B. C. (2012). Developing and assessing scientific reasoning. In B. C. Csapó & G. Szabó (Eds.), *Framework for diagnosing assessment of science* (1st ed., pp. 17-53). Budapest: Nemzeti Tanknyvkiado.

Adey, P., Farkas, M. A., Németh, M. B., Csapó, B., Csíkos, C., Korom, E., . . . Wagner, E. (2012). *Framework for diagnostic assessment of science* (B. Csapó & G. Szabó Eds.). Budapest: Nemzeti Tankönyvkiadó.

Adey, P., & Harlen, W. (1986). A Piagetian analysis of process skill test items. *Journal of Research in Science Teaching, 23*(8), 707-726.

Ahmed, A., & Pollitt, A. (2001). *Improving the validity of contextualised questions*. Paper presented at the BERA Conference, Leeds.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*, 955-959.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115-129.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1-18.

Allen, E. I., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress, 40*(7), 64-65.

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. New York, NY: John Wiley and Sons, Inc.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association (AERA) and National Council on Measurements Used in Education (NCMUE). (1955). *Technical recommendations for achievement tests*. Washington, DC: National Education Association.

American Psychological Association (APA), American Educational Research Association (AERA), & National Council on Measurement in Education (NCME). (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association, Inc.

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development, 11*, 523-550.

Andrich, D. (1982). An index of person separation in latent trait theory the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives, 9*(1), 95-104.

Andrich, D. (1988). *Rasch models for measurement*. Thousand Oaks, CA: SAGE Publications, Inc.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.), *Mathematical and theoretical systems*. North Holland: Elsevier Science Publishers B. V.

Andrich, D. (2002). A framework relating outcomes based education and the taxonomy of educational objectives. *Studies in Educational Evaluation, 28*, 35-59.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*, 1-7.

Andrich, D. (2006a). *On the fractal dimension of a social measurement II*. Unpublished paper, Social Measurement Laboratory, Murdoch University. Perth, Australia.

Andrich, D. (2006b). *On the fractal dimension of a social measurement: I*. Unpublished Paper, Social Measurement Laboratory, Murdoch University. Perth, Australia.

Andrich, D. (2010a). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika, 75*(2), 292-308.

Andrich, D. (2010b). Understanding the response structure and process in the polytomous Rasch model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications*. New York, NY: Routledge.

Andrich, D., de Jong, J. H. A. L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langerheine (Eds.), *Applications of latent trait and latent class models in the social sciences*. New York, NY: Waxmann Münster.

Andrich, D., Sheridan, B., & Luo, G. (1997-2012a). *Cronbach's α and the Person Separation Index (PSI)*. Perth: RUMM Laboratory.

Andrich, D., Sheridan, B., & Luo, G. (1997-2012b). *RUMM2030: Rasch unidimensional models for measurement*. Perth: RUMM Laboratory.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21*(3), 1105-1106.

Baird, J.-A. (2014). Valuable learning. In J.-A. Baird, T. N. Hopfenbeck, P. Newton, G. Stobart, & A. T. Steen-Utheim (Eds.), *State of the field review: Assessment and learning. Report for the Norweigian Knowledge Centre for Education* (pp. 21-29). Oxford, UK: Oxford University Centre for Educational Assessment.

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice, 24*, 317-350.

Baird, J.-A., Caro, D. H., & Hopfenbeck, T. N. (2016). Student perceptions of predictability of examination requirements and relationship with outcomes in high-stakes tests in Ireland. *Irish Educational Studies, 35*(4), 361-379.

Baird, J.-A., Hopfenbeck, T. N., Newton, P. E., Stobart, G., & Steen-Utheim, A. T. (2014). *Assessment and learning: State of the field review*. Oslo: Knowledge Centre for Education.

Baird, W. E., Perry, W. D., & Simon, M. (1989). *Correlates of student performance in the Science Olympiad: The Test of Integrated Process Skills and other variables*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., . . . Wu, N. (2009). Learning and scientific reasoning. *Science, 323*, 586-587.

Barmby, P., & Defty, N. (2006). Secondary school pupils' perceptions of physics. *Research in Science Technological Education, 24*(2), 199-215.

Beilin, H. (1992). Piaget's enduring contribution to developmental psychology. *Developmental Psychology, 28*(2), 191-204.

Bennett, R. E. (2012). Consequences that cannot be avoided: A response to Paul Newton. *Measurement: Interdisciplinary Research & Perspective, 10*(1-2), 30-32.

Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education, 94*, 765-793.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research & Perspective, 9*, 71-123.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals; Handbook I: Cognitive domain*. New York, NY: David McKay Company, Inc.

Borsboom, D. (2006). The attack of the psychometricans. *Psychometrika, 71*, 425-440.

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology, 44*(5), 1478-1490.

Braine, M., & Rumain, B. (1983). Logical reasoning. In J. Flavell, E. Markman, & P. Mussen (Eds.), *Handbook of child psychology, Volume 3: Cognitive development* (pp. 263-340). New York, NY: Wiley and Sons.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Briggs, D. C. (2017). Learning theory and psychometrics: Room for growth. *Assessment in Education: Principles, Policy & Practice, 24*, 351-358.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Brookhart, S. M., & Nitko, A. J. (2011). Strategies for constructing assessments of higher order thinking skills. In G. Schraw & D. R. Robinson (Eds.), *Assessment of higher order thinking skills* (pp. 327-359). Charlotte, NC: Information Age Publishing, Inc.

Brown, N. J. S., Furtak, E. M., Timms, M., Nagashima, S. O., & Wilson, M. (2010). The evidence-based reasoning framework: Assessing scientific reasoning. *Educational Assessment, 15*, 123-141.

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment, 15*, 142-174.

Bruner, J. S. (1960). *The process of education*. New York, NY: Vintage.

Bruner, J. S. (1964). The course of cognitive growth. *American Psychologist, 19*, 1-15.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking* (6th ed.). New York, NY: John Wiley & Sons, Inc.

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97-113.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research, 1*, 8-22.

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood: Findings from a 20 year longitudinal study* (pp. 173-197). Mahwah, NJ: Lawrence Erlbaum.

Burmester, M. A. (1953). The construction and validation of a test to measure some of the inductive aspects of scientific thinking. *Scientific Thinking, 37*(2), 131-140.

Burns, J. C., Okey, J. R., & Wise, K. C. (1985). Development of an integrated process skill test: Tips II. *Journal of Research in Science Teaching, 22*(2), 169-177.

Campanario, J. M. (2002). The parallelism between scientists' and students' resistance to new scientific ideas. *International Journal of Science Education, 24*(10), 1095-1110.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15*, 546-553.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Caravita, S., & Halldén, O. (1994). Re-framing the problem of conceptual change. *Learning and Instruction, 4*, 89-111.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences, 3*(3), 106-116.

Case, R., & Bereiter, C. (1984). From behaviourism to cognitive behaviourism to cognitive development: Steps in the evolution of instructional design. *Instructional Science, 13*, 141-158.

Castillo-Díaz, M., & Padilla, J.-L. (2012). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research, 114*, 963-975.

Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instuments. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 9-24). Switzerland: Springer.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variable strategy. *Child Development, 70*(5), 1098-1120.

