

Cognitive evidence for construct validity of the IELTS Reading Comprehension Module:
Content analysis, test taking processes, and experts' accounts

By

Raof Moeini

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfilment of the
requirements for the degree of

Doctor of Philosophy

in

Applied Linguistics and Discourse Studies

Carleton University Ottawa, Ontario

Raof Moeini©

2020

Dedication

To My Daughter Ariana
and
My Parents

Abstract

Despite the growing demand for the International English Language Test System (IELTS) Academic as a high-stakes language proficiency test, and its significant impact on the lives of test takers, only a limited amount of validity research has been published. Among the four modules of the test, the Reading Comprehension Module (RCM) has been the least researched, with few studies examining specific dimensions of the reading construct operationalized by the test (e.g., Bax, 2013; Weir, Hawkey, Green, & Devi, 2009). Following Messick's (1989) advice that construct is central in considerations of validity and informed by Khalifa and Weir's (2009) Multi-Componential Reading Comprehension Model, this qualitative case study investigated the IELTS RCM construct. It drew evidence from three phases: 1) content analysis of an RCM sample test; 2) verbal reports of the processes used during RCM performance, elicited from three groups of test takers (N= 21) with different language backgrounds (i.e., first/L1 English; second/L2, English as a Foreign Language (EFL) learners in Iran; across levels of language proficiency); and, 3) verbal reports and interviews with (N= 10) testing experts, who judged the skills, knowledge sources, processes, and strategies (SKPSs) (Gorin, 2006), that were tapped by each RCM task. Results of the content analysis showed gaps in representing features and practices of academic reading at both levels of texts and tasks. Coding (Saldaña, 2009) of the test takers' verbal accounts indicated the test construct differed across language backgrounds and levels of language proficiency. For L2 test takers, test performance was basically task-based, consisting of careful reading at sentence and inter-sentential levels rather than meaningful comprehension, and raising questions about the test as a measure of the academic reading construct. The experts' accounts reinforced findings from the test takers' reports and suggest that the RCM tasks tap into micro, lexico-grammatical features rather than macro, textual comprehension. Higher order inferential global comprehension reading skills were disproportionately underrepresented. Implications for different stakeholders are discussed. The study concludes that more research is essential in order to justify the use of IELTS RCM as an English language proficiency measure.

Acknowledgement

This study is a culmination of two years of work, which would have not been possible without the invaluable support I received from my professors and friends.

My most heartfelt gratitude goes to my thesis supervisor, Professor Janna Fox, whose inspiring and kind personality and insightful comments and feedback contributed a lot to the quality of the study. With her deep knowledge and expertise, she walked with me every step of the way and provided generous support throughout the study. I am grateful for working with you Janna.

I owe much gratitude to Dr. David Wood for the detailed and insightful feedback he provided me and Dr. Tracy Hodgson-Drysdal for reading the research draft and the suggestions she made. I am also very grateful to Dr. Scott Douglas from University of British Columbia University, Dr. Natasha Artemeva, and Dr. Randall Gess for agreeing to be my thesis examiners, for patiently reading my dissertation with enthusiasm, and for their thought-provoking and eye-opening questions during the defence.

I would also like to thank All the L1 test takers and L2 test takers who were helpful and cooperative during the long hours of data collections procedures. Special thanks should also go to Joan Grant, our program administrator, for being so supportive and helpful through these years.

I am also very grateful to family and friends for providing constant emotional support while I was conducting the research and writing this dissertation.

Very special thanks go to my dear family back in Iran, my beloved brothers and sisters for all their encouragement, emotional support and positive energy especially my parents who have always supported me. Last but not least, I should thank my daughter Ariana who has always been the very meaning of life for me.

Table of Contents

Table of Contents

<i>Dedication</i>	<i>ii</i>
<i>Abstract</i>	<i>iii</i>
<i>Acknowledgement</i>	<i>iv</i>
<i>Table of Contents</i>	<i>v</i>
<i>List of Tables</i>	<i>x</i>
<i>List of Figures</i>	<i>xiii</i>
<i>List of Appendices</i>	<i>xv</i>
<i>List of Abbreviations</i>	<i>xvi</i>
<i>Glossary</i>	<i>xvii</i>
CHAPTER ONE: INTRODUCTION	1
1.1. Statement of the Problem	1
1.2. Rationale of the Study	4
1.3. Research Questions	8
1.4. Organization of the Dissertation	10
CHAPTER TWO: THEORETICAL FRAMEWORK	11
2.1. Introduction	11
2.2. Brief History of Validity Theory	12
2.2.1. Unified validity theory	14
2.2.3. Test usefulness	18
2.3. Criticism of the Unified Validity Theory	19
2.4. Gorin's (2006) Process-based Model of Construct Validity	23
2.5. Test Validation	24
2.5.1. The quantitative approach to validity	26
2.5.2. The qualitative approaches to validity	28
2.6. Sources of Validity Evidence	31
2.6.1. Content analysis evidence.....	32
2.6.2. Test taking processes and verbal reports.....	36
2.6.3. Experts' judgements and accounts	38
2.7. Summary of the Chapter.....	41
CHAPTER THREE: LITERATURE REVIEW	42
3.1. Introduction	42

3.2. Khalifa and Weir’s (2009) cognitive reading comprehension model.....	43
3.3. Central core component: Cognitive processes	45
3.3.1. Text (discourse) comprehension processes	46
3.3.2. Brown and Abeywickrama’s (2004) dimensions of reading	48
3.4. The Metacognitive component: Metacognitive strategies.....	50
3.4.1. Metacognitive strategies and reading: Strategic Competence Theory	53
3.4.2. Metacognitive strategies and reading: Construction-Integration Theory	54
3.5. Knowledge base component: Knowledge sources.....	57
3.5.1. The Knowledge base component and Schema Theory	58
3.6. Validation model of IELTS RCM construct	60
3.7. Khalifa and Weir’s (2009) Model and Assessing Reading	62
3.8. Construct Validity of the IELTS RCM: Evidence from Empirical Research	64
3.9. Cognitive Processes and Metacognitive Strategies in Reading Comprehension	71
3.10. Influence of Readers’ Characteristics on Reading Processes	75
3.10.1. Influence of language background on reading comprehension processes	76
3.10.2. Influence of level of language proficiency on reading comprehension processes	77
3.10.3. Readers’ attitudes towards the test.....	80
3.11. Influence of Text Features on Comprehension Processes and Test Performance	81
3.11.1. Influence of linguistic features of the text	82
3.11.2. Text content: Topical/ Background knowledge	83
3.11.3. Propositional density: Readability.....	84
3.11.4. Text type	85
3.11.5. Text authenticity manipulation.....	86
3.11.6. Text coherence.....	87
3.11.7. Order of text information and test tasks	87
3.11.8. Stem structure and its impact on test performance	89
3.12. Summary of the Chapter	90
CHAPTER FOUR: METHODOLOGY.....	92
4.1. Introduction	92
4.2. Research Design	92
4.3. Sampling Design	95
4.3.1. Inclusion criteria (Phases Two and Three)	95
4.3.2. Participant recruitment.....	97
4.4. Participants.....	97
4.4.1. Undergraduate L1 English speakers	97
4.4.2. Undergraduate L2 English learners	97
4.4.3. Expert Participants	98
4.5. Instruments.....	99
4.5.1. Cloze test.....	99
4.5.2. Reading essay.....	100
4.5.3. IELTS Reading Comprehension Module: Pilot test	102
4.5.4. IELTS Reading Comprehension Module: The main test	103
4.5.5. Readability of the IELTS RCM and academic texts	104

4.5.6. Test Performance Observation Scheme (TPOS)	105
4.5.7. Construct of the IELTS RCM Questionnaire	105
4.6. Procedures (Data Collection).....	108
4.6.1. Phase one: The sample IELTS RCM.....	108
4.6.2. Phase two: Test takers' accounts	109
4.6.3. Phase three: testing experts' judgements and accounts	114
4.7. Data Analysis.....	115
4.7.1. Phase one: analysis the IELTS RCM	115
4.7.2. Phase two: Analysis of test takers' accounts	116
4.7.3. Second cycle of coding	123
4.7.4. Phase three: Analysis of testing experts' judgements and accounts	123
4.8. Summary of the Chapter	123
CHAPTER FIVE: RESULTS: CONTENT ANALYSIS	125
5.1. Introduction	125
5.2. Results of Test Task Analysis of the Sample IELTS RCM	126
5.2.1. Passage One- "Raising the Mary Rose"	126
5.2.2. Passage Two- "What destroyed Easter Island?"	135
5.2.3. Passage Three- "Neuraesthetics"	143
5.3. Comparison of text features of the IELTS RCM texts and academic texts	152
5.4. Comparison of the lexical profile of the IELTS RCM texts and academic texts	155
5.5. Summary of Answer Level, Skills Measured, and Type of Comprehension	157
5.6. Reading dimensions in the IELTS RCM (Brown and Abeywickrama, 2004).....	159
5.7. Summary of the Chapter.....	161
CHAPTER SIX: RESULTS OF TEST TAKERS' ACCOUNTS	163
6.1. Introduction	163
6.2. Construct of the IELTS RCM Test Tasks	165
6.2.1. Construct of the True, False, Not given Task.....	166
6.2.3. Construct of the Diagram Completion Task	174
6.2.4. Construct of the <i>Matching Headings Task</i>	180
6.2.5. Construct of the Summary Completion Task 1	186
6.2.6. Construct of the Multiple-Choice (two- answers) Task.....	189
6.2.7. Construct of the <i>Multiple-Choice Task</i>	192
6.2.8. Construct of the Summary Completion Task 2	197
6.2.9. Construct of the <i>Yes, No, Not given Task</i>	202
6.3. Models of IELTS RCM Test Performance of L1 and L2 Test Takers	207
6.4. Results of the First and the Second Cycle of Coding	209
6.4.1. Results of the "Reading Theme" in test performance	211
6.4.2 Results of "Searching Theme" in the IELTS RCM Test Performance	217
6.4.3 Results of Answering theme in the IELTS RCM Test Performance	229
6.4.4. Results of Strategies Used in IELTS RCM Test Performance.....	235
6.5. Results of Test Performance Observation Scheme (TPOS)	245
6.5.1. Frequency of back and forth movements between the text and test items (B&FM)	245

6.5.2. Time spent on the test tasks	249
6.6. Difficulty of the IELTS RCM Tasks	251
6.7. Summary of the Chapter	256
CHAPTER SEVEN: RESULTS OF EXPERTS' ACCOUNTS AND JUDGEMENTS.....	258
7.1. Introduction	258
7.2. The Level of Reading.....	258
7.3. The Type of Knowledge Needed.....	260
7.4. The Types of Comprehension	262
7.5. The Type of Reading (careful-expeditious).....	265
7.6. Text Difficulty.....	267
7.8. The Level of Processing.....	269
7.9. Time Needed.....	272
7.10. The Reading Skills Measured.....	273
7.11. Summary of Experts' Judgements	276
7.12. Results of experts' accounts and judgements of the test tasks	278
7.12.1. Construct of The True, False, Not Given Tasks.....	278
7.12.2. Construct of The <i>Matching Features Task</i>	278
7.12.3. Construct of The Diagram Completion Task.....	278
7.12.4. Construct of The Matching Headings Task.....	279
7.12.5. Construct of The Summary Completion Task 1	279
7.12.6. Construct of The Multiple-Choice (two answers) Task	279
7.12.7. Construct of The <i>Multiple-Choice Task</i>	279
7.12.8. Construct of The Summary Completion Task 2	280
7.12.9. Construct of The Yes, No, Not given Task	280
7.13. Is IELTS RCM a measure of academic reading?	280
7.14. Summary of the Chapter	282
CHAPTER EIGHT: DISCUSSION	283
8.1. Introduction	283
8.2. Discussion	284
8.2.1. Khalifa and Weir's (2009) model of reading comprehension	287
8.2.2. Inferencing	291
8.2.3. The types of reading.....	293
8.2.4. Local-global comprehension	295
8.2.5. Task difficulty	298
8.2.6. Reading fluency (speededness).....	300
8.2.7. Use of background knowledge.....	303
8.3. Summary of the Chapter	306
CHAPTER NINE: CONCLUSION, IMPLICATIONS,	307
AND SUGGESTIONS FOR FUTURE RESEARCH.....	307

9.1. Introduction	307
9.2. Conclusion	307
9.3. Pedagogical Implications and applications	310
9.3.1. Implications for IELTS RCM preparation course instructors and reading instructors	310
9.3.2. Implications for test takers	312
9.3.3. Implications for test developers	312
9.3.4. Implications for policy makers and test users.....	315
9.4. Limitations of the Study	316
9.5. Suggestions for Future Research	317
9.5.1. The Academic Reading Construct	318
9.5.2. Sub-construct of the IELTS RCM.....	320
9.5.3. Test task features	320
9.5.4. Test bias and item bias.....	321
9.5.5. General IELTS RCM	322
9.5.6. Research designs	322
9.5.7. Consequential basis of Validity	324
9.5.8. Test improvements	325
9.6. Concluding Remarks	326
References	329
Appendices	364
Appendix A: Test takers' invitation email	364
Appendix B: Testing Experts' invitation email	365
Appendix C: Consent form	366
Appendix D: Proficiency Cloze Test	367
Appendix E: Reading Essay	368
Appendix F: Writing Summary Task	371
Appendix G: Sample RCM IELTS	372
Appendix H: Test Performance Observation Scheme	382
Appendix I: Construct of the RCM IELTS Questionnaire	383
Appendix J: Rating scale for the summary tasks	384
Appendix K: Definitions, examples, and distribution of the processes and strategies used in Searching Theme	385
Appendix L: Definitions, examples, and distribution of the processes and strategies used in Answering Theme	392
Appendix M: Definitions and examples of the metacognitive strategies used	398

List of Tables

Table 3.1. Brown and Abeywickrama's (2004) classification of reading comprehension dimensions (types)	51
Table 4.1. Demographic details of the research participants	99
Table 4.2. Testing experts' characteristics	100
Table 4.3. Readability of the Cloze test and the reading essay	102
Table 4.4. Readability and linguistic features of the research instruments.	105
Table 4.5. Construct of IELTS RCM Questionnaire	108
Table 4.6. Rating scale for essay questions	117
Table 4.7. Results of language proficiency testing of L1 and L2 participants)	118
Table 4.8. IELTS Academic Reading marking scheme	119
Table 4.9. Results of test takers' scores on the IELTS RCM and their IELTS band scores	120
Table 4.10. Some codes emerging from first cycle of coding	123
Table 5.1. Textual features of the "Raising the Mary Rose" text	128
Table 5.2. Lexical profile of the "Mary Rose text (IELTS RCM).	128
Table 5.3. The True, False, Not given Task	129
Table 5. 4. Summary of the True, False, Not given Task features	131
Table 5.5. The Matching Features Task	131
Table 5.6. Summary of the main features of the Matching Features Task	132
Table 5.7. Summary of the Diagram Completion Task features	135
Table 5.8. Words and phrases with the same reference used in the passage two	136
Table 5.9. Textual features of the "Easter Island" text	137
Table 5.10. Lexical profile of the "Easter Island" text (IELTS RCM)	137
Table 5.11. List of headings for the Matching Headings Task	138
Table 5.12. Summary of the Matching Headings Task features	140
Table 5.13. The Summary Completion Task 1	141
Table 5.14. Relevant information in the text	141
Table 5.15. Summary of the Summary Completion Task features	142
Table 5.16. The Multiple Choice (two answers) Task	143
Table 5.17. Summary of the Multiple Choice (two answers) Task features	143
Table 5.18. Textual features of the "Easter Island" text	145
Table 5.19. Lexical profile of the "Neuroaesthetics" text (IELTS RCM)	146
Table 5.20. The Multiple-Choice Task	147
Table 5.21. The relevant information that provides the answer to Multiple Choice items	148