Chin, D. B., Dohmen, I. M., & Schwartz, D. L. (2013). Young children can learn scientific reasoning with teachable agents. *IEEE Transactions on Learning Technologies, 6*(3), 248-257.

Chin, T.-Y., & Poon, C.-L. (2014). Design and implementation of the national primary science curriculum: A partnership approach in Singapore. In A.-L. Tan, C.-L. Poon, & S. L. S. Lim (Eds.), *Inquiry into the Singapore science classroom: Research and practices* (pp. 27-46). Singapore: Springer.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1-49.

Christensen, K. B., Makransky, G., & Horton, M. (2016). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178-194.

Clark, D. B., & Sampson, V. (2008). Assessing dialogic argumentation in online environments to relate structure, grounds, and conceptual quality. *Journal of Research in Science Teaching, 45*, 293-321.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological review, 76*(4), 387-404.

Cleary, D. (2008). U.K. education reform: Too much of a good thing? *Science, 312*, 1428-1429.

Clement, J. J. (1998). Expert novice similarities and instruction using analogies. *International Journal of Science Education, 20*(10), 1271-1286.

Cloonan, C. A., & Hutchinson, J. S. (2011). A chemistry concept reasoning test. *Chemistry Education Research and Practice, 12*, 205-209.

Cohen, H. F. (2010). *How modern science came into the world: Four civilizations, one 17th-century breakthrough*. Amsterdam: Amsterdam University Press.

Collins, K. M. T., Onwuegbuzie, A. J., & Sutton, I. L. (2006). A model incorporating the rationale and purpose for conducting mixed-methods research in special education and beyond. *Learning Disabilities, 4*(1), 67-100.

Collins, S., Reiss, M. J., & Stobart, G. (2009). The effects of national testing in science at key stage 2 in England and Wales. *Education in Science, February*, 14-15.

Crombie, A. C. (1994). *Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts. Volume I*. London, UK: Gerald Duckworth & Co. Ltd.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147-171). Champaign, IL: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Csapó, B. (1997). The development of inductive reasoning: Cross-sectional assessments in an educational context. *International Journal of Behavioral Development, 20*(4), 609-626.

Curriculum Council. (2005). *Curriculum framework progress maps: Science*. Osborne Park, Western Australia: Curriculum Council of Western Australia.

Custer, M., Sharairi, S., & Swift, D. (2012). *A comparison of scoring options for omitted and not-reached items through the recovery of IRT parameters when utilising the Rasch Model and Joint Maximum Likelihood Estimation*. Paper presented at the National Council on Measurement in Education, Vancouver, British Columbia.

Darian, S. (2000). The role of figurative language in introductory science texts. *International Journal of Applied Linguistics, 10*, 163-186.

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. *Yearbook of the National Society for the Study of Education, 104*, 289-319.

Davies, A. (1985). Follow my leader: Is that what language tests do? In Y. P. Lee, C. Y. Y. Fok, R. Lord, & G. Low (Eds.), (pp. 1-12). Oxford: Pergamon Press.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194-197.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching, 37*(6), 582-601.

DeMars, C. (2002). *Missing data and IRT item parameter estimation*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. https://files.eric.ed.gov/fulltext/ED476175.pdf

Denzin, N. K. (1970). *The research act in sociology: The theoretical introduction in sociological methods*. London: Butterworths.

Department for Education. (2011). *The National Strategies 1997-2011: A brief summary of the impact and effectiveness of the National Strategies*. Retrieved from

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/175408/DFE-00032-2011.pdf

Department for Education. (2013). *The national curriculum in England: Key stages 1 and 2 framework document*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/425601/PRIMARY_national_curriculum.pdf

Department for Education. (2014). *The national curriculum in England: Framework document*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/381344/Master_final_national_curriculum_28_Nov.pdf

Department for Education. (2015). National curriculum in England: Science programmes of study. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study#contents

Department for Education and Skills, & Qualifications and Curriculum Authority. (2004). *Science: The national curriculum for England*. Retrieved from http://webarchive.nationalarchives.gov.uk/20130402163320/https://www.education.gov.uk/publications/eOrderingDownload/DfES-0303-2004.pdf

DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44*(11), S50-S59.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Diaz-bilello, E., Patelis, T., Marion, S., Hall, E., & Betebenner, D. (2014). Are the Standards for Educational and Psychological Testing relevant to state and local assessment programs? *Educational Measurement: Issues and Practice, 33*(4), 16-18.

Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. *Science Education, 64*(5), 601-608.

Downing, R. E. (1928). The elements and safeguards of scientific thinking. *The Scientific Monthly, 26*(3), 231-243.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*, 133-143.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*, 61-82.

Driver, R. (1983). *The pupil as scientist?* Milton Keynes: Open University Press.

Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. *Studies in Science Education, 5*(1), 61-84.

Driver, R., Learch, J., Millar, R., & Scott, P. (1996). *Young people's images of science*. Buckingham, UK: Open University Press.

Duit, R. (2007). Science education research internationally: Conceptions, research methods, domains of research. *Eurasia Journal of Mathematics, Science and Technology Education, 3*, 3-15.

Dunbar, K. N., & Klahr, D. (2012). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 699-718). Oxford: Oxford University Press.

Duschl, R. A. (2000). Making the nature of science explicit. In R. Millar, J. Leach, J. Osborne, & R. Driver (Eds.), *Improving science education: The contributions of research*. Buckingham, UK: Open University Press.

Duschl, R. A. (2007). Quality argumentation and epistemic criteria. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based Research* (pp. 159-175). Dordrecht: Springer.

Duschl, R. A. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32*, 268-291.

Duschl, R. A., & Grandy, R. (2013). Two views about explictly teaching nature of science. *Science and Education, 22*, 2109-2139.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*, 39-72.

Earle, S. (2014). Assessment in England: Some FAQs (and answers!). *Primary Science, 133*, 27-28.

Elwood, J. (2005). Gender and achievement: What have exams got to do with it? *Oxford Review of Education, 31*(3), 373-393.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Ennis, R. H. (1982). Children's ability to handle Piaget's propositional logic: A conceptual critique. In S. Mogdil & C. Mogdil (Eds.), *Jean Piager: Consensus and controversy* (pp. 101-130). London: Holt, Rinehart and Winston.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*(2), 24-35.

Erduran, S., & Dagher, Z. R. (2014a). Methods and methodological rules. In S. Erduran & Z. R. Dagher (Eds.), *Reconceptualizing the Nature of Science for science education: Scientific knowledge, practices and other family categories* (pp. 91-112). Netherlands: Springer.

Erduran, S., & Dagher, Z. R. (2014b). Reconceptualizing Nature of Science for science education. In S. Erduran & Z. R. Dagher (Eds.), *Reconceptualizing the Nature of Science for science education: Scientific knowledge, practices and other family categories* (pp. 1-17). Netherlands: Springer.

Erduran, S., & Dagher, Z. R. (2014c). Scientific practices. In S. Erduran & Z. R. Dagher (Eds.), *Reconceptualizing the Nature of Science for science*

education: Scientific knowledge, practices and other family categories (pp. 67-89). Netherlands: Springer.

Erduran, S., & Duschl, R. A. (2004). Interdisciplinary characterizations of models and the nature of chemical knowledge in the classroom. *Studies in Science Education, 40*(1), 105-138.