Table 5.22. Summary of the Multiple-Choice Task features	148
Table 5.23. The Summary Completion Task 2	149
Table 5. 24. Relevant information for the Summary Completion Task 2	149
Table 5.25. Summary of the features for the Summary Completion Task (2)	150
Table 5.26. The Yes, No, Not give Task and the relevant information in the text	151
Table 5.27. Summary of the Yes, No, no given Task features	152
Table 5.28. Lexical profile of academic text “book chapter”	153
Table 5.29. Lexical profile of the sample academic text “Linguistics Article”	153
Table 5.30. Summary of text features for the IELTS RCM text and the two sample academic texts	154
Table 5.31. Lexical profile of the main sample of IELTS RCM and the two academic sources	156
Table 5.32. Summary of answer level, skills measured, and type of comprehension for IELTS RCM test tasks	158
Table 5.33. Summary of comprehension level of the IELTS RCM test tasks	159
Table 5.34. Reading dimensions of the IELTS RCM test tasks	161
Table 6.1. Results of recurring reading activities in test performance of different test takers	211
Table 6. 2. Results of crosstab analysis of searching theme across the three groups of test takers	218
Table 6.3. Summary of the frequency of searching theme in different groups of test takers	221
Table 6.4. Results of chi-square goodness of fit for the mean frequency of the search category	222
Table 6.5. Degree of association between test tasks and search category	225
Table 6.6. Association of the search processes used with the test tasks	225
Table 6.7. Frequency and percentage of “answering processes” across groups	230
Table 6.8. Summary of frequency of in answering theme by different groups of test takers	231
Table 6.9. Result of chi-square goodness of fit for the mean frequency of the searching theme	231
Table 6.10. Scale for classification of degree of association between test tasks and answering theme	232
Table 6.11. Association of processes and strategies used in answering theme with test tasks	233
Table 6.12. Crosstab of strategies used across groups.	236
Table 6.13. Summary of frequency of strategies used across groups	239
Table 6.14. Result of Chi-square goodness of fit for strategies used across test takers	240
Table 6.15. Association of the strategies used with test tasks	241
Table 6.16. Strategies used in different phases of test performance	243
Table 6.17. Calculation of mean frequency of (B&FM) per task and per item	245
Table 6.18. Frequency of (B&FM) for different test tasks and test takers	245

Table 6.19. Results of the Chi Square significance tests of B&FM across proficiency groups	247
Table 6.20. Rank order of (B&FM) for different test tasks and test takers	247
Table 6.21. Time spent on the test tasks	248
Table 6.22. Time bands for different test tasks across groups of test takers	249
Table 6.23. Item difficulty levels	251
Table 6. 24. Task difficulty for the three groups of test takers based on test scores	251
Table 6.25. Difficulty of test tasks for each group of test takers	252
Table 6.26. Self-declared task difficulty	253
Table 6.27. List of new words in the IELTS RCM for each group of test takers	254
Table 7.1. Results of the experts' judgements for the reading level	258
Table 7.2. Results of type of knowledge needed to do the tasks	260
Table 7.3. Results of experts' judgements on type of comprehension	262
Table 7.4. Results of Chi Square for types of comprehension	263
Table 7.5. Results of experts' judgements of type of reading (careful-expeditious)	264
Table 7.6. Results of text difficulty judgement	266
Table 7.7. Results of experts' judgements of the task difficulty	267
Table 7.8. Results of experts' judgements of the level of processing	269
Table 7.9. Experts' Assessment of time needed for test takers' task performance	271
Table 7.10. Experts' judgments of skills measured by RCM IELTS	273
Table 7.11. Summary of experts' judgements for different classifications	276

List of Figures

<i>Figure 2.1.</i> Schematic outline of Chapter Two	13
<i>Figure 2.2.</i> Cognitive model of construct validity (adapted from Gorin, 2006)	25
<i>Figure 2.3.</i> Quantitative versus qualitative approaches to test validation	32
<i>Figure 2.4.</i> Construct validity sources of evidence	42
<i>Figure 3.1.</i> Theoretical and empirical background to reading construct	44
<i>Figure 3.2.</i> Cognitive processing model of reading comprehension (Khalifa & Weir, 2009, p.43)	45
<i>Figure 3.3.</i> Validation model of IELTS RCM (Adapted from Khalifa and Weir, 2009, and Gorin, 2006) used in the present study	62
<i>Figure 4.1.</i> Data collection and data analysis procedures	109
<i>Figure 4.2.</i> Coding procedures (Adopted from Saldaña, 2009)	113
<i>Figure 4.3.</i> Phases of data analysis Data	116
<i>Figure 5.1.</i> The <i>Diagram Completion Task</i>	133
<i>Figure 6.1.</i> Schematic representation of data analysis framework moving from data to theory: The inter-relationship of processes, strategies, categories, themes and construct	164
<i>Figure 6.2.</i> Test performance in the <i>True, False, Not given Task</i>	166
<i>Figure 6.3.</i> Test performance in the <i>Matching Features Task</i> : Pattern one	170
<i>Figure 6.4.</i> Test performance in the <i>Matching Features Task</i> : Pattern two	171
<i>Figure 6.5.</i> Test performance in the <i>Matching Features Task</i> : Pattern three	171
<i>Figure 6.6.</i> Pattern of test performance for the <i>Diagram Completion Task</i>	174
<i>Figure 6.7.</i> Pattern of test performance for the <i>Summary Completion Task 1</i>	185
<i>Figure 6.8.</i> Patterns of test performance in The <i>Summary Completion Task 2</i>	192
<i>Figure 6.9.</i> Patterns of test performance in the <i>Summary Completion Task 2</i>	198
<i>Figure 6.10.</i> Model of IELTS RCM test performance of L1 test takers	207
<i>Figure 6.11.</i> Model of IELTS RCM test performance of L2 test takers	208
<i>Figure 6.12.</i> The main themes emerging from test takers' IELTS RCM test performance	209
<i>Figure 6.13.</i> Some processes and strategies used in "Reading Theme"	211
<i>Figure 6.14.</i> Processes and strategies used by the L1 test takers in reading theme	213
<i>Figure 6.15.</i> Processes and strategies used by the L2 test takers in reading theme	214

<i>Figure 6.16.</i> Processes and strategies used by the more successful L2 test takers in reading theme	215
<i>Figure 6.17.</i> The main processes and strategies used in “Searching Theme”	218
<i>Figure 6.18.</i> Patterns of searching the relevant information	227
<i>Figure 6.19.</i> Some processes used in “Answering Theme”	229
<i>Figure 6.20.</i> Multi-modal triangulated approach to task difficulty	256
Figure. 7. 1. Experts’ judgements of task difficulty of the RCM IELTS test tasks	267
Figure. 7.2. Experts’ judgements of level of processing for IELTS RCM test tasks	270
<i>Figure 8.1.</i> Schematic representation of the research questions	282

List of Appendices

Appendix A: Test takers' invitation email

Appendix B: Testing Experts' invitation email

Appendix C: Consent form

Appendix D: Proficiency Cloze Test

Appendix E: Reading Essay

Appendix F: Summary Task

Appendix G: Sample RCM IELTS test

Appendix H: Test Performance Observation Scheme

Appendix I: Construct of the RCM IELTS Questionnaire

Appendix J: Writing scale for summary tasks

Appendix K: Definitions, examples, and distributions of the processes and strategies used in Searching Theme

Appendix L: Definitions, examples, and distribution of the processes and strategies used in Answering Theme

Appendix M: Definitions and examples of the Metacognitive Strategies

List of Abbreviations

B&FM: Back and forth movements between the text and the test item

EFL: English as a Foreign Language

IELTS: International English Language Testing System

L1: First Language

L2: Second Language

RCM: Reading Comprehension Module

SKPSs: skills, knowledge sources, processes, and strategies

TPOS: Test Performance Observation Scheme

Glossary

In the context of this study a number of several terms were repeatedly used in different chapters and sections. To help reader for quick reference to these terms, this glossary presents operational definition of this term. Elaborate and detailed definition of the terms is intentionally avoided. The definitions are kept simple for quick referencing. These definitions are consistent with the relevant literature in the field.

Academic reading: reading as integrated with other academic activities such as reading to ask, reading to write, reading to present, reading to learn, and reading to research.

Answering theme: a theme is the synthesized or consolidated outcome of coding analysis of qualitative data (from code, to category, to theme). In the present study answering is one of three themes identified in the data, along with reading and searching. Answering engages the reader's processes and strategies in order to find relevant information, for example, to get the answer to a test question.

Careful reading: Reading all or parts of the text to get complete, clear, precise meaning of the main points and relevant details of the text in order to construct text representation (Khalifa & Weir, 2009).

Category: a set of inter-related codes (processes and strategies). A category can consist of some sub-categories.

Cognitive processes: cognitive processes refer to a set of tacit automatic actions used in reading comprehension. Reading theories agree on a set of specific cognitive processes that are commonly used in reading. They include word recognition, syntactic parsing, forming a semantic proposition, developing a text representation, inferencing, and creating mental model, etc.

Construct: the knowledge, skill, or ability that a test is designed to measure. For example, reading comprehension is a complex construct. Test developers specify the construct by drawing on theory and research. The specification of the construct leads to developments of items or tasks that represent it. Construct validity is "the degree to which a test reflects the essential aspect of the theory on which the test is based" (Mousavi, 2009, p.138).

Construction-Integration Model: a *bottom-up* model of reading that suggests there are three levels of text representation: 1) surface level (words and letters), 2) propositional level (making meaning from the words), and 3) the situation level (a mental representation connected to prior experience and expectations, Kinstch, 1988; Kintsch, & van Dijk, 1978).

Constructivist view: considers reading to be a process of social construction; meaning derived from written text in the practice of reading is situated, historically and culturally, as the reader interacts with a text to construct its meaning.

Construct validity: Messick's (1989) definition of construct validity is adopted for this dissertation. He defines construct validity as the degree to which the decisions made on the basis of test scores are meaningful, appropriate, and useful.

Evaluative meaning: meaning that requires readers/test takers to move beyond text, e.g., critically look at the ideas in the text, analyze them, identify the position of the author, justify what they think about it, argue for a particular viewpoint.

Expeditious reading: quick, selective, strategic, and efficient reading of parts of the text to access relevant information. Skimming and scanning are examples of expeditious reading. (Khalif & Weir, 2009; Urquhart & Weir, 1998).

Global comprehension: the focus on text macro-propositions by incorporating topical knowledge and text information.

Inferential meaning: meaning that is not explicitly expressed in the text and requires readers/test takers to draw on their prior background knowledge or identify relevant clues in a text in order to understand. It is reading between the lines to understand implied meaning.

Knowledge (sources): refers to different types of linguistic knowledge such as vocabulary knowledge, grammatical knowledge, textual knowledge and non-linguistic knowledge (topical knowledge).

Level of processing: the level at which a text, or test item/task is processed. It may involve processing a few sentences, a paragraph, several paragraphs, or the whole text.

Lexical density: the total number of the content words (noun, verb, adverb, and adjective) divided by the total number of words in the text.

Literal meaning: surface level meaning that is explicitly expressed in words and sentences and requires no inferencing.

Local comprehension: the focus on micro-propositions. In a test, items that require a quick search of some section of the text to locate specific information or details.

Metacognitive strategies: strategies that are used to plan, execute, monitor, and modify reading processes or test performance. Strategies are capacities that help test takers/readers use their language competence in the context of reading a text and doing the test items

Nomological network: a representation of the constructs (concepts) of interest in a test; their observable manifestations, and the interrelationships among and between these; developed from multiple theoretical and empirical sources.

Process: a tacit, automatic operation such as word recognition and syntactic parsing, that is engaged to accomplish reading or do a test task.

Process-oriented model of validity: a model of construct test validation that operationalizes construct validity as alignment between Skills, Knowledge, Processes, and Strategies (SKPSs) of the intended and measured construct (Gorin, 2006)

Propositional density: the number of propositions in a text which is calculated by dividing the number of propositions in a given text by the total number of words.

Protocol coding: applying pre-established codes, classifications, or categories from a previously developed system to analyse qualitative data (Saldaña, 2009)

Reading Theme: a theme is the synthesized or consolidated outcome of coding analysis of qualitative data (from code, to category, to theme). In the present study, reading is one of three themes identified in the data, along with searching, and answering.

Scanning: searching for specific words, phrases, dates, etc. to retrieve relevant information for achieving a reading purpose or answering a specific test item.

Schema Theory: a system that describe how knowledge is stored and used and suggests that all knowledge is organized into units which store information.

Searching Theme: a theme is the synthesized or consolidated outcome of coding analysis of qualitative data (from code, to category, to theme). In the present study searching is one of three themes identified in the data, along with reading and answering. Searching involves processes and strategies used to locate the relevant information in text through careful reading, skimming, and/or scanning.

Search reading: processing the text in order to locate information relevant or necessary to achieving a reading purpose or answering specific test questions.

Situation Model: the mental representation (model) of the situation (e.g., actions, events) presented in the text

SKPSs: Skills, Knowledge sources, Processes, and Strategies used in test performance (Gorin, 2006).

Skimming: reading rapidly to get a general overview of the reading material.

Strategy: deliberate goal-directed operations or actions; used to accomplish reading and responding to a test task.

Strategic Competence: The ability to use language interactively in different contexts and situations.

Test management: a *test-wiseness* strategy that deals with managing test performance activities such as time management; understanding the relative value of an item or task in relation to points, etc.

Test-wiseness: a test taker's capacity to use characteristics and formats of the text and/ or test to guess the correct answer.

Test content analysis: analysis of linguistic and textual features of the texts and test items in the Reading Comprehension Module of IELTS based on classifications such as type of reading, level of processing, difficulty, etc.

Text type: the main structure of a particular text or one of its parts according to its dominant properties.

Validation: an ongoing process of examining (verifying/falsifying) a particular interpretation of a test score by building a logical case and providing supporting evidence for a validity argument.

CHAPTER ONE: INTRODUCTION

1.1. Statement of the Problem

The main purpose of the current research is to study the construct validity of the Reading Comprehension Module (RCM) of the International English Language Testing System (IELTS), administered globally as a high stakes English language proficiency test. Put simply, a *construct* is, “The trait (traits) or underlying ability that we intend to measure through assessment” (Cheng & Fox, 2017, p. 224). *Validity* (Messick, 1989), is an evaluative judgement of the degree to which a test measures what it was intended to measure. (Please see the glossary at the beginning of this dissertation for additional information).

The study examines cognitive evidence (i.e., skills, knowledge sources, processes, and strategies used during test performance, Gorin, 2006) from test content, test taker retrospective verbal accounts of skills, processes, and strategies applied while taking the RCM, and testing experts’ accounts and judgements of the test. In the context of the unified view of construct validity (Messick, 1989, 1998), these diverse sources of construct validity evidence can provide detailed information about the construct of the test and help policy makers and test users make more informed decisions about the meaningfulness, appropriacy and usefulness of the test as a measure of academic reading.

In the last two decades, the economic, social, and educational changes in contexts of globalization have led to internationalization of higher education, on the one hand, and an increasingly mobile international workforce migration, on the other. The number of people who needs to study or work outside their home country have continually increased. Parallel with these developments, demand for high stakes English language proficiency tests such as Canadian Academic English Language (CAEL), Test of English as a Foreign Language (TOEFL), and IELTS has also increased. These tests have served as gate keepers and helped countries, businesses, and academic institutes to screen and choose the most qualified candidates; and, at the same time, have also helped test takers fulfill their dreams. Such high stakes tests exert high impact on the lives of stakeholders including test takers, academic departments, content instructors, language teachers, parents, and other test users. Arguably, then, these tests deserve more research and understanding. With increased demand for language assessment, “there is greater demand for language testers to be accountable to stakeholders” (Bachman 2013, p.1), and

testing organizations and corporations need to provide further evidence and arguments for the claims they make about their tests. Furthermore, test assessment standards, ethical considerations, and professionalism (Bachman and Palmer, 2010) all obligate more accountable language assessment practices especially for high stakes language proficiency testing. This necessitates further research into the construct validity of these tests.

With regard to the high stakes test that is the focus of this dissertation, among different modules of the test, the IELTS RCM is the least researched module and there exist only a few studies that have delved into its construct validity (e.g., Bax, 2013; Green & Hawkey, 2012; Weir, Hawkey, Green, Ünalı, & Devi, 2012; Moore, Morton, & Price, 2012; Weir, Hawkey, Green, & Devi, 2009). These few studies, taken together, have highlighted only a few features related to the construct validity of the RCM such as level of engagement (local-global) of the test items, types of reading (careful-expeditious) involved, and type of comprehension (literal-evaluative), but have not examined the construct validity of the whole test or each of its nine test tasks. The approaches adopted by these studies were singular, relied on one single source of evidence, and did not look at the multiple sources of validity evidence. Findings of these studies provide some insightful, detailed information about specific dimensions of the test, but do not directly address the wide range of cognitive activities including the knowledge sources, skills, processes, and strategies elicited by each test task. In addition, other sources of validity evidence such as test content and perceptions and judgements of testing experts and professionals have been mostly neglected. Even the IELTS RCM validity statements that are officially cited (www.IELTS.org, 2019) are very general, brief, and limited to just naming some reading component skills. They do not go beyond some skill-based statements in describing the construct, such as reading for the gist, reading for main ideas, reading for details, and skimming, which suggest the test not much different from other typical reading comprehension tests. Construct definition and task explanations lack process-based statements, which raises questions about what the test actually measures. The texts used are also claimed to be “appropriate for people entering university” (ibid). Most of these claims remain unsubstantiated and the existing literature on the construct validity of the test does not address these gaps, suggesting the need for further research.

Furthermore, key to the accountability standard in language assessment is providing validity justification for test use and interpretation, which requires a theoretical rationale and

empirical evidence to justify use of a particular assessment (e.g., Messick, 1989). Empirical evidence collected from multiple sources of evidence can reveal detailed information about the test construct and enrich the test validity argument, which helps test users justify its appropriate, meaningful, and useful interpretation (Messick, 1988). As suggested by Messick's unified validity theory (1988, 1998) construct validity is an all-encompassing validity concept which can accommodate different sources of validity evidence. Quantitative experimental evidence such as multi-traits/multi-methods and qualitative evidence such as test takers' perceptions, and experience (Fox and Cheng, 2007, 2016), experts' accounts (Downing & Haladyna, 1995, 1996; Fulcher, 2012; Pill & McNamara, 2016) and test content evidence (Green & Hawkey, 2012) can address different dimensions of the test construct and enrich the validity argument of the test. Introspective/retrospective verbal reports of the processes and strategies used by the test takers, in particular, can shed light on the internal dynamics of test processes. They can highlight the processes and strategies (Cohen, 2006, Gorin, 2006) that result from text-readers interactions on a test. Furthermore, any validity argument for a language test depends on the correspondence between the processes and strategies used during test performance and the processes and strategies of the target language use domain (TLU) (Bachman & Palmer, 2010). In the absence of such correspondence, the test may fall short of justifying the appropriacy of the decisions made on the basis of test scores. In sum, examining construct validity of a language test in the context of these diverse dimensions of evidence can contribute to the construct validity argument of the test.

Due to the high impact of the IELTS test on different stakeholders especially student test takers, and the paucity of empirical evidence that demonstrates the specific processes and strategies tapped by different IELTS RCM test asks, the current study aims at addressing this void and providing further cognitive evidence for the construct of the IELTS RCM. Researching the cognitive processes in terms of the skills, knowledge sources, processes and strategies (SKSPs) involved in the IELTS RCM in three independent yet inter-related sources of evidence can provide detailed information about the construct of the test. Test content, test takers' processes and strategies, and experts' accounts and judgements provide a more comprehensive and detailed picture of the test construct.

1.2. Rationale of the Study

As discussed above, the main purpose of the current study was examining the construct validity of the IELTS RCM by examining three diverse but complementary sources of validity evidence. The rationale for this study was manifold. First, the test has had high impact on the lives of many stakeholders especially test takers who were learners of English as a foreign language (EFL) or learners of English as a second language (ESL) test takers. Second, I had firsthand experience and encounters as an English for Academic Purposes (EAP) instructor with EFL learners who were preparing to take General or Academic IELTS. They were struggling to pass the minimum requirement of the test and get an admission from their desired university. Spending months and years preparing for the test, many of them felt desperate not knowing what more they needed to do to pass the test and get a minimum of 6 on the test. Third, the relevant literature and findings of empirical studies on the validity of high stakes tests in general, and the IELTS RCM in particular, had all the indications of a paucity of research and called for further research. These issues and dimensions motivated me to conduct this research and address some of the questions and gaps related to the construct validity of the IELTS RCM. Some more specific reasons that encouraged me to conduct this study included the followings.

First, the success stories of high stakes tests such as IELTS and TOFEL are usually highlighted and promoted by their associated organizations, but there are many untold failure stories as well. EFL/ESL students who take the test spend huge sums of money and months and perhaps years of their lives preparing for such tests. For them and their families, there is a lot at stake. They may take the tests twice or more before they can meet the minimum requirements. When they fail, they lose faith in their abilities and lose their self-confidence. All their investment for further education abroad depends on the results of the test, which impacts them personally and professionally. Results also impact other stakeholders and test users such as families, admission departments and university content teachers, who teach the students who have made it, and get admission for university. Testing organizations and bodies are socially responsible and accountable for the consequences of the test (Shohamy, 2001). Therefore, presenting validity evidence for high stakes tests such as IELTS, which significantly impacts test users, is the least testing organizations and testing researchers need to present to the public. Validity evidence, however, needs research and scrutiny on an ongoing basis (Cronbach, 1988).

Second, as an EAP instructor and language testing researcher and being a second language learner myself, I could see firsthand, how families and test takers were frustrated and distressed going through preparation and registration procedures, which did not always end happily for them. At that point, they had to start over and take the test at least once more. They struggled to pay for the test fees, which were twice as much as the average monthly income of a household. They spent months of their time and sums of money attending IELTS preparation courses to learn different tricks and techniques that the instructors advised them to use for improving their scores on the test. Most of them worked hard to learn these techniques, hoping that applying them would increase their chances to succeed in the test. I could see many of them chose to focus on *test wiseness* strategies (i.e., using characteristics and formats of the test to guess the correct answer) instead of investing in improving their English language proficiency. I heard a few success stories of participants who could improve their score from 5.5 or 6 to 6.5 by attending IELTS preparation courses, but I was confident these tricks and techniques could hardly help them. I frequently heard from the test candidates, talking about the IELTS Reading Module, as a module which does not really require much reading skills, rather, they often described applying test taking strategies and techniques in responding to the tasks. I wondered if and how this could be true and assumed that test wiseness strategies work for those who enjoy good reading skills and help them to improve their scores on the test. I assumed the use of strategies did not make up for low reading skills and could not help those who struggled reading and comprehending a text within the allowed time limit. I assumed test wiseness strategies are additional strategies that can only support reading comprehension skills and cannot substitute it. I also wondered if test takers with different language proficiency rely on their reading skills or if they use their test wiseness strategies and the techniques they learn in the preparation courses.

I thought that gaps in research relating to my own assumptions and questions about the test might best be addressed by diverse groups of test takers who could provide a broad perspective of the actual experiences and processes of test performance and reveal more evidence on dimensions of the test construct that are not brought to light by quantitative approaches to validation. I thought the differences and similarities of the processes and strategies used by test takers at different levels of language proficiency could provide a broader picture in mapping the core components of the test construct. I assumed diverse groups of test takers could provide rich sources of evidence to address these questions. With these questions and

uncertainties about the reading module in mind, I could not stop thinking about huge number of the students who had failed to meet the minimum requirements after months of preparation.

Third, as claimed in the test webpage (www.IELTS.org, 2019), the IELTS RCM measures a wide range of reading skills, including reading for gist, reading for main ideas, reading for detail, skimming, understanding logical argument and recognising writers' opinions, attitudes and purpose. However, these reading skills are claimed by almost all reading tests and they do not provide much information about the actual reading construct operationalized by the RCM. I assumed that understanding what the test actually measures may not be possible without more evidence to examine what specific reading skills the test tasks measure and whether these skills really measure academic reading skills. Reviewing the literature, I could see the huge gap in this area of research and could recognize the need for further research to address these gaps. To my surprise, in spite of the impact of IELTS on the lives of so many test takers and other stakeholders, as discussed in the previous section, there were only a few studies on the construct validity of the IELTS RCM. This type of classroom experience is evidence of washback from the test on teaching and learning (Cheng, 2002, 2005; Cheng & Curtis, 2004; Cheng, Sun, & Ma, 2015; Cheng, Watanabe, & Curtis, 2004).

Fourth, I reviewed the relevant literature with a focus on studies that have addressed construct validity of the IELTS RCM. Based on the review of the literature (e.g., Bax, 2013; Green & Hawkey, 2012; Moore, Morton, & Price, 2012; Weir, Hawkey, Green, Unaldi, & Devi, 2009; Weir, Hawkey, Green, & Devi, 2009; Weir & Urquhart, 1998; see Chapter Three for a review of this and other relevant literature), the following conclusions were drawn. Although, most of the studies adopted Khalifa and Weir's (2008, 2009) model of reading comprehension and were designed and conducted within the confinements of the mode, the main focus of the studies was limited to the type of reading involved (careful-expeditious) and level of processing (local-global), as conceptualized in the model (see Chapter Three or the glossary for further information on these distinctions). The studies did not address the full spectrum of cognitive processes (i.e., word recognition, syntactic parsing, formation of semantic propositions, inferencing, developing text representation, creating a mental model, Khalifa and Weir, 2009), strategies (deliberate operations used to read and comprehend a text or do a test task), and knowledge stores (i.e., linguistic knowledge such as vocabulary knowledge, grammar, textual knowledge and background

knowledge involved in test performance). In addition, the research participants recruited were at two levels of language proficiency, higher level and lower levels. Further, the studies considered the IELTS RCM as one whole task, did not address each test task separately, and made no distinction among the nine test tasks and what each type of test task measures. Test takers' accounts of their test performance and reading processes were mainly collected through questionnaires, with the possibility of adding some comments when necessary. None had used verbal reports of the test takers. Bax (2013), for example, included retrospective verbal reports as complementary data to eye-tracking data.

Administration of the tests in these research studies also differed from the actual administration used in most 'live' tests, namely pencil and paper test administration. In another study conducted by Weir et al., (2009) participants took the test and after answering each item they had to stop and record their accounts of the item by choosing from the list of options in the questionnaire. Or they answered the test on computer screen which can alter the construct of test (Bax, 2013). These practices of the test did not reflect IELTS administration protocols (IELTS.org), and thus lacked performance authenticity. In some studies, test takers took only one text and items, not the full sample of the test. In addition, in a few studies, two different sources of validity evidence were used. For example, in one study (Moore, Morton, & Price, 2012) content drawn from instructors' accounts of academic reading in real life was included as a source of evidence for test validity. In sum, the studies as a whole did not adhere to the skill model of reading, rather they conceptualized and examined the test construct in terms of a more functional perspective and notions such as type of reading involved, namely, expeditious-careful (Weir et al., 2009), level of processing, local-global, (Urquhart & Weir, 1998), and type of engagement, literal-inferential (Moore, Morton, & Price, 2009). Further, only Weir et al.'s (2009) study included L1 test takers for data collection. Most participants were a sample of EFL or ESL students. In my view, including L1 speakers could provide more information and insight about the test construct.

In addition to my personal rationale, the available literature broadened my perspective for the examination of the construct validity of the test by including other sources of evidence such as analysis of test content, the test takers' verbal reports of the processes, and accounts and judgements of testing experts. I argue that these sources of evidence could better inform our understanding of the construct of the IELTS RCM. These gaps and limitations helped formulation of the research questions and research design of the present study. Based on the paucity of research

in the validity of high stakes language proficiency tests and more specifically, the dearth of research on the construct validity of the IELTS RCM, and the limitations of the reviewed studies, the current study examined the construct validity of the IELTS RCM by collecting validity evidence from three diverse sources of evidence, including: 1) text content analysis, 2) test takers' accounts and experience of their test taking processes, and 3) testing experts' account and judgements of the test tasks and items.

1.3. Research Questions

The overarching focus of the study was an examination of the cognitive evidence for the construct validity of the reading section of the RCM IELTS as operationalized by its nine test tasks. However, the focus was broken down into more specific research questions. The following research questions guided the study;

1. What construct of reading comprehension in terms of (Knowledge, Skills, Processes, and Strategies, (KSPSs) is measured by the IELTS RCM tasks?
 - 1.1. What does content analysis of the IELTS RCM test tasks in terms of linguistic, textual, and topical features of the texts, the test tasks, and test items reveal about the construct of the test?
 - 1.2. Do adequately proficient test takers versus less-proficient test takers versus native test takers use different (KSPSs) in taking the IELTS RCM test tasks?
 - 1.3. What are the testing experts' (e. g., EAP teachers, language test developers, test preparation instructors) accounts and judgements of the SKSPs tapped by each type of task in the IELTS RCM?
 - 1.4. Is there evidence of congruence or incongruence among different sources of validity evidence (i.e., content analysis of the test, cognitive processes and strategies used by test takers, experts' accounts and judgements, and the existing literature) with regard to the IELTS RCM tasks?

Arguably, the cognitive perspective (Gorin, 2006) across these three sources of evidence (i.e., test content, test takers' accounts, and testing experts' accounts and judgements) allowed for a deeper understanding of the test construct. The study broadened the range of evidence for test validity by diversifying both the sources of evidence used and the sample of test takers who were selected from different language backgrounds (L1 and L2) and levels of language

proficiency. In my view, the test construct could be better understood in the context of the test takers' experience of the test and what they actually did during their test performance. Examining the validity of each type of test task and providing details about the processes and strategies used during test performance should be helpful for test developers and item writers in illuminating how test tasks and items actually function. This would allow them to compare evidence from testing processes with their intended purpose, to evaluate item quality, and thereby to improve their test development. Results could also help them decide if they should make some modifications in test format or test task. Without process details, no possible change or modification is possible. As the test collects data from diverse sample of test takers performing on nine different tests tasks each consisting of 4-6 items, it can provide an array of skills, knowledge sources, processes, and strategies that are involved in test performance. The study also proposes to address the full range of cognitive activities, skills, knowledge stores, cognitive and metacognitive processes and strategies used during test taking processes, not focusing on one specific cognitive aspect of test performance. In light of the rich pool of cognitive processes of the whole test derived from the verbal reports of the test takers, cognitive dimension of the test construct can be more comprehensively understood. Results of the experts' judgements address another dimension of test construct, i.e., perception of testing professionals, and provide further evidence to the test construct.

Results of the study have also several implications and applications for different stakeholders. First, it can examine the current theories of reading comprehension construct by providing evidence from multiple sources. Second, test taking processes and strategies used by the test takers can provide close-up details of the test processes and show how or if the test is relevant to the construct of the academic reading comprehension which can help decision makers and test users interpreting test result. Third, results can enrich the literature on the construct of academic reading comprehension tests by exemplifying the way academic reading comprehension is conceptualized and operationalized by the IELTS RCM. This can help decision makers and other test users make a more informed and evidence-based understanding of the test construct and its use. Fourth, methodologically, the study can exemplify how different sources of construct validity evidence can be integrated in one study for examination of construct validity and set an example for further studies. Fifth, the study construct validity evidence from two of the least studied validity sources, i.e., test content and experts' accounts and judgements.

Results of these two sources can provide information that is usually less attended to in validation studies. Results of content analysis, in particular, can help decision makers decide if the test they choose or develop for a given context is relevant and appropriate for their testing context. In brief, in light of the findings reading instructors, IELTS RCM instructor, test developers, item writers, and decision makers can benefit from details of the findings and improve their use of the test. The findings provide a pool of information and can help different test users make more informed decision about their use of the test

1.4. Organization of the Dissertation

This dissertation was organized in nine chapters. Chapter One has presented some background to the study, rationale of the study, and the main research questions related to the construct validity of the IELTS RCM. Chapter Two is reserved for discussion of the theoretical framework of the study which consists of the unified construct validity theory (Messick, 1989) and the process-oriented model of construct validity (Gorin, 2006). Chapter Three reviews the cognitive model of reading comprehension (Khalifa and Weir, 2009) which defined the reading construct considered in this study, and the empirical studies related to the factors and variables that influence reading comprehension processes and test performance. This includes theories and research findings related to the influence of: 1) cognitive processes and metacognitive strategy use in reading comprehension, 2) readers' variables such as ~~their~~ language background and level of language proficiency, and 3) the influence of text features and task features of the test on reading processes and reading construct. Chapter Four presents an overview of the methodology employed in the study, including information on participants, instruments, data collection procedures, coding cycles, and data analysis procedures. Next, results of the study are reported in three separate chapters. Chapter Five is devoted to the results of the content analysis of the IELTS RCM. Chapter Six reports the findings that emerged from test takers' account of the SKSPs used in their test performance, and Chapter Seven presents results of the testing experts' accounts and judgements of the test tasks. Chapter Eight discusses the results overall, integrating each source of evidence, in response to the research questions. The final chapter, Chapter Nine, concludes by summarizing the findings, discussing implications, and suggesting directions for further research.

CHAPTER TWO: THEORETICAL FRAMEWORK

2.1. Introduction

The current study sought to examine cognitive evidence for the construct validity of the IELTS RCM. Chapter One presented the core research problem, rationale of the study, the research questions, and potential contribution of the findings to the scholarly work in the field. As shown in Figure 2.1., the present chapter is devoted to the discussion of Messick's (1989) unified construct validity theory which was adopted as the theoretical framework of the study. Messick's unified construct validity theory and process-oriented model of construct validity (Gorin, 2006) which is a working model that fits into unified theory were adopted as core theories and models that conceptualized construct validity definition of the study. In the first part, a brief history of validity theories is presented which includes unified validity theory, further developments related to validity theory such as validity argument and test usefulness. It also includes discussion of threats to construct validity, test validation, and sources of validity evidence such as content analysis, verbal reports, and experts' judgements which are used as main validity sources in the current study. The next chapter will discuss Khalifa and Weir's (2009) reading comprehension model which served as a map of the construct of reading comprehension which the IELTS RCM operationalizes.

In any research study, the theoretical framework adopted by the researcher plays a crucial role in the conceptualization of the research problem and the methodology adopted, "By placing the research into a more general conceptual framework or theoretical orientation, a rationale is provided for the research questions. Essentially the intellectual or scholarly perspective in which the problem is embedded is described" (McMillan & Schumacher, 2001, p. 74). The significance of the theoretical framework lies in its contribution to the enrichment and support of the purpose of research by relating it to the body of theoretical and empirical knowledge discussed in the field and by framing research variables within existing theories. It also helps researchers choose an appropriate design and discuss their findings in relation to the body of related knowledge, theories, and research findings. A sound theoretical framework of validity that describes components, assumptions, procedures, and limitations of the theory can inform researchers in adopting appropriate research designs and procedures in test validation. Hence, the unified construct validity theory and the process-oriented validation model adopted in this study

informed different dimensions of the current research in terms of sources of validity evidence, research focus, research design, method of data collection, and discussion of the findings.