Erduran, S., & Jiménez-Aleixandre, M. P. (2012). Argumentation in Science education research: Perpectives from Europe. In K. Tobin, C. Milne, & D. Jorde (Eds.), *Science education research and practice in Europe: Retrospective and prospective*. Rotterdam: Sense Publishers.

Erduran, S., Ozdem, Y., & Park, J.-Y. (2015). Research trends on argumentation in science education: A journal content analysis from 1998–2014. *International Journal of STEM Education, 2*, 5.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Science Education, 88*, 915-933.

European Commission. (2013). *European encyclopedia on national education systems: United Kingdom (England)*. Retrieved from https://eacea.ec.europa.eu/national-policies/eurydice/content/united-kingdom-england_en

Feist, G. J. (2006). Origins of the scientific thinking. In G. J. Feist (Ed.), *The psychology of science and the origins of the scientific mind* (pp. 186-218). New Haven, CT: Yale University Press.

Feyzíoglu, B., Demírdag, B., Akyildiz, M., & Altun, E. (2012). Developing a science process skills test for secondary students: Validity and reliability study. *Educational Sciences: Theory & Practice, 12*(3), 1899-1906.

Fields, D. C. (2000). The impact of Gagné's theories on practice. In R. C. Richey (Ed.), *The legacy of Robert M. Gagné* (pp. 147-181). Syracuse, NY: ERIC Clearinghouse on Information and Technology.

Finley, F. N. (1983). Science processes. *Journal of Research in Science Teaching, 20*(1), 46-54.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., . . . Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 4*, 28-45.

Fisher, W. P. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions, 6*(3), 238.

Flavell, J. H. (1963). *The developmental psychology of Jean Piaget*. New York, NY: Van Nostrand Reinhold Company.

Flavell, J. H., Miller, P. H., & Miller, S. A. (1993). *Cognitive development* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive development* (4th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Forster, M., & Masters, G. (2010). Progression and assessment: Developmental assessment. [electronic resource]. In P. Peterson, E. Baker, & B. McGaw

(Eds.), *International encyclopedia of education* (3rd ed.). Oxford, UK: Academic Press.

Fraser, B. J. (1980). Development and validation of a test of enquiry skills. *Journal of Research in Science Teaching, 17*(2), 7-16.

Gagné, R. M. (1963). The learning requirements for enquiry. *Journal of Research in Science Teaching, 1*, 144-153.

Gagné, R. M. (1968). Learning hierarchies. *Educational Psychologist, 6*(1), 1-9.

Garcia-Mila, M., & Anderson, C. (2008). Cognitive foundations of learning argumentation. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 29-45). Dordrecht: Springer.

Gelman, S. A., & Markman, E. A. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.

Germann, P. J. (1989). The processes of biological investigations test. *Journal of Research in Science Teaching, 26*(7), 609-625.

Gibbs, B. C. (2015). Reconfiguring Bruner: Compressing the spiral curriculum. *Phi Delta Kappan, 95*(7), 41-44.

Giere, R., Bickle, J., & Maudlin, R. F. (2006). *Understanding scientific reasoning* (5th ed.). Belmont, CA: Thomas Wadsworth.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for *g* factor theory and the brief measurement of *g*. *Intelligence, 52*, 71-79.

Gilbert, J. K. (2004). Models and modelling: Routes to more authentic science education. *International Journal of Science and Mathematics Education, 2*, 115-130.

Girotto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition, 106*(1), 325-344.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521-563.

Gold, R. (1987). *The description of cognitive development: Three Piagetian themes*. New York, NY: Oxford University Press.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments, 3*, 125-156.

Gopinathan, S. (1991). Education. In E. C. T. Chew & E. Lee (Eds.), *A history of Singapore*. Oxford, UK: Oxford University Press.

Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 173-201). New York, NY: Cambridge University Press.

Gorin, J. S., & Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*(5), 394-411.

Gorin, J. S., & Embretson, S. E. (2013). Using cognitive psychology to generate items and predict item characteristics. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice*. New York, NY: Routledge.

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE Life Sciences Education, 11*, 364-377.

Goswami, U. (2004). *Models of cognitive development*. Malden, MA: Blackwell Publishing Ltd.

Goswami, U. (2010). Inductive and deductive reasoning. In U. Goswami (Ed.), *Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 399-419). Hoboken, NJ: Wiley.

Gott, R., & Duggan, S. (1996). Practical work: Its role in the understanding of evidence in science. *International Journal of Science Education, 18*(7), 791-806.

Gott, R., & Duggan, S. (2003). *Understanding and using scientific evidence: How to critically evaluate data*. London, UK: SAGE Publications, Inc.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (2015). *Research into understanding scientific evidence*. Retrieved from https://community.dur.ac.uk/rosalyn.roberts/Evidence/CofEv_Gott%20et%20al.pdf

Gott, R., & Mashiter, J. (1991). Practical work in science: A task-based approach. In B. Woolnough (Ed.), *Practical science* (pp. 179-191). Milton Keyes: Open University Press.

Gotwals, A. W., & Alonzo, A. C. (2012). Introduction: Leaping into learning progressions in science. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 3-12). Rotterdam: Sense Publishers.

Gotwals, A. W., Songer, N. B., & Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions*. Rotterdam: Sense Publishers.

Grandy, R. (1997). Constructivisms and objectivity: Disentangling metaphysics from pedagogy. *Science and Education, 6*, 43-53.

Grayson, H., Houghton, E., O'Donnell, S., & Sargent, C. (2014). *Curriculum structures and stages in primary education: Audit of policy across jurisdictions. Technical appendix: Consolidated country tables*. Slough, UK: National Foundation for Educational Research.

Green, M. G. (1978). Structure and sequence in children's concepts of chance and probability: A replication study of Piaget and Inhelder. *Child Development, 49*(4), 1045-1053.

Gross, P., Buttrey, D., Goodenough, U., Koertge, N., Lerner, L., Schwartz, M., & Schwartz, R. (2013). *Final evaluation of the Next Generation Science Standards*. Retrieved from http://edex.s3-us-west-2.amazonaws.com/publication/pdfs/20130612-NGSS-Final-Review_7.pdf.

Gulliksen, H. (1987). *Theory of mental tests*. New York, NY: Routledge.

Hacking, I. (1992). 'Style' for historians and philosophers. *Studies in History and Philosophy of Science Part A, 23*(1), 1-20.

Hagell, P. (2014). Testing rating scale unidimensionality using the Principal Component Analysis (PCA)/t-test protocol with the Rasch model: The primacy of theory over statistics. *Open Journal of Statistics, 04*(06), 456-465.

Hager, P. J. (2006). Rethinking learning. In P. J. Hager & J. Halliday (Eds.), *Recovering informal learning: Wisdom, judgement and community* (7th ed., pp. 109-157). Dordrecht: Springer.

Haglund, J., Jeppsson, F., & Andersson, J. (2012). Young children's analogical reasoning in science domains. *Science Education, 4*(96), 725-756.

Hagquist, C., Bruce, M., & Gustavsson, J. P. (2009). Using the Rasch model in nursing research : An introduction and illustrative example. *International Journal of Nursing Studies, 46*, 380-393.

Haladyna, T. M. (2004). The future of item development and item response validation. In T. M. Haladyna (Ed.), *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Routledge.

Haladyna, T. M. (2013). Automatic item generation: A historical perspective. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice*. New York, NY: Routledge.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-333.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hall, R. A. (1983). *The revolution in science 1500 - 1750*. London, UK: Routledge.

Halwes, T. (2000). *The myth of the magical scientific method*. Retrieved from http://www.dharma-haven.org/science/myth-of-scientific-method.htm

Hambleton, R. K., & Russell, W. J. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hamers, J., de Koning, E., & Sijtsma, K. (1998). Inductive reasoning in third grade: Intervention promises and constraints. *Contemporary educational psychology, 23*, 132-148.

Han, J. (2013). *Scientific reasoning: Research, development, and assessment.* (Doctor of Philosophy), Ohio State University, Columbus, OH.

Hanushek, E. A., & Woessmann, L. (2010). *How much do educational outcomes matter in OECD countries?* Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17*, 267-321.

Harlen, W. (2001). Research in primary science education. *Journal of Biological Education, 35*, 61-65.

Harlen, W. (2006). *Teaching, learning and assessing science 5-12* (4th ed.). London, UK: SAGE Publications, Inc.

Harlen, W. (2007). The SPACE legacy. *Science, 97*, 13-15.

Harlen, W. (2008). Science as a key component of the primary curriculum: A rationale with policy implications. *Perspectives on Education (Primary Science), 1*, 4-18.

Harlen, W. (2010). Teaching primary science: How research helps. *Primary Science*(114), 5-8.

Harrison, A. G., & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education, 22*(9), 1011-1026.

Harrison, C., & Howard, S. (2010). Issues in primary assessment: 1 Assessment purposes. *Primary Science, 115*, 5-7.

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology, 4*, 246.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York, NY: The Free Press.

Henry, J. (2008). *The scientific revolution and the origins of modern science* (3rd ed.). Hampshire, UK: Palgrave Macmillan.

Her Majesty's Stationery Office. (1988). *Education Reform Act 1988 Chapter 40*. Retrieved from http://www.legislation.gov.uk/ukpga/1988/40/pdfs/ukpga_19880040_en.pdf

Hewson, P. W. (1981). A conceptual change approach to learning science. *European Journal of Science Education, 3*(4), 383-396.

Hill, C. T. (2007). The post-scientific society. *Issues in Science and Technology, 24*(1), 78-84.

Hofstein, A., & Mandler, V. (1985). The use of Lawson's test of formal reasoning in the Israeli science education context. *Journal of Research in Science Teaching, 22*(2), 141-152.

Hogan, D., Towndrow, P. A., & Koh, K. (2009). Instructional and assessment practices in Singapore. In E. L. Grigorenko (Ed.), *Multicultural psychoeducational assessment* (pp. 253-286). New York, NY: Springer Publishing Company, LLC.

Hogan, K., & Fisherkeller, J. (2005). Dialogue as data: Assessing students' scientific reasoning with interactive protocols. In J. J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), *Assessing science understanding: A human constructivist view* (pp. 95-127). Burlington, MA: Elsevier Academic Press.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 234-259). Oxford, UK: Oxford University Press.

Holyoak, K. J., & Morrison, R. G. (2012). Thinking and reasoning: A reader's guide. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 1-9). Oxford, UK: Oxford University Press.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y. H., Cantrell, K., Ryan, J., & Baird, J.-A. (2017). Lessons learned from PISA: A systematic review of peer-reviewed

articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research, 62*(3), 333-353.

Howe, C., Nunes, T., & Bryant, P. (2011). Rational number and proportional reasoning: Using intensive quantities to promote achievement in mathematics and science. *International Journal of Science and Mathematics Education, 9*, 391-417.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*, 10-16.

Hricko, M. (2008). Gagné's nine events of instruction. In L. A. Tomei (Ed.), *Encyclopedia of information technology curriculum integration*. Hershey, PA: IGI Global.

Huff, T. E. (2003). *The rise of early modern science: Islam, China, and the West* (2nd ed.). Cambridge, UK: Cambridge University Press.

Humphry, S. M. (2009). Modeling the effects of person group factors on discrimination. *Educational and Psychological Measurement, 70*(2), 215-231.

Huxley, R. (2003). Challenging the dogma: Classifying and describing the natural world. In K. Sloan (Ed.), *Enlightenment: Discovering the world in the eighteenth century* (pp. 70-79). London, UK: The British Museum Press.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York, NY: Basic Books.

Jaffe, K., Caicedo, M., Manzanares, M., Gil, M., Rios, A., Florez, A., . . . Davila, V. (2013). Productivity in physical and chemical science predicts the future economic growth of developing countries better than other popular indices. *PloS one, 8*(6), 1-10.

James, M. (2006a). Assessment and learning. In S. Swaffield (Ed.), *Unlocking assessment, understanding for reflection and application* (pp. 20-35). Abingdon, UK: Routledge.

James, M. (2006b). Assessment, teaching and theories of learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 47-60). London, UK: SAGE Publications, Inc.

James, M. (2017). (Re)viewing assessment: changing lenses to refocus on learning. *Assessment in Education: Principles, Policy & Practice, 24*, 404-414.

James, M., & Lewis, J. (2012). Assessment in harmony with our understanding of learning: Problems and possibilities. In J. Gardner (Ed.), *Assessment and learning* (pp. 187-205). London, UK: SAGE Publications, Inc.

Jiménez-Aleixandre, M. P., & Erduran, S. (2007). Argumentation in science education: An overview. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 3-25). Dordrecht: Springer.

John-Steiner, V., & Souberman, E. (1978). Afterword. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes. L. S. Vygotsky* (pp. 121-133). Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (2012). Inference with mental models. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning*. Oxford, UK: Oxford University Press.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29*, 95-110.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*, 112-133.

Jones, L. R., Wheeler, G., & Centurino, V. A. S. (2013). TIMSS 2015 Science Framework. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 29-59). Boston, MA: TIMSS & PIRLS International Study Centre, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Joyce, B., & Weil, M. (1972). *Conceptual complexity, teaching style and models of teaching*. Paper presented at the National Council for the Social Studies, Boston, MA. https://files.eric.ed.gov/fulltext/ED073965.pdf.

Joyce, B., Weil, M., & Calhoun, E. (2011). *Models of teaching* (8th ed.). Boston, MA: Pearson Education International.

Kamens, D. H., & Mcneely, C. L. (2009). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review, 54*(1), 5-25.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.

Kapon, S., & diSessa, A. A. (2012). Reasoning through instructional analogies. *Cognition and Instruction, 30*(3), 261-310.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., . . . Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42*(3), 139-151.

Keating, D. (1980). Thinking processes in adolescence. In J. Adelson (Ed.), *Handbook of adolescence* (pp. 211-246). New York, NY: John Wiley & Sons, Inc.

Kim, M., & Roth, W.-M. (2014). Argumentation as/in/for dialogical relation: A case study from elementary school science. *Pedagogies: An International Journal, 9*(4), 300-321.

Kim, M., Tan, A.-L., & Talaue, F. T. (2013). New vision and challenges in inquiry-based curriculum change in Singapore. *International Journal of Science Education, 35*(2), 289-311.

Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching, 50*(5), 530-560.