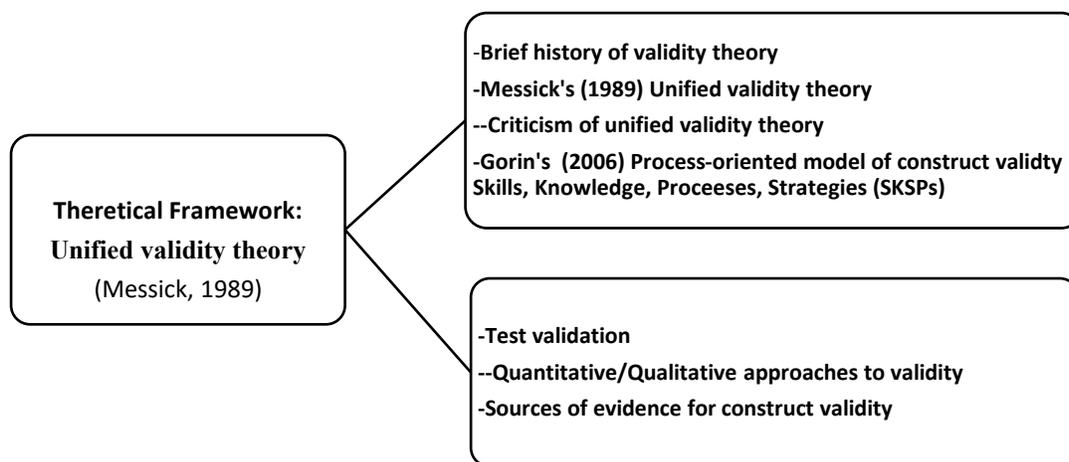


Figure. 2.1. Schematic outline of Chapter Two

2.2. Brief History of Validity Theory

In its short history, validity theory has undergone dramatic changes and developments over time. These changes can be best observed in the conceptual definition of the term. (Cronbach, 1988, 1989; Goodwin, 1997, 2002; Kane, 2002, 2004, 2011; Messick, 1989). In the 1940s, the early conceptualization defined validity as measuring what is supposed to be measured and assumed validity as a static characteristic of the test itself that could be assessed by the correlation of the test with some other external measures (Goodwin & Leech, 2003). Cronbach & Meehl (1955) introduced their view of construct validity as a “network of associations or propositions in which it occurs” (p. 300) and argued for predictability of the propositions that can be examined with empirical evidence collected through many types of evidence including content validity, interitem correlations, intertest correlations, test-"criterion" correlations,

stability over time, and stability under experimental intervention. Their evidence-based conceptualization of construct validity could allow test developers redefine the test construct in light of the empirical evidence or search for a fresh body of evidence to support the test construct. They emphasized that

construct validity cannot generally be expressed in the form of a single simple coefficient.... The integration of diverse data into a proper interpretation cannot be an entirely quantitative process. (p. 300)

Their conceptualization of construct validity moved beyond pure description of relations observed among variables to highly theoretical constructs that can involve hypothesized entities and processes. Cronbach and Meehl (1955) turned test validation into a set of serious scientific procedures and methodologies that can enjoy scientific legitimacy and logical rationale. Following Cronbach and Meehl (1955), in the first edition of *Standards and Psychological Testing* which was published in 1966, the concept of validity shifted to test use and was defined as the extent to which the test produces information that is useful to the specific use of the test and suggested three types of validity evidence: criterion-referenced validity, content validity and construct validity (Brown, 2000; Brualdi, 1999). First, criterion-referenced validity examined external correlates as evidence of validity. The assumption was that measures of different skills produce low correlation coefficient and measures of the same skills produce high correlation coefficient. Second, content validity was defined in terms of the correspondence between the test content and the target content it is supposed to represent. Finally, construct validity was then defined as the experimental evidence that indicates a test is measuring the construct it claims to measure (Bachman & Palmer, 1996; Brown, 2000). Construct validity was then used as a post hoc experimental procedure to collect validity evidence. Based on the positivistic paradigm, which was adopted by psychometricians, a test should produce some evidence that indicate the examinee possesses a hypothetical trait (construct). As the trait could not be directly observed, empirical procedures of different types were employed to validate the hypothetical construct. However, the experimental evidence obtained still needed explanatory concepts and discussions.

The main shift in the conceptualization of validity which changed the face of language test validation occurred in the 1980s and 1990s. Cronbach (1980) and Messick (1989) were the pioneers in the shift. They associated validity with the inferences and decisions that are made on

the basis of test scores and defined validity as “the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores” (p. 40). Test validation was seen as “the process of accumulating evidence to support such inferences (AERA, APA, & NCME, 1985, p.40) where test developers collect data from different sources to explore or support for test validity. In this model of validity, the three traditional discrete types of validity discussed above were not seen as separate or discrete types of validity, rather as different facets of a single unified form of construct validity. Another main feature of the new conceptualization was the need for evidence to justify the social consequences of test use (Cronbach, 1988; Messick, 1989, 1994). Some of these early conceptualizations of types of validity evidence were later modified. For instance, content validity which was narrowly operationalized in terms of item correlations was modified to “a representative sample of the tasks from a well-defined target domain” (Bachman 1990, p. 310). The unified validity theory (Messick, 1989) is discussed in more details below.

2.2.1. Unified validity theory

Perhaps the most important development in validity theory was made by Messick who proposed unified construct theory (Kane, 2006, 2011; Messick, 1989) which has dominated conceptualization of validity as a superordinate concept encompassing all other traditional validity types. It has since influenced professional standards, testing legislation, and validation studies conducted by testing researchers. Messick (1989) defines construct validity as

...an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment (p. 41).

The theory subsumed and unified multiple sources of evidence under a generalized notion of construct validity as one single validity argument with multiple lines of evidence. Messick (1989) recognized the multiple sources of evidence as “complementary forms of evidence to be integrated into an overall judgment of construct validity” (p.37). Today, it is considered as a general approach to validity that seeks multiple sources of evidence, e.g., including, content evidence criterion-based evidence, reliability, and a wide range of other sources of evidence. These facets look at both the internal and the external aspects of construct validity. Content

validity looks at one of the internal components of validity by analysing the test content in terms of the linguistic and psychometric characteristics of the text and test items while criterion-related evidence makes use of some experimental, quasi-experimental, or correlational designs as part of the “old orthodoxy” to search for some of the external evidence for construct validity. Results of these studies can provide some convergent and discriminant correlations for/against the construct under study Campbell and Fiske (1959). Examining the test takers’ accounts of the test processes provides another window to look at the internal dynamics of test processes. Experts’ accounts, on the other hand, provide a perspective external to the test which can highlight some other features of the test.

One main development in Messick’s conceptualization of construct was the shift from validity as the characteristic of the test to validity as a property of its interpretation and use. Messick (1989) maintained that “what needs to be valid are the inferences made about score meaning, namely the score interpretation and its action implications for test use” (p.37). Based on his definition, test validation required collecting evidence from multiple sources of evidence to support the claims made by the test or other modes of assessment. The argument is that the inferences made on the basis of test scores cannot be valid without enough supporting evidence. Multiplicity of validity evidence provides different perspectives to look at different dimensions of test use and its interpretation.

Messick (1989) considered validity as a function of what test users do with the results. Examining test validity, then, required collecting different types of data to validate its use. For example, knowing what the test content taps into can help test users decide the context in which the test results can be used. In the context of the IELTS RCM, if the reading module claims to represent a measure of academic reading comprehension, there should be evidence that supports the match between characteristics of the texts, test task features, and other qualities of the test with academic reading tasks and activities. Furthermore, empirical evidence should ascertain the skills, processes, and strategies used in the test match those used in academic reading. Without such evidence, the test fails to support its claim as a measure of academic reading ability. Evidence should also demonstrate that those who do not meet the minimum requirement fail in successfully achieving academic reading.

Messick also discussed some threats to construct validity under two broad categories; construct-irrelevant variance and construct underrepresentation variance (Messick, 1989; Young, 2008). As pointed out by Markus and Borsboom (2013)

According to Messick's theory, all validity comes down to demonstrating that a test assesses as much as possible of what it should (minimize construct underrepresentation) and as little as possible of what it should not (minimize construct-irrelevant variance). All sources of invalidity, in this view, fall into one of these two exhaustive types. (p. 14)

As these two threats might be at work in any language test and construct validity of the test depends on the degree these two threats are controlled, it is important to consider them in discussing and researching validity. At this juncture, they are briefly discussed.

Construct irrelevant variance relates to all the undesirable variables and conditions that are not directly related to the test construct, but they influence test scores (Bakker, et al., 2008; Messick, 1989; Young, 2008) and contaminate the test construct. Hence, the observed test scores represent factors and variable that are irrelevant to the intended construct. Controlling for error due to the test, the testee, test administration, and scoring can reduce the size of the construct irrelevant variance and increase the size of true variance (Brown, 2000). Construct underrepresentation variance, as another threat to construct validity, refers to the degree test domain represents the target domain. when the test domain is narrowly defined and operationalized and fails to include the critical components of the construct, the test construct is assumed to be underrepresented. (Messick, 1989; Young, 2008). For some language tests, detailed description of the target domain and its components and careful planning of the test task can help avoid construct underrepresentation. Construct irrelevant variance and construct under-representation have implications for test developers to design, operationalize, and develop tests that fully represent the target construct and provide error-free instrument.

Consideration of construct irrelevant variance and construct under-representation is relevant to the current study. In the IELTS RCM several factors might threaten construct validity. Variation in the format of the test tasks format of the test tasks, test takers' unfamiliarity with the test tasks, ambiguity of the test task, text topics, and nature of the test tasks and what they measure can allow for construct irrelevant variance. There are still serious questions about the extent to which the IELTS RCM test tasks represent academic reading skills. Research has

demonstrated some of the gaps in the test (Weir, Hawkey, Green, Unaldi, & Devi, 2009). The type of reading (expeditious or careful reading, global versus local reading), type of comprehension (literal and inferential meaning), and cognitive processes and strategies used in the two domains need further research (Hawkey, 2006; Weir, et al. 2007; Rosenfeld, Oltman, & Sheppard, 2004; Weir, et al. 2009) (Weir, Hawkey, Green, & Devi, 2009; Weir, Hawkey, Green, Unaldi, & Devi (2009) Sarojani, 2011; Katalayi & Sivasubramaniam, 2013). Controlling for construct irrelevance is relevant to the current study because test content or task features of the IELTS RCM may include some construct irrelevance variance. Furthermore, it is quite possible the test task may measure a skill or knowledge which was not intended by the test or test tasks may underrepresent academic reading construct. For instance, if a reading test task does not define topical knowledge as part of the test construct and contrary to what is expected, test takers fail to answer the task due to lack of topical knowledge, then topical knowledge functions as construct irrelevant to the test construct and influences test score. Content analysis of the test at text and the task features levels on the one hand, and analysis of test takers' processes and performance, on the other, can reveal some potential variables related to construct irrelevant variance.

2.2.2. Validity argument

In the last two decades, further developments were made in validation theory. To address the more abstract and theoretical nature of unified validity theory, testing scholars tried to operationalize his theory in test validation studies and testing projects. Kane (2006) argued that validity is never assumed, rather, it is established through relating empirical evidence to the suggested score interpretations and meanings in a meaningful way (Kane, 2002, 2004, 2006, 2011). He suggested validation argument should aim at reaching a conclusion about the adequacy of test score interpretation and use and “provide conceptual tools for specifying the connection between limited samples of observations and proposed interpretations and uses” (Kane, 2006, p. 17).

A key component of the validity argument, he asserted, is the concept of “score interpretations” which systematically examines different types of inferences in score interpretation process including scoring inference, generalization inference, extrapolation inference, and implication inference which are then integrated into the interpretive argument to

form the core to the validity argument of the test. Argument-based validity has close affinity with unified validity theory which views validity as multi-faceted argument where test developers establish validity through collecting evidence from different sources. Downing (2003) argued that “Validity requires multiple sources of evidence to support or refute meaningful score interpretation” (p.831). In line with unified validity theory, Kane (2002, 2004, 2006) contends that validity argument can be established through an argument that relates theory, predicted relationships and empirical evidence to the suggested score interpretations and meanings in such a way that these relations are meaningful and provide a “cogent presentation of all of the evidence to the proposed interpretations” (Kane, 2002; p. 31). Validity argument downplays the need to define the construct and serves as a conceptual tool for expressing the multifaceted meaning of test scores.

It helps form an interpretive argument for score interpretation which is, according to Kane, a set of inferences and the supporting evidence leading from scores to decisions made. The interpretive argument also specifies how the validity argument can be questioned, weakened, limited or refuted by research that supports rebuttals (Chapelle, 2011). Kane’s interpretive argument adopted Toulmin Model of Argumentation that proposes a layout with six interrelated components for analysing arguments; claims, grounds, warrants, backing, rebuttals, and qualifiers (Toulmin, 1958).

Unlike other models of validity, the argument-based conceptualization of validity is more complicated and requires more professional knowledge and skills on the part of the test developers (Chapelle, Enright, & Jamieson, 2010). However, the level of sophistication involved in developing a validity argument in terms of inferences, assumptions, and claims helps researchers see where more empirical research and evidence is needed. At the same time, the opportunity to include more finer-grain levels of specification, evidence and rationales clarifies the challenges applied linguists face in understanding substantive issues about language assessment” (Chapelle, 201, p. 20).

2.2.3. Test usefulness

Further developments in construct validity conceptualization was suggested by Bachman and Palmer (2010) who adopted a different approach to construct validity. They suggested test usefulness as an all-encompassing feature of a good test which can be assessed in terms of six features of reliability, construct validity, authenticity, interactivity, impact, and practicality. In

their model, construct validity is only one of the features of a useful test. The six features addressed both internal and external aspects of test use. They defined construct validity as the extent to which a given test score can be interpreted as an indicator of the intended abilities or constructs. Interactivness as another feature of test usefulness addresses the socio-cognitive characteristics of the test taker and test context and includes features such as motivation and affective variables in test performance. If construct validity was the unifying concept in Messick's (1989) unitary validity theory, for Bachman and Palmer (2010) the unifying term was test usefulness which seems to have more practical orientation to it.

One feature of the usefulness model is integration of evidential basis such as reliability and construct validity evidence with the consequential basis of validity. Overall, they argued that usefulness is an aggregate of all these features and cannot be independently evaluated. They also maintained that test usefulness should be assessed for every specific test and situation. Unlike unified validity theory, test usefulness seems to be more practical for test application and test developers can use it at local school level or more global level.

These developments in the short history of validation theory can be seen through three interacting processes of *expansion*, *unification*, and *partition* (Markus & Borsboom, 2013). First, expansion can be observed in the emergence of other types of validity such as convergent validity and discriminant validity where relatively higher correlations between scores intended to assess the same construct provide support for convergent validity and lower correlations between scores intended to assess different constructs provide support for discriminant validity. Second, a unification process can be observed in Messick's (1989) unified theory which subsumed different types of validity evidence, considered as distinct and discrete types of evidence, under one unifying general theory or in Bachman and Palmer's (2010) test usefulness validity argument which was the umbrella term for a set of feature. Third, partitioning is best reflected in the traditional tripartite types of validity which deemed appropriate for different types of tests.

2.3. Criticism of the Unified Validity Theory

The unified validity theory has not been without challenges and problems. The main limitation of the theory has been its impracticality and lack of clarity especially for educational settings (Borsboom, Mellenbergh, & van Herdeen, 2004; Brennan, 1998; Embretson, 2011; Fremer, 2000, Lissitz & Samuelsen, 2007). Critics have argued that, in practice, collecting

diverse sources of evidence especially evidence related to the consequential basis of validity which is one of the two pillars of construct validity is too demanding. Consequences of test interpretation and use, as mentioned earlier, contribute to the integrated judgment on the appropriateness of test score interpretation and use (Borsboom & Mellenbergh, 2007; Messick, 1998; Popham, 1997; Reckase, 1998). For test developers who are already overwhelmed with the standards of the profession, it is a challenging undertaking and demands more resources and expertise. Critics have also argued the theory lacks sufficient details to be applied to testing programs in educational settings which is the main reason why it leads to wrong application of the theory or making it “out of reach” for test developers (Fremer, 2000). Kane’s (2004, 2006). In line with this criticism, argument-based approach can be seen as a response to the impracticality of the theory which confused testing bodies prioritizing different kinds of evidence. Borsboom, Mellenbergh, and van Herdeen (2004) maintained that the theory “fails to serve either the theoretically oriented psychologist or the practically inclined tester who advance a much more simplified version of the current theory” (p.1061).

The main criticism, however, was made by Lissitz and Samuelsen (2007) who argued that the unified theory of validity fails to provide adequate guidance for test validation and suggested that the best way to establish the validity of a test is by using appropriate operational definitions and item development procedures. They contend that what educators and researchers want are practical tools for validation not a theoretical framework to conceptualize validity. They considered external aspects irrelevant to test validity per se and conceptualized validity as an internal aspect of the test which can be established by content analysis. They suggested that validity should be established through operational definitions and test development procedures which can be examined through experts’ judgements, task analysis, and completeness of test specifications regarding content representativeness. They also questioned if construct validity should be core to the theory and argued that validity is about test scores, not test interpretations. They actually moved back to the traditional view that validity as a property of the test independent of any proposed interpretation or use of the test scores. They suggested that any questions concerning validity must be answered by showing the extent to which the test provides a representative sample of tasks from some content domain. Their criticism caused a lot of reactions in testing research circles who responded to their points.

In response to their criticism, Gorin (2007) rejected equating test construct with test task and content analysis on the ground that it would blur the distinction between construct and content and would take us back to traditional conception of content validity and would withhold many of the advances resulting from unified theory of validity and centrality of construct validity. Kane (2007) argued that Lissitz and Samuelson's proposals

would either reduce validity to a very narrow concern about representativeness of test's content (...) or, more likely, take us back to the situation in the 1970s and 1980s, when we had a profusion of specialized validation methods (shortcuts), each designed for a specific kind of application. (p. 278).

As to the criticism of establishing validity through operational definitions, Gorin (2007) rejected their argument on the ground that it narrows construct validity to results from experts' domain, task analyses, and completeness of test specifications. She argued that these sources of validity are necessary but not sufficient evidence of validity. Gorin also rejected Lissitz and Samuelson's position on validity as an internal aspect of the test on the ground that they placed less attention to substantive examination of test taker's processing and cognition which further narrowed their concept of internal validity. She maintained that test development based on operational definition would lead to "empirically unsubstantiated claims regarding score meaning" (p. 457) and advocated correlations of test scores with external variables as a useful source of additional evidence to consider in the validity argument. She also suggested that in the larger context of score meanings, information about content relevance and content representativeness as well as criterion-relatedness all contribute to score interpretations and construct validity of the test. In regard to Lissitz and Samuelson's position on validity as a feature of test score, not interpretation, most testing scholars emphasize that both the test and the specific purpose for which the test is used should be valid and these two dimensions cannot be separated (Gorin, 2007, Kane, 2008; Sireci, 2007). They contend that test scores should be referenced to the purpose for which they are being used. In this regard, Gorin (2007) maintained that "Validity evidence, if considered only in terms of the test itself, is meaningless" (p.406).

Based on the arguments for and against the unified validity theory, I can discern the challenges and impracticalities of the theory for test developers. Even Kane's (2002, 2006, 2011) validity argument model which expanded the theory and developed a working model for

validation studies within the unified validity theory is still a bar too high for most test practitioners to reach. The technicalities involved in understanding and practicing the model are beyond the capacities available to teachers and test developers especially those who are using test at local levels. More work is still needed to address the practicality gap. The unitary validity theory model seems to be more practical for testing programs and testing bodies especially high-stakes or large-scale public tests, but I wonder how the theory might be used in specific validation studies or low-scale tests.

All in all, in the context of this study, the unified validity theory informed the current study in different ways. It helped design the study by incorporating different sources of evidence each presenting a different perspective and contributing to the understanding of the test construct. It also provided a broader context to understand and interpret the IELTS RCM test scores and decide how meaningful and appropriate the decisions are. In fact, the theory provided a broad context to search for different sources of evidence. The diverse sources of evidence could enrich the evidence for test use and reveal possible gaps and shortcomings in the construct of the IELTS RCM. The theory justified collecting data from diverse sources such as test content, test takers' accounts of the processes and strategies used during test performance, and experts' judgements to provide some evidence for validity of the test. Each of these sources addressed some internal or external dimensions of test validity. Test content analysis and task features analysis can identify some internal aspects of test validity while test takers' processes provide both an internal and external angle to the test and examine it in the context of the test takers' knowledge and reading skills. These sources of evidence could be cross validated to enrich the sources of validity used in the study. The theory also accommodated different dimensions of test use and integrated them into a meaningful whole. The theory also highlighted the need to search for cognitive evidence that justifies use of the IELTS RCM as a representative of academic reading ability for university admission of ESL/EFL learners. In light of the multiple sources of evidence decision makers may find justifications for modifying test tasks and/or test format or continue using the test as is. It can provide insight in re-operationalization of the construct or changing the minimum (cut off score) requirement for university admission.

2.4. Gorin's (2006) Process-based Model of Construct Validity

Another component of the theoretical framework that could help frame the validity model of the study was adopted from Gorin (2006) who suggested a process-oriented cognitive model of construct validity. Unlike the unified validity theory, which seems to be more a theoretical model, Gorin's (2006) model (See Figure 2.2) served as a working model that allowed examine some specific propositions about the test construct (Cronbach & Meehl, 1995). The model adopts a cognitive perspective in the study of reading comprehension tests and centers cognitive processes and strategies used in the test performance as key in defining the test construct. She operationalized construct validity as a set of skills, knowledge sources, processes and strategies (SKSPs) that are used during test performance. The match between the cognitive processes involved in the test performance and those intended by test developers provide evidence for test validity. Hence, she operationalized validation involving collecting evidence to examine the relationship between the intended and enacted construct. She maintained that the first step in the comparison process is to build a model of the enacted (observed) construct using the same level of description as the model of the intended construct (i.e., the construct definition). Building on Ferrara et al.'s (2004) cognitive processes, she suggested that the match can be examined terms of 1) the alignment between the intended Knowledge, Skills, Processes, and Strategies (KSPSs) used in test taking processes and the observed KSPSs, and 2) the alignment between the intended inferences and the possible inference that can be drawn from test performance. Accordingly, validity of a test can be established by examining the degree of (mis)match between the intended KSPSs and the observed KSPSs on the one hand, and the intended inference and the possible inferences, on the other. With strong alignment, valid score interpretations are supported. Weak alignments indicate that the observed KSPSs may include additional skills or may lack those specified in the intended KSPSs. She argues that not all the KSPSs used in responding to an item can be individually and directly observed and suggests close study of the test by use of introspective and retrospective verbal reports for inferring the underlying processes. She also cautioned that results for construct validity depend on the accuracy of the inferences made by decision makers about the knowledge, skills, strategies, and processes implied or stated in the collected data.

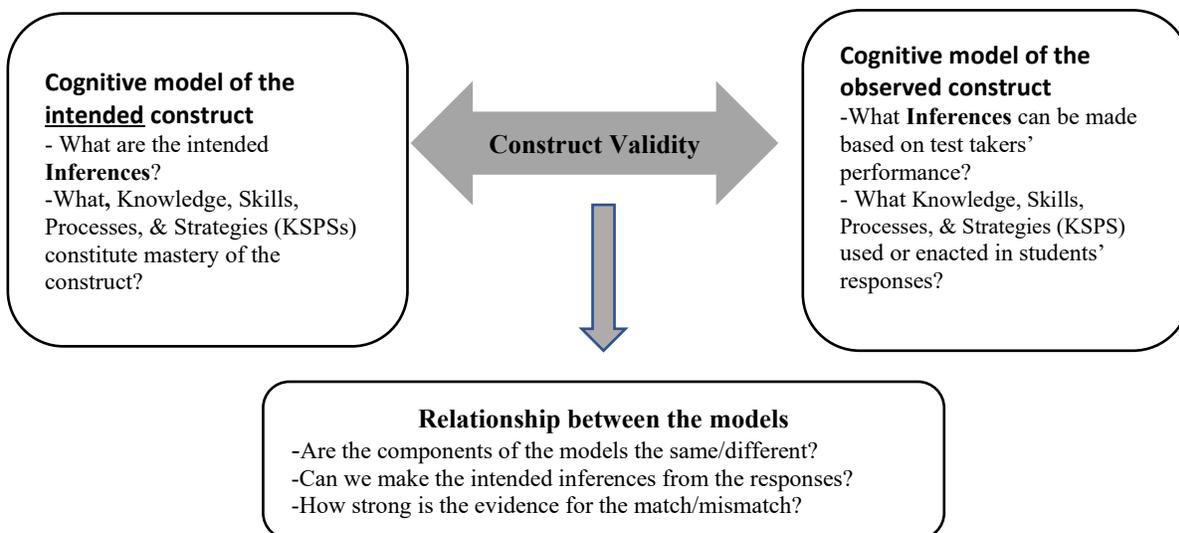


Figure 2.2. Cognitive model of construct validity (adapted from Gorin, 2006)

Gorin's (2006) model was adopted as a working model for the current study. It enjoys the potential to examine the multi-componential construct of reading comprehension tests and lends itself for detailed description of the academic reading tasks and activities. In the context of the IELTS RCM test tasks, the model best fits the purpose of the current research and can capture several aspects of the enacted construct for each test task and item. However, it is worth noting that the intended construct of the IELTS RCM, as stated in the test webpage is expressed in terms of a set of reading skills such as reading and understanding details, reading for gist, reading for main ideas, skimming, understanding logical argument and recognising writers' opinions, attitudes and purpose. They are not expressed in terms academic reading skills or in terms of SKPSs used during test performance. Practically, the claims made by the test can be found in almost all reading comprehension tests. They are too general and lack any details and sophistication. The intended SKPSs are then, not available for comparison with the actual SKPSs observed in test takers' performance. To make up for such shortcoming, the observed SKPSs will be compared against the experts' accounts and judgements of the test tasks, test content, and the available literature.

2.5. Test Validation

Another aspect of validity relevant to the current study had to do with the procedures and methods used in test validation studies. Just as validity theories developed over time, so did the methodologies used in test validation studies. In this section, two methodological aspects of

validation are discussed, namely, quantitative and qualitative approaches and use of verbal reports in validation studies.

As discussed above, test developers need to provide different sources of evidence to support the appropriateness, meaningfulness, and usefulness of the inferences drawn and decisions made on the basis of the test scores. Immigration decisions, entry and exit decisions, admission, employment, and promotion decisions all depend on test score interpretation. Test validation helps test developers and testing researchers collect, synthesize, and integrate diverse sources of evidence for test score interpretation (Kane, 2004, 2011). What is implied is the complementary nature of diverse sources of validity evidence. Using multiple sources of evidence help testing researchers, triangulate the sources of data to establish validity claim of the test. These sources of evidence can be collected by using different techniques and methods and procedures. Data for test validation can be collected by the test developer or a third-party outsider. However, it is more likely that test developers are biased and supply supportive evidence for test validity while a testing researcher does not usually have an interest to support or reject a certain interpretation of a given test and seem to be more objective and trustworthy. Kane (2011) cautions test developers not to take an advocacy role in validation processes. In either case, the evidence collected needs to be integrated and cross-checked to establish validity argument (Kane, 2006). In the following section qualitative and quantitative approaches to test validation are discussed.

In his unified construct validity theory, Messick (1989) defined test validity as an integrated evaluative judgment which examines “the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p. 14). He viewed all the three traditional types of validity (criterion-referenced validity, content validity, and construct validity) as three different aspects of a single unified form of construct validity which are “complementary forms of evidence to be integrated into an overall judgment of construct validity” (Messick, 1998, p. 37). Collecting different types of evidence involve using different quantitative and qualitative approaches and language testing researchers have extensively studied the evidential basis of language tests by applying quantitative statistical procedures to the scores obtained from the tests under study. In this section, the main features of quantitative and qualitative approaches to construct validity and the rationale for adopting a qualitative approach are briefly presented.

2.5.1. The quantitative approach to validity

The main feature of quantitative statistical studies which range from correlational studies to more complicated statistical procedures regression analysis, factorial analysis, and Rasch analysis, is their focus on the test scores (Cheng et al., 2007; Xie, 2010). As Bachman (1990) pointed out,

A... critical limitation to correlational and experimental approaches to construct validation ... is that these examine only the products of the test taking process, the test scores, and provide no means for investigating the processes of test taking themselves (p. 269).

In this approach, test scores as the final product of test performance serve as the basis for all the statistical procedures applied during the analysis and interpretation procedures. The meanings and characteristics that are associated with the test under study tend to be nothing but the function of test scores. Results of such statistical procedures are used to label a test as a measure of a given construct. As Song (2008) pointed out, the quantitative approach to validity just looks at the test score which is narrowly defined as correct-incorrect and provides no information about the undercurrents of the test scores or how test takers did the test, nor the actual processes used by the test takers in their test performance. Such an approach to construct validity is insufficient and problematic as it does not capture the core activities and processes used by the test takers in responding to a test task/item (Purpura, 2011). Test scores fall short in demonstrating the skills and processes that are actually involved in the target construct and leave out the socio-cognitive variables and context in which the test is used.

Another main feature of the quantitative approach which has dominated SLA research including language testing and validation studies for decades is that it often examines the external correlates of the test (criterion-related validity and concurrent validity), but not the internal dynamics of test performance. Quantitative validation studies adopt convergent-discriminate designs and multi-trait multi-method designs to demonstrate the extent to which the tests used in the study measure the same construct or not, assuming that a set of highly correlated tests measure the same construct and tests that correlate low measure different constructs. Another shortcoming of quantitative approach of validation studies using correlational and factor analytical procedures lies in its circularity. Sternberg (1979) pointed out that in quantitative studies one test is validated against another test which has already been tested against another

test and this cycle never ends. Therefore, the nature of the construct is dependent on the validity of the criterion tests selected and remains very general without any specifications (McNemar, 1964).

In the context of the assessment of reading comprehension which is the main focus of the current study, most validation studies of large-scale reading comprehension tests have widely adopted quantitative research designs and have relied on the comparison of scores of one reading test with scores of other reading tests by applying different statistical procedures such as correlational analysis and factorial design for data analysis (Cutting and Scarborough, 2006; Francis, Fletcher, Catts, & Tomblin, 2005; Keenan, Betjemann, & Olson, 2008; Nation, & Snowling, 1997; Uccelli, Galloway, Barr, Meneses, & Dobbs, 2015). In these reading validation studies, a set of reading tests used with children or adult learners at school level, university level, or in special education programs were selected and administered to a large group of participants. Take, for example, Vellutino, Tunmer, Jaccard, & Chen (2007) who studied components of reading ability by administering a large battery of tests assessing reading skills and subskills and reading-related cognitive abilities to define latent constructs representing reading skills and subskills; or, a recent work by Uccelli, Galloway, Barr, Meneses, & Dobbs (2015) who studied the construct of academic language proficiency in the context of a linguistically and socioeconomically diverse cross sample of 218 students (grades 4-6). Uccelli, et al. administered four reading assessment instruments; the Core Academic Language Skills Instrument (CALSi), a standardized reading comprehension assessment (Gates- Reading Test), an academic vocabulary test (Vocabulary Association Test), and a word reading fluency test (Test of Silent Word Reading) and analyzed the students' scores by applying the general linear model of analysis of variance (ANOVA).

These studies provide convergent-discriminant evidence for the test under study, but they fail to capture experiences of the test takers and the actual processes and strategies they used in test performance. Statistical analyses are silent about what the construct actually is and provide no information about the *what* and *how* of the test takers' reading behaviors and activities used during test performance. They label a test measuring a given construct based on the results of test scores and the statistical procedures that are used in interpreting them. These analytic procedures, however, are not enough to know what is actually measured by the test because test scores do not hold any of the information that define the actual components of a construct. In

addition to such statistical evidence, a test construct needs to be examined in the context of actual test performance of real test takers. Gorin (2006) highlighted the limitation of quantitative approaches and argued that in score-based approach to validity, the observable examinee–item interactions, are limited to examinees’ answers with little observable information regarding the underlying cognitive processes and strategies used. In such a context a lot of information that needs to be built into the definition of the construct will be missed. Xie (2011) contends that what test designers think they are measuring may not necessarily be the same as what is actually measured by the test. Test takers may use a variety of processes and strategies to get the answer. They may make a correct or wrong hypothesis and get the answer or they may simply use wild guessing. Thus, understanding the processes and strategies test takers use during test performance via verbal reports and other appropriate self-reported methods can reveal the reality of what the construct is actually measuring. Results of process-oriented validation studies have implications for test development. In light of results of process-oriented validation studies, test developers can refine a test and make it more appropriate for the purpose they intend to measure (Storey, 1997). Furthermore, Cohen (2006) advocated use of qualitative approach and use of verbal reports in validation studies “can provide valuable information about what test actually are measuring” (p. 325). These arguments and studies indicate the need to consider diverse sources of evidence such as processes and strategies used during test performance for examining construct validity of language tests including reading comprehension tests.