Kind, P. M., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education, 101*(1), 8-31.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

Klahr, D., & Carver, S. M. (1995). Commentary: Scientific thinking about scientific thinking. *Monographs of the Society for Research in Child Development, 60*(245), 137-151.

Klahr, D., Chen, Z., & Toth, E. (2001). Cognitive development and science education: Ships that pass in the night or beacons of mutual illumination? In S. Carver & D. Klahr (Eds.), *Cognition and Instruction* (pp. 75-119). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

Klahr, D., & Fay, A. L. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25*, 111-146.

Klahr, D., & Simon, H. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 125*(5), 524-543.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science, 333*, 971-975.

Klauer, K. J., Willmes, K., & Phye, G. D. (2002). Inducing inductive reasoning: Does it transfer to fluid intelligence? *Contemporary educational psychology, 27*(1), 1-25.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London, UK: Routledge.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development, 86*(1), 327-336.

Koretz, D., Lewis, E., Skewes-Cox, D., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress*. Retrieved from https://files.eric.ed.gov/fulltext/ED378220.pdf

Korom, E., Nagy, L., Németh, M. B., Radnóti, K., Makádi, M., Farkas, M. A., . . . Wagner, E. (2012). Detailed framework for diagnostic assessment of science. In B. C. Csapó & G. Szabó (Eds.), *Framework for diagnostic assessment of science*. Budapest: Nemzeti Tankönyvkiadó.

Korom, E., Németh, B., Antal, E., & Csapó, B. (2012). Diagnostic assessment frameworks for science: Theoretical background and practical issues. In B. Csapó & G. Szabó (Eds.), *Framework for diagnostic assessment of science* (pp. 147-174). Budapest: Nemzeti Tankönyvkiadó.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: The MIT Press.

Koufetta-Menicou, C., & Scaife, J. (2000). Teachers' questions - Types and significance in science education. *The School Science Review, 81*(296), 79-84.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: The classification of educational goals; Handbook II: Affective domain*. New York, NY: David McKay Company, Inc.

Kreiner, S. (2013). Analysing and reporting Rasch models. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health*. London, UK: ISTE Ltd & John Wiley & Sons Inc.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141-164). New York, NY: John Wiley and Sons, Inc.

Kuhn, D. (2008). Formal operations from a twenty-first century perspective. *Human Development, 51*, 48-55.

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education, 94*, 810-824.

Kuhn, D. (2011). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 497-523). Oxford, UK: John Wiley & Sons, Inc.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. San Diego, CA: Academic Press, Inc.

Kuhn, D., & Dean, D. J. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development, 5*(2), 261-288.

Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how). *Handbook of Child Psychology, Volume 2*, 953-993.

Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development, 1*, 113-129.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development, 74*(5), 1245-1260.

Kuhn, T. S. (1962/2012). *The structure of scientific revolutions* (4th ed.). Chicago, IL: University of Chicago Press.

Kwa, C. (2014). *Styles of knowing: A new history of science from ancient times to the present*. Pittsburgh, PA: University of Pittsburgh Press.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*, 389-433.

Lakoff, G., & Johnson, M. (1980/2003). Afterword. In G. Lakoff & M. Johnson (Eds.), *Metaphors we live by* (pp. 243-276). Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York, NY: Basic Books.

Lane, S. (2011). Issues in the design and scoring of performance assessments that assess complex thinking skills. In G. Schraw & D. R. Robinson (Eds.), *Assessment of higher order thinking skills* (pp. 265-302). Charlotte, NC: Information Age Publishing, Inc.

Lawley, J., & Tompkins, P. (2000). *Metaphors in mind: Transformation through symbolic modelling*. London: The Developing Company Press.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*(1), 11-24.

Lawson, A. E. (1985). A review of research on formal reasoning and science teaching. *Journal of Research in Science Teaching, 22*(7), 569-617.

Lawson, A. E. (2000). The generality of hypothetico-deductive reasoning: Making scientific thinking explicit. *The American Biology Teacher, 62*(7).

Lawson, A. E. (2002). *Science teaching and development of thinking*. Belmont, CA: Thomson Custom Publishing.

Lawson, A. E., Karplus, R., & Adi, H. (1978). The acquisition of propositional logic and formal operational schemata during the secondary school years. *Journal of Research in Science Teaching, 15*(6), 465-478.

Leach, J., & Scott, P. (2003). Individual and sociocultural views of learning in science education. *Science & Education, 12*, 91-113.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching, 39*(6), 497-521.

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching, 51*, 581-605.

Lee, Y.-J., Kim, M., & Yoon, H.-G. (2015). The intellectual demands of the intended primary science curriculum in Korea and Singapore: An analysis based on revised Bloom's taxonomy. *International Journal of Science Education, 37*, 2193-2213.

Leighton, J. P. (2004). The assessment of logical reasoning. In J. Leighton & R. Sternberg (Eds.), *The nature of reasoning* (pp. 291-312). Cambridge, UK: Cambridge University Press.

Leighton, J. P. (2013). Learning sciences, cognitive models, and automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice*. New York, NY: Routledge.

Leighton, J. P., & Gokiert, R. J. (2008). Identifying potential test item misalignment using student verbal reports. *Educational Assessment, 13*, 215-242.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*(3), 294-304.

Lenton, G., & Stevens, B. (2000). Numeracy in science: Understanding the misunderstandings. In J. Sears & P. Sorensen (Eds.), *Issues in science teaching* (pp. 80-88). London, UK: RoutledgeFalmer.

Lester, F. (2005). On the theoretical, conceptual, and philosophical foundations for research in mathematics education. *Zdm, 37*(6), 457-467.

Lim, L. (2013). Meritocracy, elitism, and egalitarianism: A preliminary and provisional assessment of Singapore's primary education review. *Asia Pacific Journal of Education, 33*(1), 1-14.

Linacre, J. M. (1994). Sample size and item calibration or [person measure] stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (1998a). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*, 266-283.

Linacre, J. M. (1998b). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions, 12*, 636.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Lindberg, D. C. (2007). *The beginnings of western science: The European scientific tradition in philosophical, religious, and institutional context, prehistory to A.D. 1450* (2nd ed.). Chicago, IL: University of Chicago Press.

Linn, M. C., Clement, C., & Pulos, S. (1983). Is it formal if it's not physics? (The influence of content on formal reasoning). *Journal of Research in Science Teaching, 20*(8), 755-770.

Linn, M. C., & Rice, M. (1979). A measure of scientific reasoning: The springs task. *Journal of Educational Measurement, 16*(1), 55-58.

Linn, R. L. (2006). *The Standards for Educational and Psychological Testing: Guidance in test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lipton, P. (2004). *Inference to the best explanation*. New York, NY: Routledge.

Liu, O. L., Lee, H.-S., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment, 13*(1), 33-55.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and non-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*(4), 615-630.

Lutkehaus, N. C., & Greenfield, P. (2003). From *The Process of Education* to *The Culture of Education*: An intellectual biography of Jerome Bruner's contributions to education. In B. J. Zimmerman & D. H. Schunk (Eds.), *Educational psychology: A century of contributions* (pp. 409-430). Mawah, NJ: Erlbaum.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382-385.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46.

Mahootian, F., & Eastman, T. E. (2012). Complementary frameworks of scientific inquiry: Hypothetico-deductive, hypothetico-inductive, and observational-inductive. *World Futures: Journal of General Evolution, 65*(1), 61-75.

Marais, I., & Andrich, D. (2008). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*, 200-215.

Markus, K. A. (2016). Alternative vocabularies in the test validity literature. *Assessment in Education: Principles, Policy & Practice, 23*, 252-267.

Markus, K. A., & Smith, K. M. (2010). Content validity. In N. J. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 239-243). Thousand Oaks: SAGE Publications, Inc.

Martens, K., Niemann, D., & Teltemann, J. (2016). Effects of international assessments in education – a multidisciplinary review. *European Educational Research Journal, 15*(5), 516-522.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*, 1332-1361.

Masterson, M. J., & Rainer, J. R. K. (2009). The nomological network and the research continuum. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (2nd ed., pp. 2827-2833). Hershey, PA: Information Science Reference.

Maurer, T. J., & Pierce, H. R. (1998). A comparison of Likert scale and traditional measures of self-efficacy. *Journal of Applied Psychology, 83*(2), 324-329.

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29*, 43-55.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299-324). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis, and application of psychological and educational tests*. The Hague: Eleven International.

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70*, 308-312.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (4th ed., pp. 13-103). New York, NY: American Council on Education/Macmillian.

Messick, S. (1993). *Foundations of validity: Meaning and consequences in psychological assessment*. Retrieved from Princeton, NJ: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1993.tb01562.x

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5-8.

Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Millar, R. (1989). What is 'scientific method' and can it be taught? In J. Wellington (Ed.), *Skills and processes in science education: A critical analysis* (pp. 47-62). London, UK: Routledge.

Millar, R. (1991). A means to an end: The role of processes in science education. In B. Woolnough (Ed.), *Practical science: The role and reality of practical work in school science*. Buckingham: Open University Press.

Millar, R., & Driver, R. (1987). Beyond processes. *Studies in Science Education, 14*(1), 33-62.

Millar, R., Lubben, F., Gott, R., & Duggan, S. (1994). Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance. *Research Papers in Education, 9*, 207-248.

Millar, R., & Osborne, J. (1998). *Beyond 2000: Science education for the future* (1 871984 78 5). Retrieved from https://www.nuffieldfoundation.org/sites/default/files/Beyond%202000.pdf.

Miller, J. M., & Penfield, R. D. (2005). Using the score method to construct asymmetric confidence intervals: An SAS program for content validation in scale development. *Behavior Research Methods, 37*(3), 450-452.

Miller, P. H. (2010). Piaget's theory: Past, present, and future. In U. Goswami (Ed.), *Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 649-672). Hillsdale, NJ: Wiley-Blackwell.

Ministry of Education Singapore. (2001). *Science Syllabus Primary*. Singapore: Author.

Ministry of Education Singapore. (2009). Syllabus reviewed every six years. Retrieved from https://www.moe.gov.sg/news/forum-letter-replies/syllabus-reviewed-every-six-years

Ministry of Education Singapore. (2013). *Science Syllabus Primary 2014*. (978-981-05-8232-6). Retrieved from http://www.moe.gov.sg/education/syllabuses/sciences/files/science-primary-2008.pdf.

Molitor, L. L., & George, K. D. (1976). Development of a test of science process skills. *Journal of Research in Science Teaching, 13*(5), 405-412.

Morris, B. J., Croker, S., Masnick, A. M., & Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B. J. Morris, & J. L. Amaral (Eds.), *Current topics in children's learning and cognition* (pp. 61-82). London, UK: InTech.

Morse, J. (2010). Procedures and practice of mixed method design: Maintaining control, rigor, and complexity. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: SAGE Publications, Inc.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research, 40*(1), 120-123.

Murphy, P. (2000). Are gender differences in achievement avoidable? In J. Sears & P. Sorensen (Eds.), *Issues in science teaching* (pp. 165-174). London, UK: RoutledgeFalmer.

National Research Council. (1996). *National Science Education Standards*. Washington DC: The National Academies Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

National Research Council. (2007). Learning progressions. In R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.), *Taking science to school: Learning and teaching science in grade K-8* (pp. 213-250). Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas* (9780309217422). Retrieved from Washington, DC: https://www.nap.edu/catalog/13165/a-framework-for-k-12-science-education-practices-crosscutting-concepts

Nersessian, N. J. (1995). *Opening the black box: Cognitive science and history of science* (2nd ed. Vol. 10). Chicago, IL: The University of Chicago Press on behalf of the History of Science Society.

Netz, R. (1999). *The shaping of deduction in Greek mathematics*. Cambridge, UK: Cambridge University Press.

Newton, P. E. (2012a). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research & Perspective, 10*(1-2), 1-29.

Newton, P. E. (2012b). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research & Perspective, 10*(1-2), 110-122.

Newton, P. E., & Baird, J.-A. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice, 23*(2), 173-177.

Newton, P. E., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education, 21*(5), 553-576.

Newton, P. E., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London, UK: SAGE Publications, Inc.

Ng, Y. P. D. (2017). A somewhat brief history of research on scientific thinking and reasoning. In E. Schultz (Ed.), *St. Anne's academic review 7* (pp. 81-90). Oxford, UK: St. Anne's College Middle Common Room, University of Oxford.

Ng, Y. P. D. (in press). Assessment of argumentation in chemistry: A model for designing items. In S. Erduran (Ed.), *Argumentation in chemistry education: Research, policy and practice*. London, UK: Royal Society of Chemistry.

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states. Volume 1: The standards - Arranged by disciplinary core ideas and by topics* (978-0-309-27227-8). Retrieved from Washington, DC: https://www.nap.edu/catalog/18290/next-generation-science-standards-for-states-by-states

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students* (5th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Nowak, K. H., Nehring, A., Tiemann, R., & Upmeier zu Belzen, A. (2013). Assessing students' abilities in processes of scientific inquiry in biology using a paper-and-pencil test. *Journal of Biological Education, 47*(3), 182-188.

Nudzor, H. P. (2009). A critical commentary on combined methods approach to researching educational and social issues. *Issues in Educational Research, 19*(2), 114-127.

O'Donnell, J. R. (2011). *Creation of national norms for scientific thinking skills using the Classroom Test of Scientific Reasoning.* (Master of Science in Education), Winona State University, Winona, MN.

OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving* (Revised ed.). Paris: PISA, OECD Publishing.

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research, 4*(1), 56-78.

Onwuegbuzie, A. J., & Combs, J. P. (2010). Emergent data analysis techniques in mixed methods research: A synthesis. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (2nd ed., pp. 397-430). Thousand Oaks, CA: SAGE Publications, Inc.

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning - a review of test instruments. *Educational Research and Evaluation*, 1-24.

Oppenheim, A. N. (2005). *Questionnaire design, interviewing and attitude measurement*. London: Continuum.

Osborne, J. (2011). Science teaching methods: A rationale for practices. *School Science Review, 93*, 93-103.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265-279.

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education, 25*, 177-196.

Osborne, J., Collins, S., Ratcliffe, M., Millar, R., & Duschl, R. (2003). What "ideas-about-science" should be taught in school science? A Delphi study of the expert community. *Journal of Research in Science Teaching, 40*(7), 692-720.

Osborne, J., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*, 821-846.