2.5.2. The qualitative approaches to validity

Qualitative approaches, which began as a reaction to the shortcomings and limitations of the traditional statistical approaches to validation, are not new to validation research (See, for example, Fox & Cheng, 2007, 2016; Gorin, 2005; 2009; Hilden & Pressley, 2011) but they have been much less evident in test validation studies. They examine the test construct in a broader context of test use by collecting evidence from more stakeholders and focusing on qualitative dimensions of test use. They focus on the undercurrents of test score and examine different dimensions of test performance such as the evidence from the internal mental processes and strategies used by test takers during test performance, test takers’ experience and perceptions of the test, experts’ accounts and judgements of the test, and content analysis of the texts and task features. They adopt different socio-cognitive perspectives and examine both the cognitive processes and the social context in which the test is administered. Moreover, they look into the

test takers' perceptions, experiences of the test, their familiarity with the test, and their motivation and produces detailed description of participants' experience, feelings, and opinions to interpret the meanings of their actions (Denzin, 1989). In addition to the test product, qualitative approaches also highlight the interaction between the test taker and the test and to collect the evidence necessary to define, interpret, or modify the construct of the test. Most importantly, they focus on the processes and strategies claimed by the test takers used during test performance which serve as key source of evidence to the construct of the test.

Qualitative approaches have the potential of opening a window that provides some information about the dynamic nature of test performance and the features that form the construct or influence it. Qualitative research describes different features of test performance that are internal to the test task, and hence provide a most informative source in examining the underlying processes and strategies of the construct. Using qualitative approaches, testing researchers can also examine the influence of L1, level of language proficiency, and cross-cultural differences on the test taking processes. In light of qualitative approach, validity researchers can examine the socio-cognitive processes that impact test performance and inform construct validity of the test. In support of a cognitive, process-based approach, Gorin (2006) recommended examining construct representation of the test task by detailed analysis of test taking processes. She suggested establishing test validity through appropriate conceptual and operational definition of the test construct and item development procedures. In the same vein, Embreston (1985) adopted a cognitive approach to test validity and defined construct representation as "the processes, strategies, and knowledge stores that are involved in item responses. Construct representation is understood when the various components, metacognitive components, strategies and knowledge stores that are involved in solving psychometric items are explained," (p.196) The need for use of cognitive evidence in test validation is best summarized by Pellegrino (2009), "I do not see how we can effectively pursue issues of construct validity without some principled applications of cognitive theory in the design and validation of tests." (P. 54).

In terms of the methodology, language testing researchers and test validation researchers who adopt qualitative approach have mostly relied on different forms of verbal reports such as introspective and retrospective verbal report, checklist, strategy questionnaire, stimulated immediate recall, and cognitive interview. These self-reported verbal accounts allow the

readers/test takers to describe what they do while performing on the test. However, since the readers have no access to unconscious automatic processes, the method cannot capture unconscious processes (Green and Gilhooly, 1996, van den Broek et al., 2005). These techniques have great potential to open a window into the conscious mental processes used by the test takers and offer researchers a unique opportunity to examine the cognitive reading processes that readers/test takers can access and verbalize (Afflerbach, 2000; Graesser, Wiemer-Hastings, & Wiemer-Hastings, 2001; Pressley & Afflerbach, 1995, Xie, 2011). Verbal reports can capture the contents of readers' working memory that are available for verbal report (Cohen, 2006; Ericsson and Simon, 1993; Hilden & Pressley, 2011; Miller & Brewer, 2003). Further discussion of the use of verbal reports in reading research and testing reading will be presented in Chapter Three.

Thus, the qualitative approach in language testing and test validation can contribute to the understanding of test construct. In light of the detailed description of the processes, strategies used, the experiences, feelings, and perceptions of test takers obtained through qualitative approach methods, testing researcher can look at test performance in a more comprehensive context and understand the complex features of test constructs, the test takers' activities and possible cultural influence on their test performance. Qualitative approach is employed to achieve deeper insights into issues related to designing, administering, and interpreting language tests (Chalhoub-Deville & Deville, 2008). The insights gained through qualitative approach is not limited to validity studies. They can be used for designing, administering and interpreting language assessment.

In brief, quantitative approaches to test validity tend to adopt a narrower view of validity and simplify a test construct in terms of statistical significance and certain patterns observed in test scores. As Gorin (2007) argued, quantitative approaches adopt an external score-based definition of a construct which is often based on its correlation with another test and ignore the validity as a function of the internal content of tests and the cognitive processes used by the test takers. Quantitative approaches to validation can be also seen as top-down, deductive experimental approaches which seek to support/refute what is hypothesized to be a construct by the test developer. They treat a test as a hypothesis which can be tested with a sample of test takers. Test score analysis can provide some statistical evidence for or against the test developers' assumptions about the intended construct. Quantitative approaches to test validity, in

fact, examines the *potential* of the test to tap the intended construct by testing it against an external criterion. Qualitative approaches, on the other hand, adopt a more inclusive exploratory bottom-up inductive position and examine diverse sources of evidence such as perceptions, and experiences of other stakeholders' activities, knowledge sources, processes, and strategies used by the test takers. It deals with the *actual* construct of the test as it is processed and manifested by the test takers' cognitive activities. It is concerned with what the test construct is, not what it can be or should be. Figure 2.3 presents some of the difference between quantitative and qualitative approach in test validation.

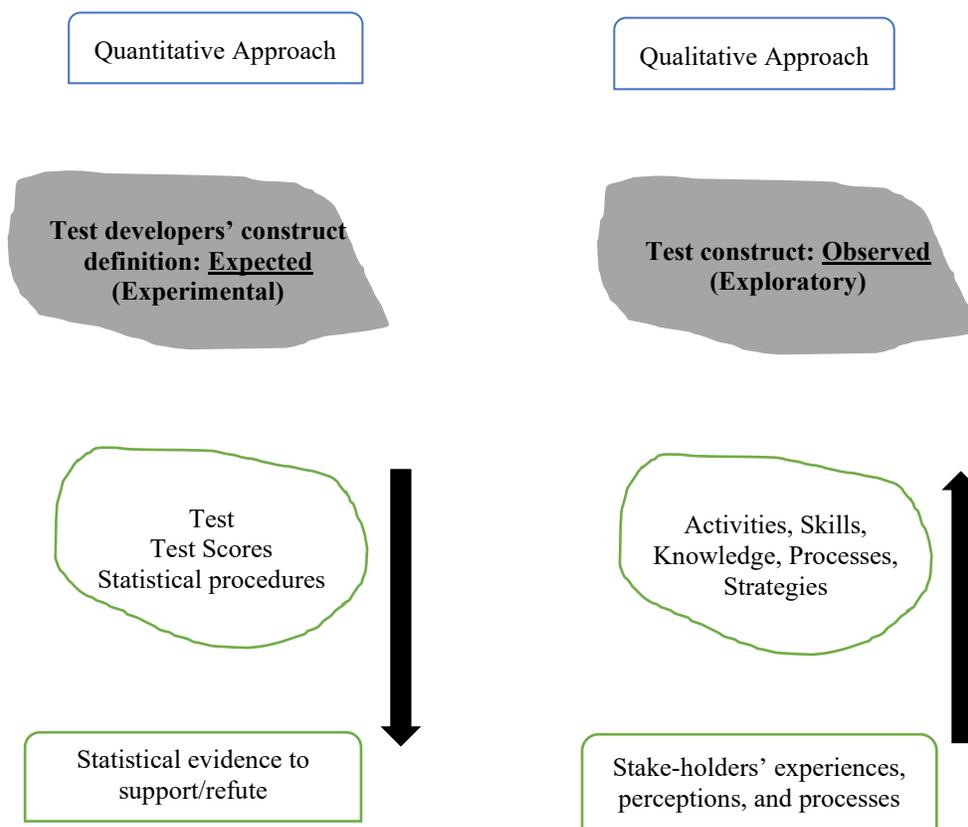


Figure 2.3. Quantitative versus qualitative approaches to test validation

2.6. Sources of Validity Evidence

A key feature in any validation study is the question of validity approach adopted by the researcher. As mentioned in the previous section, researchers may adopt quantitative, qualitative, or a mix of these two approaches for test validations. Due to the importance of validity approach adopted in validation studies, they are briefly presented here.

Contrary to the quantitative statistical approach to validity which relies just on test scores, qualitative approach to test validation tends to incorporate multiple sources of evidence by examining perceptions, perspectives, and accounts of different stakeholders to open a new window for examination of test construct. Intuitively, the more sources of evidence used in validity argument, the more room to argue for validity of the test and its use (Bachman & Palmer, 2010). For instance, test content holds a rich source of information about what the test intends to measure. Also, test takers' accounts of the processes and the strategies they use in test performance can provide a different context for the test developer to see how the test takers use different processes and strategies in their interaction with each test item. Their accounts provide more insight into what the item or the test task is actually measuring. Furthermore, language testing experts, with their knowledge and awareness of the key concerns and issues in language testing, and their perception of the construct of each test task can add to the validity evidence of the test. Significance of these less-researched sources of validity evidence which were adopted in the current study are briefly discussed.

2.6.1. Content analysis evidence

The least frequently studied source for test validation is test content. Content analysis as a research method can be used to identify the presence of certain words, concepts, and features within texts or sets of texts. Stemler (2001) defined it as a method for making inferences by objectively identifying specified characteristics of texts. As a method of analyzing the content of documents, it helps researchers examine features or theoretical issues to enhance understanding of the document/text (Elo & Kyngas, 2007). In the context of language testing, it can be a rich source of test score interpretation and has a long history in language test evaluation. For example, Cronbach (1971) recognized the importance of examining and establishing test content and its relevance in test development and interpretation. He contends that investigating the construct validity of a test based on test scores is not sufficient if it lacks information on the actual performance on the test. In the traditional view of validity, content validity was the main source of validity. However, with emergence of unified construct validity theory, most attentions shifted from test content to the technical and statistical features of construct validity. Content validity seemed to be old fashioned and was less frequently used in validation studies and projects. During the last two decades, it remained the neglected child in the family of different sources of validity evidence.

Lazaraton and Taylor (2007) and Lazaraton and Frantz (1997) argued that task features cannot be seen as independent of the test takers' performance because they can influence task processes and performance. They concluded that test content and task features can still provide further evidence for construct validity of the test. Bachman, Davidson and Milanovic (1996) also recommended test developers pay due attention to both abilities to be measured and the characteristics of the tasks utilized for demonstrating the validity of test score interpretation. They held that characteristics of the test tasks can be best explored by careful analysis of the test content. Recently, some researchers went so far to suggest that the best way to establish the validity of a test is by using appropriate operational definitions and item development procedures and conceptualized validity as an internal aspect of the test which can be established by content analysis (Lissitz & Samuelsen, 2007). They suggested that validity can be established through operational definitions and test development procedures which can be examined through experts' judgements, task analysis, and completeness of test specifications regarding content representativeness. However, this view was questioned on the ground that it ignores other dimensions of validity such as internal dynamics and external criterion. Based on the analysis of test content and the linguistic, topical, cultural features of the texts and task, test developers will have a broader and more detailed picture of the language skills, sub-skills, and processes being measured. Content analysis of the test items and the test tasks can also provide a useful context for understanding test takers' performance. For example, in the context of assessing reading comprehension, the processes used by test takers can make more sense linguistic, textual, topical demands, and task features of the test are precisely analyzed and identified. In addition to studying the actual test processes, content analysis can help test developers evaluate the quality of the items or the test task to define and identify some key components and features of the construct of the test and the context of its use.

Another factor that justifies the significance of test content is related to the congruence of test content the target domain. Test tasks are expected to represent the target language use domain (TLU). Bachman (1990) and Buchman and Palmer (2010) maintained that one main source of validity evidence is the congruence between the cognitive processes used in the test tasks/items and cognitive processes used in the target language use domain (TLU) task. By implication, in the context of assessing academic reading comprehension, linguistic, textual, topical, cultural, and task features of the actual academic domain should be represented by the

texts and test tasks. In fact, validity of the inferences made on the basis of test scores of an academic reading test depends on the degree to which they represent the same cognitive processes and features of the academic domain. Therefore, one best way to establish and examine correspondence of the test domain and target language used domain is through detailed analysis of the test content and its task features. Content analysis can, then, help identify the potential skills and knowledge stores required in responding to the test tasks. Evidence of such correspondence cannot be ignored in the examination of construct validity of an academic test. In this regard, Green et al. (2010) argued for the comparability of texts used in reading tests with texts used in real domain and suggested some key features for comparison and analysis.

In the context of assessing reading comprehension, analysis of the test content by itself is a rich source of information that can support validity argument of the test and can reveal useful information about the linguistic, textual and topical features of the reading text, its layout and readability, the test tasks, and the individual items. Inclusion of test content analysis allows focus “on some entity that is external to the reader-the reading task” (Moor, Morton, & Price 2009 p. 9). It can pinpoint some of the reading skills that might be tapped by the test tasks, the specific parts of the text that contain the relevant information required for answering the item and the way the text and the test tasks integrate to achieve assessment purpose. These details of test content can be integral to the use of test scores and its interpretation. In fact, test score interpretation and analysis can make sense only if there is ample information about the nature of the reading text and the test tasks. Results of content analysis can establish a meaningful context for test score interpretation and use. In view of test content analysis, test score users can deduce the concepts, intentions, and ideas that test developers had in mind in planning and operationalizing the test construct. It allows testing experts evaluate the quality of the items and the whole test based on the specific linguistic, textual and task features of the test tasks. For example, knowing whether the test is tapping into micro level of lexico-grammatical features of the text or measuring macro elements of text structure and discoursal features such as cohesion and coherence, test score users can have a better idea what the test scores mean. Or in a reading comprehension test, if test takers know the type of comprehension (literal or inferential) is tapped by a test task, they can better see how different test takers process the item. Hence, in the light of content analysis, test score users can make more informed and valid decisions based on the test score.

Another factor that contributes to the significance of test content analysis is related to congruence of test tasks the target language use (TLU) domain tasks (Bachman, 1990; Bachman & Palmer, 2010). They contend that cognitive processes elicited by a test task should correspond to those elicited by tasks in the TLU domain. In this respect, content analysis of the test can provide some description of the possible processes involved in test performance which can be later compared with TLU domain processes. In the context of assessing academic reading comprehension, validity of an academic reading test depends on the degree to which it represents the same cognitive processes and features of the academic domain and the linguistic, textual, topical, and cultural features that characterize the actual academic domain. One best way to examine the correspondence of the test domain and target domain is through content analysis. Detailed description of the test tasks is the stepping-stone to start such cross comparison. Evidence of such correspondence cannot be ignored in examination of construct validity of an academic test. In this regard, Green et al. (2010) argued for the comparability of texts used in tests with target real life domain texts and suggested some key features of text comparability for analysis of test content such as text length, grammatical characteristics such as vocabulary, cohesion and rhetorical organization, genre and rhetorical task, topical knowledge, cultural knowledge, and text abstractness for text comparison. These features can provide a framework for analysis of test content.

In brief, there exist a lot information in a test that need exploration. A test cannot explain itself by few descriptive statements. Content analysis of the test can serve as a basis for understanding some of the main features of the test that characterize or influence its construct. Content analysis of a test can help explaining how different test takers interact with each item and test task. Test developers can better explain why test takers did what they did in their test performance if and only if some details of task features of the test are part of the picture. Neither test content nor test performance processes by themselves are sufficient for test validity argument. Results of test content analysis should be seen as complementary sources where test performance is seen in light of the specific characteristics of the test content. In other words, knowing the main features of an item and the test task beforehand can give more context to the understanding of the processes and strategies used by the test takers.

2.6.2. Test taking processes and verbal reports

Another source of validity evidence comes from the cognitive processes and metacognitive strategies used by test takers during test performance. Test takers verbal reports of the cognitive processes and metacognitive strategies used during test performance can serve as a very rich source of information for test validation. Cognitive processes and metacognitive strategies have been frequently used in reading research and validation of reading tests (Caldwell & Leslie, 2010; Cummins & Stallmeyer-Gerard, 2011; Fisher, Frey & Diane, 2011; Jackson, 2016; Meyers, Lytle, Palladino, Devenpeck, & Green, 1990; Smith, 2006, to cite a few). These studies provide detailed information about the specific processes, strategies, and knowledge sources used and the challenges faced by the test takers during test performance. Cognitive dimensions of the test task performance present a window to the test developers to look at the internal dynamic of the test-test taker interaction across test tasks and items and explore the actual skills and abilities measured. They provide information that is essential for defining and operationalizing test construct (Gorin, 2006) and help test developers be more cognizant of what each test item actually measures. In brief, studying test taking processes and strategies, as suggested by Cohen (2006) can help test developers compare test results from different test methods and test takers and identify the extent to which test processes and strategies are learner related strategies, task-related strategies, and test wiseness strategies and decide if they match the what they want to measure.