Osborne, R. (1985). Children's own concepts. In W. Harlen (Ed.), *Primary science: Taking the plunge* (pp. 75-90). London, UK: Heinemann Educational Books.

Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children's social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology, 53*(3), 450-462.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance and other formats*. Boston, MA: Klewer Academic Publishers.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1-18.

Passmore, C., & Stewart, J. (2002). A modeling approach to teaching evolutionary biology in high schools. *Journal of Research in Science Teaching, 39*(3), 185-204.

Pellegrino, J. W. (2017). The two disciplines problem – 'it's like Déjà vu all over again!'. *Assessment in Education: Principles, Policy & Practice, 24*, 359-368.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century* (978-0-309-25649-0). Retrieved from https://www.nap.edu/catalog/13398/education-for-life-and-work-developing-transferable-knowledge-and-skills

Pellegrino, J. W., & Wilson, M. (2015). Assessment of complex cognition: Commentary on the design and validation of assessments. *Theory Into Practice, 54*, 263-273.

Penfield, R. D. (2003). A score method of constructing asymmetric confidence intervals for the mean of a rating scale item. *Psychological Methods, 8*(2), 149-163.

Penfield, R. D., & Giacobbi Jr, P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213-225.

Penfield, R. D., & Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement in Education, 17*(4), 359-370.

Penuel, W. R., & Shepard, L. A. (2016). *Assessment and Teaching* (5th ed.). Washington, DC: American Educational Research Association.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18*, 16-25.

Phipps, R., & Duggan, S. (1995). Practical work in science in the primary school. *Education 3-13, 23*(2), 23-27.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15*(1), 1-12.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children* (J. Leake, L., P. Burrell, & H. D. Fishbein, Trans.). London, UK: Routledge & Kegan Paul.

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research, 63*(2), 167-199.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health, 29*, 489-497.

Pollitt, A., & Ahmed, A. (2001). *Science or reading?: How students think when answering TIMSS questions*. Paper presented at the International Association for Educational Assessment, Rio de Janeiro, Brazil.

Pollitt, A., & Ahmed, A. (2008). *Outcome space control and assessment*. Paper presented at the Association for Educational Assessment - Europe, Hissar.

Poon, C.-L. (2014). Five decades of science education in Singapore. In A.-L. Tan, S. L. S. Lim, & C.-L. Poon (Eds.), *Inquiry into the Singapore science classroom: Research and practices* (pp. 1-25). Singapore: Springer.

Popham, W. J. (1992). Appropriate expectations for content judgements regarding teacher licensure tests. *Applied Measurement in Education, 5*, 285-301.

Posner, G., Strike, K., Hewson, P., & Gertzog, W. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227.

Pruitt, S. L. (2014). The Next Generation Science Standards: The features and challenges. *Journal of Science Teacher Education, 25*, 145-156.

Qualifications and Curriculum Development Agency. (2009). *The National Strategies: Assessing pupils' progress in primary science*. Retrieved from http://webarchive.nationalarchives.gov.uk/20110812092626/http://nsonline.org.uk/node/20683

Rabacal, J. S. (2016). Test of science process skills of biology students towards developing of learning exercises. *Asia Pacific Journal of Multidisciplinary Research, 4*(4), 9-16.

Rasch, G. (1960/1993). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Raven, J. (2008). *Manual for Standard Progressive Matrices - plus version and Mill Hill Vocabulary Scale*. London, UK: Pearson.

Reddy, M. (1978). The conduit metaphor: A case of frame conflict in our language. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed., pp. 164-201). Cambridge, UK: Cambridge University Press.

Resnick, L. B. (1983). Mathematics and science learning: A new conception. *Science, 220*(4596), 477-480.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116.

Rips, L. J., & Conrad, F. G. (1989). Folk psychology of mental activities. *Psychological review, 96*(2), 187-207.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools, 45*(2), 158-176.

Roberts, R., & Gott, R. (2004). A written test for procedural understanding: A way forward for assessment in the UK science curriculum? *Research in Science and Technological Education, 22*(1), 5-21.

Roe, A. (1953). *The making of a scientist*. New York, NY: Dodd, Mead & Company.

Romine, W. L., Sadler, T. D., & Kinslow, A. T. (2017). Assessment of scientific literacy: Development and validation of the Quantitative Assessment of Socio-Scientific Reasoning (QuASSR). *Journal of Research in Science Teaching, 54*(2), 274-295.

Ross, J. A., & Maynes, F. J. (1983). Development of a test of experimental problem-solving skills. *Journal of Research in Science Teaching, 20*(1), 63-75.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*, 49-60.

Rowe, S. M., & Wertsch, J. V. (2004). Vygotsky's model of cognitive development. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 538-554). Malden, MA: Blackwell Publishing Ltd.

Rudolph, J. L., & Horibe, S. (2016). What do we mean by science education for civic engagement? *Journal of Research in Science Teaching, 53*(6), 805-820.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching, 49*(6), 691-712.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*(2), 99-141.

Rusilowati, A., Kurniawati, L., Nugroho, S. E., & Widiyatmoko, A. (2016). Developing an instrument of scientific literacy asessment on the cycle theme. *International Journal of Environmental and Science Education, 11*(12), 5718-5727.

Şahin, A., & Anil, D. (2016). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice, 17*(1), 321-335.

Salzberger, T. (2013). Reporting a Rasch analysis. In K. B. Christensen, S. Kreiner, & M. Mounir (Eds.), *Rasch models in health* (pp. 347-362). London: ISTE Ltd & John Wiley and Sons Inc.

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*, 447-472.

Sanderson, B. A., & Kratochvil, D. W. (1971). *Science - A process approach*. Retrieved from https://files.eric.ed.gov/fulltext/ED064066.pdf.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education, 89*, 634-656.

Sandoval, W. A., & Çam, A. (2011). Elementary children's judgments of the epistemic status of sources of justification. *Science Education, 95*, 383-408.

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88*, 345-372.

Sandoval, W. A., Sodian, B., Koerber, S., & Wong, J. (2014). Developing children's early competencies to engage with science. *Educational Psychologist, 49*(2), 139-152.

Sanz de Acedo Lizarraga, M. L., Sanz de Acedo Baquedano, M. T., & Oliver, M. S. (2010). Psychological intervention in thinking skills with primary education students. *School Psychology International, 31*(2), 131-145.

Scaife, J. (2000). Learning in science. In J. Wellington (Ed.), *Teaching and learning secondary science: Contemporary issues and practical approaches* (pp. 61-108). London, UK: Routledge.

Scheffler, I. (1960). *The language of education*. Springfield, IL: Charles C. Thomas.

Schraw, G. (2011). An overview of thinking skills. In G. Schraw, M. T. McCrudden, S. Lehman, & B. Hoffman (Eds.), *Assessment of higher order thinking skills* (pp. 19-45). Charlotte, NC: Information Age Publishing, Inc.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*(2), 162-176.

Schunk, D. H. (2014). *Learning theories: An educational perspective* (6th ed.). Essex: Pearson Education Limited.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., . . . Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632-654.

Searle, J. R. (1990). *Is the Brain a Digital Computer?* Paper presented at the Proceedings and Addresses of the American Philosophical Association.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27*(2), 4-13.

Shayer, M. (1999). Cognitive acceleration through science education II: Its effects and scope. *International Journal of Science Education, 21*(8), 883-902.

Shayer, M., Adey, P., & Wylam, H. (1981). Group tests of cognitive development ideals and a realization. *Journal of Research in Science Teaching, 18*, 157-168.

Shepard, L. A., Penuel, W. R., & Davidson, K. L. (2017). Design principles for new systems of assessment. *Phi Delta Kappan, 98*(6), 47-52.

Shoemaker, D. M. (1971). *Principles and procedures of multiple matrix sampling*. Southwest Regional Educational Lab. Inglewood, CA. Retrieved from https://files.eric.ed.gov/fulltext/ED057100.pdf.

Siegel, H. (1995). Why should educators care about argumentation? *Philosophy Articles and Papers, 11*, 159-176.

Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika, 71*(3), 451-455.

Sireci, S. G. (1998a). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Sireci, S. G. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299-321.

Sireci, S. G. (2015). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, 1-10.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100-107.

Sireci, S. G., & Geisinger, K. F. (1995). Using subject-matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement, 19*, 241-255.

Sirotnik, K., & Wellington, R. (1977). Incidence sampling: An integrated theory for "matrix sampling". *Journal of Educational Measurement, 14*, 343-399.

Smith, E. V. J. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

Smith, E. V. J. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*(2), 205-231.

Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3*(2), 115-163.

Smith, K. (2016). *Working scientifically: A guide for primary science teachers*. Oxford, UK: Routledge.

Sodian, B., & Bullock, M. (2008). Scientific reasoning - Where are we now? *Cognitive Development, 23*, 431-434.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*(4), 753-766.

Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical teacher, 28*, 566-568.

Stefanich, G. P., Unruh, R. D., Perry, B., & Phillips, G. (1983). Convergent validity of group tests of cognitive development. *Journal of Research in Science Teaching, 20*(6), 557-563.

Su, J. M., Lin, H. Y., Tseng, S. S., & Lu, C. J. (2011). Opass: An online portfolio assessment and diagnosis scheme to support web-based scientific inquiry experiments. *Turkish Online Journal of Educational Technology, 10*(2), 151-173.

Sundre, D. L. (2008). *The scientific reasoning rest, version 9 (SR-9)*. Harrisonburg, VA: The Centre for Assessment and Research Studies.

Tamir, P. (1974). An inquiry oriented laboratory examination. *Journal of Educational Measurement, 11*(1), 25-33.

Tan, K. P. (2008). Meritocracy and elitism in a global city: Ideological shifts in Singapore. *International Political Science Review, 29*(1), 7-27.

Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods to the social and behavioral sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 3-50). Thousand Oaks, CA: SAGE.

Teddlie, C., & Tashakkori, A. (2012). Common "core" characteristics of mixed methods research. *American Behavioral Scientist, 56*(6), 774-788.

Tennant, A., & Pallant, A. (2006). Unidimensionality matters! (A tale of two smiths). *Rasch Measurement Transactions, 20*(1), 1048-1051.

Thelk, A. D., & Hoole, E. R. (2006). What are you thinking? Postsecondary student think-alouds of scientific and quantitative reasoning items. *The Journal of General Education, 55*(1), 17-39.

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a 'universal' rubric for assessing undergraduates' scientific

reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education, 36*(5), 509-547.

Tobin, K. G., & Capie, W. (1981). The development and validation of a group test of logical thinking. *Educational and Psychological Measurement, 41*, 413-423.

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy Press.

Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (1981). *On scientific thinking*. New York, NY: Columbia University Press.

Urbach, P. (1987). *Francis Bacon's philosophy of science: An account and a reappraisal*. La Salle, IL: Open Court Publishing Company.

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28*, 317-331.

van Wyke, J., & Andrich, D. (2006). *A typology of polytomously scored mathematics items disclosed by the Rasch model: Implications for constructing a continuum of achievement*. Murdoch University. Perth, Western Australia.

Vickers, A. J. (1999). Comparison of an ordinal and a continuous outcome measure of muscle soreness. *International Journal of Technology Assessment in Health Care, 15*, 709-716.

Vosniadou, S. (2013). Conceptual change in learning and instruction: The framework theory approach. In S. Vosniadou (Ed.), *The international handbook of conceptual change* (2nd ed., pp. 11-30). New York, NY: Routledge.

Vosniadou, S. (2014). Examining cognitive development from a conceptual change point of view: The framework theory approach. *European Journal of Developmental Psychology, 11*(6), 645-661.

Vygotsky, L. (1994). The development of thinking and concept formation in adolescence. In R. van der Veer & J. Valsiner (Eds.), *The Vygotsky reader* (pp. 185-265). Oxford, UK: Blackwell.

Watkins, C. (2003). *Learning: A sense-maker's guide*. London: Association of Teachers and Lecturers.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences* (Revised ed.). Oxford: Rand Mcnally.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (2000). *Unobtrusive measures* (Revised ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states. Research monograph no. 18*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25.

Weld, B. J., Stier, M., & Mcnew-Birren, J. (2011). The development of a novel measure of scientific reasoning growth among college freshmen: The Constructive Inquiry Science Reasoning Skills Test. *Journal of College Science Teaching, 40*(4), 101-107.

Wellington, J. (2000). *Teaching and learning secondary science: Contemporary issues and practical approaches*. London, UK: Routledge.

Wenger, E. (1998). *Communities of practice*. Cambridge, UK: Cambridge University Press.

Wertheimer, M. (1945). *Productive thinking*. New York, NY: Harper.

Wertsch, J. V. (1998). *Mind as action*. New York, NY: Oxford University Press.

Wiliam, D. (2014). What do teachers need to know about the new Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice, 33*(4), 29-30.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Woodcock, B. A. (2014). "The scientific method" as myth and ideal. *Science & Education, 23*(10), 2069-2093.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA press.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range, Inc.

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 119-145). New York, NY: Cambridge University Press.

Yore, L. D., Hand, B., Goldman, S. R., Hildebrand, G. M., Osborne, J. F., Treagust, D. F., & Wallace, C. S. (2004). New directions in language and science education research. *Reading Research Quarterly, 39*(3), 374-352.

Zeineddin, A., & Abd-El-Khalick, F. (2010). Scientific reasoning and epistemological commitments: Coordination of theory and evidence among college science students. *Journal of Research in Science Teaching, 47*(9), 1064-1093.

Zembal-Saul, C. L., McNeill, K. L., & Hershberger, K. (2012). *What's your evidence? Engaging K-5 children in constructing explanations in science*. Boston, MA: Pearson Education.

Zhai, J., Jocz, J. A., & Tan, A.-L. (2014). 'Am I like a scientist?': Primary children's images of doing science in school. *International Journal of Science Education, 36*, 553-576.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99-149.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172-223.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zumbo, B. D. (2014). What role does, and should, the test standards play outside of the United States of America? *Educational Measurement: Issues and Practice, 33*(4), 31-33.

Zumbo, B. D., & Chan, E. K. H. (2014). Setting the stage for validity and validation in social, behavioral, and health sciences: Trends in validation practices. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 3-8). Switzerland: Springer.

Zumbo, B. D., & Hubley, A. M. (2017). Response processes in the context of validity. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1-12). Switzerland: Springer International Publishing AG.