Significance of the task processes in validation studies lies in the use verbal reports which provide a different window to study test construct. Use of verbal reports in reading research is not new to second language acquisition research and it has been widely used in different areas of teaching, researching, and testing reading comprehension. For instance, different forms of verbal reports have been extensively used as a techniques in teaching reading to improve different aspects of reading comprehension (Caldwell & Leslie, 2010; Crain-Thoreson, Lippman, & McClendon-Magnuson, 1997; Cummins & Stallmeyer-Gerard, 2011; Fisher, Frey & Diane, 2011; Grace, 2016; Jackson, 2016; Kelley & Clausen-Grace, 2008; Meyers, Lytle, Palladino, Devenpeck, & Green, 1990; Smith, 2006; Walker, 2005; White, 2016; Wilhelm, 2001, to cite a few). Verbal reports have been also used in testing different language skills including reading, writing, speaking, and listening (Anderson, Bachman, Perkins, & Cohen, 1991; Azevedo & Cromley, 2004; Buck 1991; Kendeou, Muis, & Fulton, 2011; Kendeou & van den Broek, 2005;

Magliano & Millis, 2003; McNamara, 2004; Yamashita, 2003; Yi'an, 1998; Wijgh, 1996; Wade, 1990). Verbal reports have been also used in the study of reading strategies and processes by L2 readers (Carrell, 1989; Cohen, 1986; Cohen & Cavalcanti, 1987; Jiménez et al., 1996; Malcolm, 2009; Pressley & Afflerbach, 1995; Pritchard, 1990; Zhang, Gu, & Hu, 2008). More specifically in reading, verbal reports have been used in studying strategies used by more successful and less successful readers and younger and older readers (Earthman, 1992; Folger, 2001; Gordon, 1990). Using verbal reports method, researchers have focused on gaining understanding of test-takers' actual processes and strategies that underlie successful completion of the task (Anderson, Bachman, Perkins & Cohen, 1991; Cohen & Upton 2006; Ferne & Choi, 2006; Goa & Gu, 2008; Grotjahn, 1987; Homburg & Spaan, 1981; Klein-Braley, 1985; Klein-Braley & Raatz, 1984; Phakiti, 2003; Raatz & Klein-Braley, 1981; Rupp, Yamashita, 2003; Yi'an, 1998;). Finally, in the last two decades, validation studies have increasingly used verbal reports for validation of different assessment and research instruments; validating questionnaire items in the PISA assessment of student self-efficacy in mathematics Pepper et al. (2018); development and validation of the English writing strategy inventory, Hwang and Lee (2017); development and preliminary validation of measures of L2 language-skill-specific anxiety Cheng (2017).

Introspective and retrospective think aloud data have produced invaluable data on cognitive studies of language processes and test-takers' processing of different test tasks. However, concurrent and retrospective verbal reports have been criticized on two grounds; 1) reactivity and 2) veridicality. Reactivity effect can influence and change the thought process used may interfere and alter the interaction between test takers and the task from what would normally transpire. This interference may be more problematic depending on the nature of the task and the extent to which verbalization capitalizes on similar cognitive functions as the target task. For example, spatial reasoning tasks as compared with verbal problem-solving tasks are not equally affected by introspection think aloud procedures because they engage different cognitive systems (Gorin, 2006). One way to overcome this limitation is to use retrospective protocols. Retrospective approach resolves the cognitive interference issue, but it introduces challenges related to veridicality risks which have to do with the degree to which the verbal reports accurately reflect reality. Critics argue that test takers' may not be able to accurately recollect and provide an accurate description of their own cognition during processing and they may add or delete some of the processes actually used (Bowles, 2010; Ericsson & Simon, 1993; Yoshida, 2008; Pressley &

Afflerbach, 1995; Kuusela & Paul, 2000). Furthermore, some critics of both concurrent and retrospective procedures argue that some processing is implicit or unconscious and students are not aware of, nor can they verbalize behavior related to these skills. In spite of these limitations, which are recognized in language testing research, think aloud procedures especially retrospective reports have been widely used as valid and viable research tools in the examination of some hidden aspects of language test processes because they provide access to some invaluable information that would not have been available otherwise.

In summary, in the course of the last three decades, verbal reports have been fully credited as a valid research approach in different areas of applied linguistics including validation studies. The potentials of verbal reports can provide evidence not only for the cognitive dimension of language processing, but it can also delve into the personal and social dimensions of language processing.

2.6.3. Experts' judgements and accounts

Another source of validity evidence is experts' accounts and judgements of the test. It can serve different functions in language assessment. First, reviewing test tasks and items by an expert or team of experts has been recommended as part of a test development cycle (Brown, 2000). External review of test development procedures by an independent evaluator can contribute to the development of more qualitative test tasks and items. Experts' judgements can be seen as a more serious and professional form of test review that can provide more evidence about the construct of test tasks (Downing & Haladyna, 1995, 1996). Second, expert reviews have also been used as a more practical alternative to avoid the high cost of pilot testing. As test takers can provide some evidence about the test construct so can testing experts. Third, shared understanding and consensus or lack of consensus over the cognitive demands of each item/task can provide further evidence to assess the construct validity of a given test or test task. Fourth, just as perceptions and experiences of test takers matter in understanding the test construct (Fox & Cheng 2016, 2007) so do perceptions and accounts of testing experts. Their accounts can show possible matches or mismatches between what test developers defined as the construct of the test and how it is perceived and understood by testing experts. Experts' accounts and perceptions of the construct of test tasks and what they potentially measure can be used for cross validation of the intended purpose and validity statements of the test against an external source and help better improve test quality. Insights gained from experts' accounts and judgements can help test developers and item

writers better define and operationalize the test construct. Fifth, experts' and test takers' accounts can also support a kind of reverse engineering for retrieving or extracting test specifications. Finally, such accounts can be a very useful tool in language testing research as another source of evidence.

However, experts' judgements are not without problems. It has been shown that even experienced test developers are not necessarily accurate in their predictions of item difficulty (Alderson, 1993; Bachman, 2002; Bejar, 1983; Elder, Iwashita, & McNamara, 2002; Hambleton & Jirka, 2006; Hamp-Lyons & Mathias, 1994). For instance, Alderson & Lukmani (1989) asked nine university instructors to assess "lower," "middle," or "higher" order reading abilities. The findings showed that for a majority of the items there was very little agreement on classification between judges. Similarly, when judges were asked to describe what they thought each item tested, their account of what each item measured also often varied considerably. More interesting finding showed that for items where the judges actually did reach some agreement, there was little relationship between these items either in terms of difficulty or in terms of process level. Other researchers, however, have argued that expert judgements can be improved with training (Fortus, Coriat, & Fund, 1998; Hambleton & Jirka, 2006; MacGregor, Kenyon, Christenson, & Louguit, 2008). Moreover, to achieve higher reliability indices of experts' judgements, the judgement classification and categories need to be clearly and explicitly defined. Vague and ill-defined concepts and categories can cause confusion and add error to judgments. In spite of skepticism about reliability and usefulness of experts' judgements for validation of test content, it is still commonly used in applied linguistics and language testing research and practice as a reliable approach to establish item validity, (Alderson & Kremmel, 2013); Bachman, Davidson, & Milanovic, 1996; Brutten, Perkins, & Upshur, 1991; Lumley, 1993). In sum, reaching an agreement on the nature of a test task/item and the cognitive dimension of a test task is necessary for test item specification, item writing, test score interpretation, and classroom teaching of the reading skills. Such agreements can help define the reading construct and the process categories involved.

In spite of some reported inconsistencies, experts' judgements and occupational experts have proved conducive in standard setting or judgement of acceptability of performance across professions and disciplines (Pill & McNamara, 2015). In applied linguistics and language assessment, it has been used and continues to be used as an independent source for construct

validity of different tests and task types. It has been also frequently used in different activities and practices of language testing such as developing word lists and phrase lists (Ackermann & Chen 2013; Simpson-Vlach & Ellis, 2010), defining language ability or level of language proficiency (Brindley, 1991), setting passing standards in general English language proficiency (Wendt, Woo, & Kenny, 2009), scaling descriptors for scales of language proficiency (Alderson, 1993), standard setting (Tannenbaum & Katz, 2013), and developing tests and measures for professional settings (Lumley, 1995; Pill & McNamara, 2015) judgements of raters in English for health professionals and test validation studies (Qian, Woo, & Banerjee, 2014). Additionally, they have been used for studying different topics in SLA and in language testing including standard setting in standards-based assessment and certification (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006, Pill & McNamara, 2015; Tiffin-Richards & Pant, 2013), reverse engineering (Davidson & Lynch, 2002; Fulcher & Davidson, 2007), estimating item difficulty (Hambleton & Jirka, 2006), and measuring cognitive task demands (Révész, Michel & Gilabert, 2016).

Building on unified validity theory which suggests use of multiple sources of evidence, the present study was an attempt to look at the construct validity of the IELTS RCM by incorporating these three qualitative sources of evidence; 1) content of the test, 2) the skills, knowledge sources, processes, and strategies (SKPSs) test takers used during test performance and 3) experts' accounts and judgements of the test. In light of the content analysis of the test, accounts of a diverse sample of test taker participants, and testing experts' accounts and judgements extensive evidence will be provided that can inform the construct of each test task as well as the whole test. Figure 2.4 presents the diverse sources of evidence that can be built into the validity argument for any given language test.

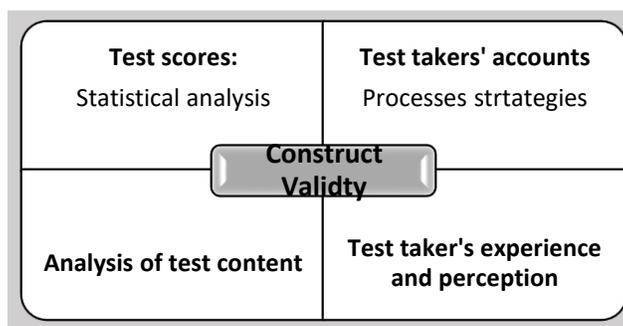


Figure 2.4. Construct validity sources of evidence

2.7. Summary of the Chapter

This chapter provided the theoretical framework adopted in the study which consisted of Messick's (1989) unified validity theory and Gorin's (2006) process-oriented validity theory. First, a brief history of validity theories with more focus on Messick's (1989) unified validity theory was presented followed by some criticism of the theory, and its relevance to the current study. Messick's unified validity theory argued for multi-faceted evidence-based interpretation and use of test scores. He maintained that validity is one overall judgement about the inferences drawn from test scores; that the validity of such inferences depends on the integration of different sources of evidence (e.g., as content evidence, psychometric evidence, and, criterion-related evidence); and that, "the evidential basis of test interpretation is construct validity" (Messick, 1989, p.34). The chapter also introduced and discussed Gorin's (2006) process-oriented model presented a working model for test validation studies, consistent with Messick's view. Her model proposes validating the match between the intended and the actual skills, knowledge sources, processes, and strategies (SKSPs) that test takers use during test taking processes.

In the next chapter, Khalifa and Weir's (2009) cognitive reading comprehension model and the literature relevant to the factors that influence reading comprehension will be presented, This includes the use of cognitive processes and metacognitive strategies in reading, influence of readers' variables such as their language background, level of language proficiency, and their motivation on reading will be discussed. The chapter will also discuss text features that influence reading comprehension.

CHAPTER THREE: LITERATURE REVIEW

3.1. Introduction

In the context of this study, reading was viewed as a multi-componential socio-cognitive process where a constellation of factors and variables impacts the reading and comprehension process (Khalifa and Weir, 2009). Consistent with this view, testing and assessment of reading should reflect not only the cognitive and social aspects of reading, but also the complexity and multiplicity of factors and their interactions in operationalizing the reading construct on a test. As discussed in Chapter Two, Cronbach and Meehl (1955) conceptualized construct validity in terms of a nomological network whereby a test construct emerges from the inter-relationship between a set of coherently related theoretical concepts and different sources of empirical evidence. The current study sought to examine the construct validity of the IELTS RCM in the context of different sources of empirical evidence.

The last two chapters presented the research questions, rationale of the study, and the theoretical framework adopted. The focus of this chapter which is divided into two main parts is on the construct of reading comprehension. The first part discusses Khalifa and Weir's (2009) reading comprehension model which, informed by the unified validity theory of Messick (1989), was adopted as the reading model for the study, along with other relevant reading theories that support the model. Khalifa and Weir's (2009) model of reading incorporates information processing concepts and variables of human cognition in its description of the multi-dimensionality of reading comprehension and accounts for the intricacies involved in the construct of reading comprehension. Because the model defines and characterizes the reading construct, it can help in better understanding and interpreting empirical findings related to the reading comprehension processes of the IELTS RCM. The first part ends with a discussion of empirical evidence for validity that supports Khalifa and Weir's model.

The second part reviews empirical research related to the reading comprehension construct and factors that influence reading comprehension processes. As cognitive processes and metacognitive strategies are key components of the comprehension construct, the first section of second part presents research findings related to the construct of the IELTS RCM. Then, the role of metacognitive strategies in reading comprehension processes and reading test performance are reviewed. Next, readers' variables that influence reading processes and test

performance, including their language background, level of language proficiency, and attitudes towards test will be presented. This is followed by a review of empirical findings that have established the impact of textual features on reading processes and reading test performance. Textual features are discussed under different categories including linguistic features, text content and topic, text authenticity and coherence, text types, and the propositional density of text. Finally, the relevance of these studies to the current research will be presented. Figure 3.1 provides a schematic representation of the theoretical and empirical background that informed the definition of the construct of reading comprehension that was considered in the present study.

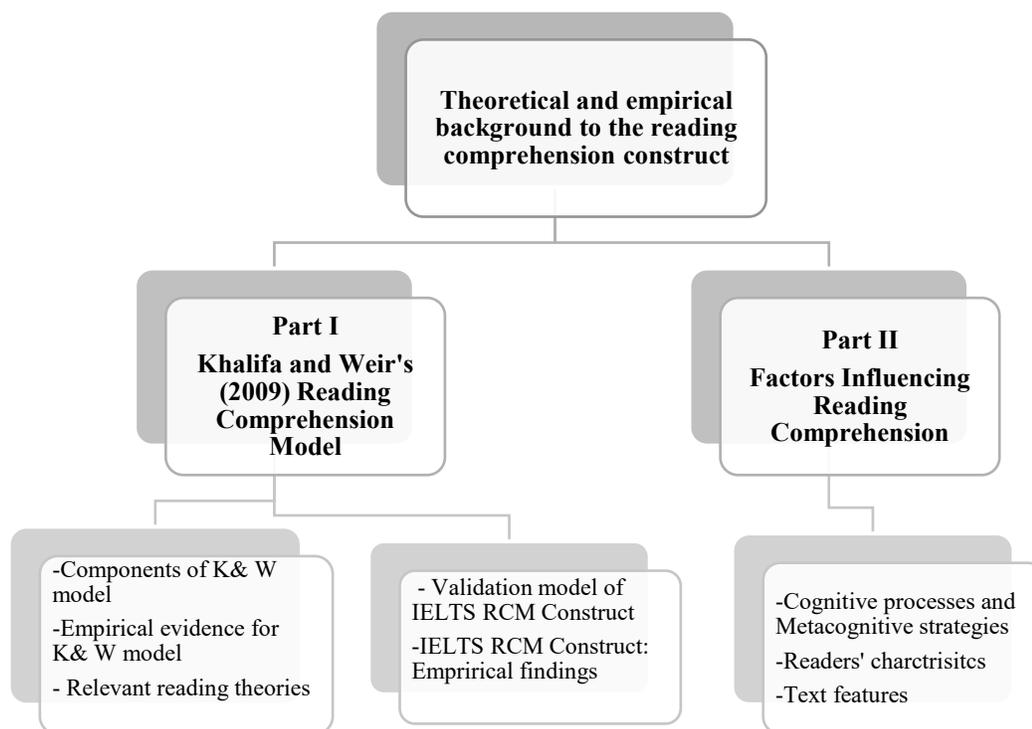


Figure 3.1. Theoretical and empirical background to reading construct

Part I

3.2. Khalifa and Weir's (2009) cognitive reading comprehension model

Consistent with the process-oriented validity model (Gorin, 2006), Khalifa and Weir's (2009) model (Figure 3.2) identifies three main components for reading comprehension: 1) cognitive processes (the central core component), 2) metacognitive strategies (the metacognitive component), and 3), different types of language knowledge and general world knowledge (the knowledge base component). In addition to some unique distinctions between different types of reading such as *local-global* (defined as attending to lexico grammatical features at low level of

sentence versus attending to textual comprehension at high level of discourse) and *careful-expeditious* (defined as extracting complete clear meanings within or beyond sentences right up to the level of the entire text, in order to construct the macrostructure text representation versus quick, selective, strategic, and efficient reading to access relevant information) the model frames cognitive processes, metacognitive strategies, and the knowledge sources (linguistic and non-linguistic) as a single unified model. These distinctions are fully discussed in Section 3.4 below.

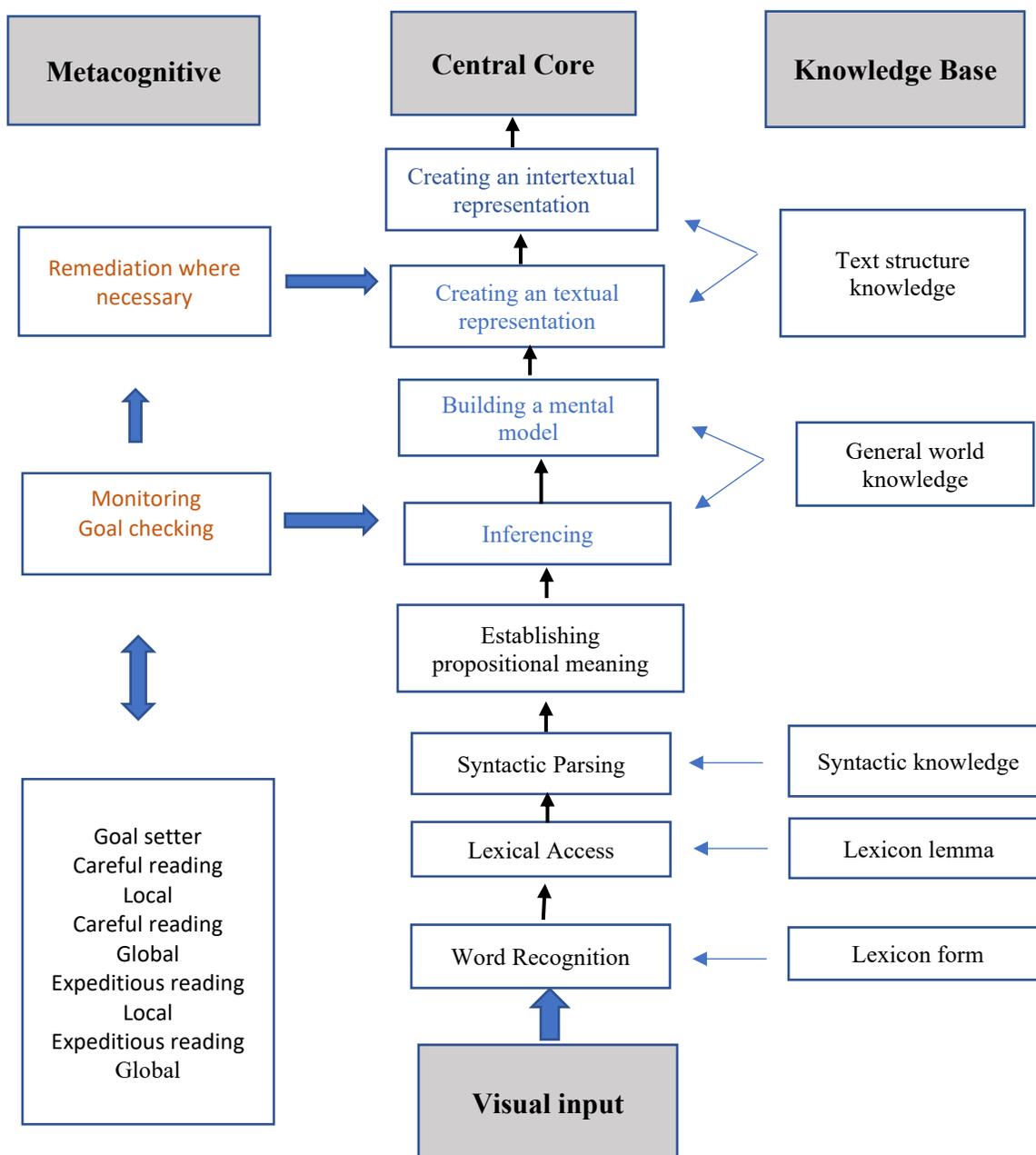


Figure 3.2. Cognitive processing model of reading comprehension (Khalifa & Weir, 2009, p.43)

3.3. Central core component: Cognitive processes

The central core component is the main component of Khalifa and Weir's (2009) model which is informed by other, different cognitive models of reading, and extensive, well-established findings in reading research (e.g., Katalayi, 2018; Sarojani and Krishnan; 2011). The Khalifa and Weir model deals with the main reading processes that are commonly recognized in the literature: word recognition, lexical access, sentence parsing, establishing propositional meaning, inferencing, building a mental model, creating a text level representation, and finally creating a situational intertextual representation (Phakiti, 2003; Rayner, et al, 2012). The relevance of these cognitive processes in reading research and test validation research lies in their potential to show the bottom-up linguistic and text processing as well as the top-down inferencing processes of comprehension involved in reading and reading test performance. As one main question regarding validity of the IELTS RCM was related to the specific process(es) the RCM's tasks measured, Khalifa and Weir's (2009) cognitive processing model provided an ideal framework to find out which of these specific cognitive processes are tapped by the test tasks. The central core component of Khalifa and Weir's model is built on rich theoretical and empirical findings related to reading theories and reading research. In this section, some of these theoretical and empirical foundations are reviewed.

Reviewing a massive body of research regarding the cognitive processes of reading conducted over three decades, Rayner et al. (2012) provided a detailed account of cognitive reading processes. The first process of reading, they argued, is visual word recognition (also referred to as lexical access), which serves as the foundation of reading and provides word meaning and syntactic information for further processing and comprehension. Accurate and automatic lexical access is essential for comprehension (Perfetti & Hart, 2001) and disruption in word recognition disrupts comprehension processes in one way or another.

Lexical access is where form meets meaning and it functions as the basis that helps a reader achieve higher order semantic and comprehension processes. Research findings have shown that semantic knowledge of words assists in connecting word recognition to comprehension. It is the bridge that leads to comprehension. For example, Berends and Reitsma (2006) observed that practice with printed words, with specific instructions focusing on the semantic characteristics of the words, promoted reading acquisition. Moreover, research findings

support the view that vocabulary size is directly related to the ability to understand printed words (Freebody & Anderson, 1983; Proctor, August, Carlo, & Snow, 2006).

Upon completion of word recognition, a reader's attention shifts to the next word which is integrated into the tentative representation of the phrase and sentence. In sentence parsing, different sources of information, such as syntactic information from lexicon, knowledge of grammar, and discourse context, feed the reader's/parser's development of a syntactic representation of the sentence. In essence, sentence parsing (as the second process) identifies the appropriate syntactic constituents of a sentence and their relationships. Based on the *interactive constrained-based model* (McRae & Matsuk, 2013), readers simultaneously use multiple sources of information and sentence parsing to build understanding of grammatical structure from detailed information about individual lexical items (Ford et al., 1982; Marslen-Wilson & Tyler, 1987). Parsing involves searching among all possible cues (Fodor et al., 1974), which compete with each other, to determine the meaning of the sentence (McClelland & Rumelhart, 1986). It is important to note that sentence parsing processes operate at different rates for different readers: skilled readers, native L1 readers, children, older people, and L2 readers.

To understand the overall meaning of the text, readers also need to go beyond literal meaning to fully understand a text by developing networks of propositions to construct mental models of the text (Kintsch, 1988, 1998). The knowledge sources used for development of mental representations are extensive and amorphous. Readers also need to use anaphoric references to link sentences together, use connectives to relate propositions to one another, and use the information structure of sentences, such as given and new information (Gordon & Chan, 1995). Text comprehension also involves appeals to knowledge about the world, attending to linguistic devices that can explicate the text, and careful knitting of sentences together, which requires pronoun interpretation (Kintsch & Rawson, 2005). Through these processes, readers arrive at a text representation. Due to their significance in reading comprehension, these processes of meaning making are further discussed below.

3.3.1. Text (discourse) comprehension processes

Another elaborate and more frequently cited model of text comprehension which is relevant to the high-level processes in Khalifa and Weir' (2009) model is Kintsch's model of discourse representation (Kintsch 1974, 1988, 1994, 1998; Kintsch & Rawson, 2005). In his model, text comprehension is achieved at different levels of processing: *textbase* (microstructure,

macro structure) and *situation model*. Textbase or microstructure and macrostructure is related to the meanings expressed in the text. Microstructure is constructed by forming propositional units at the linguistic level of the text and analysis of coherence relations among them. This stage involves conferencing, anaphora resolution, making some bridging inferences, and pronoun identification. Different models of microstructure processing have been suggested. One of the most frequently cited models of micro-structure processing is the *Construction-integration (CI) Model* (Kintsch, 1988, 1998). Macrostructure, on the other hand, deals with the global structure of text processes, global topics, and their interrelationships. One key aspect of macrostructure formation is topic identification, which is achieved by using some signaling devices or prior knowledge in the text.

Another key component of comprehension of Kintsch's model is the situation model, which is essentially a mental model of the situation presented in the text. It accounts for integration of the information provided by the text with relevant background and prior knowledge of the reader and the reader's goal for reading the text. A critical process of the situation model is inferencing, which is an indispensable process of reading and accounts for a major component of discourse comprehension. Inferencing which is also one of the key components of central core of Khalifa and Weir's (2009) model is not just a reading skill, rather it is used for understanding of the world and people's actions in daily life. It is a basic "cognitive mechanism that connects what we are currently attempting to understand with memory resources that provide our background knowledge" (Grabe, 2009, p. 68).

Regarding inferencing as one of the key processes of text comprehension, Grabe (2009) mentions several functions of inferencing in developing textbase and situation models: 1) integration of new information with prior information, 2) interpretation of decontextualized information, 3) synthesis of information from different sources 4) evaluation of the information in terms of the reader's goal, and 5) interpretation of information that contradicts prior expectations. Researchers have suggested some taxonomies of inference types (Zwann and Rapp, 2006; Zwann and Radvansky, 1998; van den Broek, 1994). Quick automatic inferencing is simple and easy inferencing while controlled inferencing is considered highly resource demanding. Inferences might also be knowledge based, relying on topical knowledge, or text based, relying on the text (Kintsch & Rawson, 2005). Another distinction worth mentioning is the difference between bridging inference that help maintain coherence in the text

representations of the text being read, and elaborate inferences that add information to the situation model, provide additional information beyond what is explicitly stated in the text, but are not needed for maintaining text coherence. Readers, however, differ in their use of one type of inference or another. For example, some readers might be minimalist and make only necessary inferences while other readers who pay more attention to details use inferencing more frequently (McKoon & Ratcliff, 1992).

Finally, discourse comprehension also involves attending to local and global text coherence and cohesion. Some coherence models discuss that propositions are encoded and connected on the basis of argument overlap or causal relations (Kintsch, 1983; van Dijk, 1978; van Dijk & Kintsch, 1983). They argue that parts of the propositions are available in the working memory for integration with the rest of the text. Integration of the proposition revolves around the principle of *coherence*. Two types of coherence are assumed; local coherence and global coherence (O'Brien & Cook, 2015). Local coherence is achieved by integrating incoming information with the information that is in the working memory, while global coherence is achieved by “establishing connections between incoming information and related information that is no longer within the active portion of the text representation” (O'Brien & Cook, 2015, p. 218).

3.3.2. Brown and Abeywickrama's (2004) dimensions of reading

Khalifa and Weir's (2009) cognitive processes are closely related to Brown and Abeywickrama's (2004) conceptualization of reading dimensions. They suggest four dimensions of reading comprehension. The dimensions correspond with the central core component of Khalifa and Weir's (2009) model and are very relevant to the construct of reading comprehension. These dimensions are: 1) perceptive reading, 2) selective reading, 3) interactive reading and 4) extensive reading (See the *processes* identified in Khalifa and Weir's model).

In perceptive reading the reader attends to the components of smaller stretches of discourse such as letter, words, punctuation, and other graphemic symbols. In selective reading, the reader attends to grammatical or discourse features of language within a very short stretch of text. Interactive reading, on the other hand, focuses on identifying lexical, symbolic, grammatical, and discourse features within texts of moderately short length with the objective of retaining the information that is processed. On tests, this dimension of reading is usually tapped by tasks such as the *True, False, Not Given*, the *Diagram Completion*, and the *Multiple-Choice*.

Finally, extensive reading involves reading larger stretches of text in professional articles, essays, technical reports and books. This dimension of reading best characterizes some of the main features of the domain of academic reading where readers engage in reading, discussing, reporting, presenting, and using several academic sources such as books, book chapters, and research articles within in their disciplines.

As shown in Table 3.1., Brown and Abeywickrama's (2004) classification matches Khalifa and Weir's (2009) micro processes and macro processes as described and ordered in the central core component of their model. Perceptive and selective reading match bottom-up processes such as word recognition, lexical access, syntactic parsing and forming semantic propositions, while interactive reading matches text representation. Extensive reading, on the other hand, matches the highest components in Khalifa and Weir's (2009) model, that is, building mental models and creating textual representations.

Table 3.1.
Brown and Abeywickrama's (2004) classification of reading comprehension dimensions (types)

Brown & Abeywickrama's reading dimensions				Khalifa & Weir: Cognitive processes
Type of reading	Reading material	Focus	Processing level	
Perceptive	Letters, words, sentence level reading	Phoneme-Grapheme relationship, Literacy skills	Bottom-up, form-based	-Word recognition, - Lexical access -Simple phrases -Short sentences
Selective	Short paragraphs	formal aspect: lexical, grammatical, and few comprehension features	Bottom-up form-based and a little meaning-based	-Syntactic parsing -Propositional meaning -Sentence comprehension
Interactive	Few paragraphs, as much as a page with somewhat complex format with charts and graphs	Literal comprehension, at sentence paragraph and short text level	Mix of form-focused and meaning-focused with more emphasis on meaning focus. It tends towards top-down processes	-Literal comprehension at paragraph level
Extensive	More than a page, essay, full article, book chapter, books	Discourse features, main idea, inferencing	Top-down, meaning-based	-Inferencing -Building mental model ---Creating textual representation

In conclusion, the central core component of Khalifa and Weir's model (2009) accommodates the bottom-up micro processes and the top-down macro processes of comprehension. These processes are extensively discussed and empirically supported by the literature on reading comprehension, some of which is discussed above. The central core component does not isolate cognitive processes from other dimensions of reading processes and integrates cognitive processes with the knowledge sources that feed them.

3.4. The Metacognitive component: Metacognitive strategies

Cognitive processes are core to reading comprehension. However, they must be consciously monitored and regulated by readers. In other words, to accomplish a reading task, readers also need to know what to do, when, where, and why. They need to monitor their

cognition and control their reading performance through the application of metacognitive strategies (Stroller and Grabe, 2011). Khalifa and Weir (2009) identified goal setting, goal checking (monitoring), and remediation as key to the management of reading activities. Metacognitive strategies help readers/test takers self-interrogate, self-introspect, and self-interpret their ongoing cognitive processes and make judgments about what they know or do not know to fulfill a task (Phakiti, 2003). They help control and manage different dimensions of reading and reading test performance.

One key function of the metacognitive component, as discussed in the model by Khalifa and Weir is goal setting, which helps readers decide on the reading purpose, the scope, and the speed of reading required. The goal setter makes a distinction between local and global comprehension which influences the type of reading to adopt for accomplishing reading. Features attributed to goal setting are highly relevant to validation research because they theorize some of the basic characteristics of reading comprehension processes. For example, test tasks that tap into local lexical-phrasal levels call for light and less demanding processes while an inferential item requires scrupulous processing of larger portions of the text. Based on the purpose of reading and other factors such as text difficulty and the time available, readers choose to read either carefully or expeditiously (Brown & Abeywickrama, 2004).

The metacognitive component of the Khalifa and Weir model (2009) model conceptualizes reading in a four-cell matrix, with *careful reading* and *expeditious reading* (on the x axis) that can be conducted at both *global* and *local levels* (on the y axis). The careful-expeditious distinction drawn by the model is a significant improvement in theorizing the construct of reading. Most reading models (Bernhardt, 1991; Hoover & Tunmer, 1993; Rayner & Pollatsek, 1989) are all premised on a careful reading model, assuming reading type as constant across reading tasks and ignoring the different purposes of reading. Extending reading types to expeditious reading can explain how readers/test takers may variably read text to manage different types of reading including scanning, skimming, or search reading. According to Khalifa and Weir (2009), careful reading refers to different operations where the reader attempts to extract complete clear meanings within or beyond sentences right up to the level of the entire text, in order to construct the macrostructure text representation. They also argue that careful reading may be used at local or global levels and is operationalized through tasks that require word recognition, analysis of syntax, developing an accurate comprehension of explicit

meanings and drawing inferences. Some of these processes such as word recognition and grammar take place at local level or at global level while developing an accurate comprehension of explicit meaning and making inferences take place at global level. Careful reading at the local level involves processing a text until the basic meaning of a proposition is established whereas careful reading at the global level involves processing the text until its macro-structure is built.

Expeditious reading, on the other hand, involves quick, selective, strategic, and efficient reading to access relevant information. Khalifa and Weir argue that in expeditious reading, the text is not necessarily linearly read and processed. Readers may sample the text in order to extract pieces of relevant information necessary to achieve reading purpose or answer specific test items (See also Urquhart & Weir, 1998). Just like careful reading, expeditious reading can be conducted at global levels (for search reading and skimming) or at local levels (scanning). In *search reading*, the reader processes the text in order to locate information relevant or necessary to achieving reading purpose or answering specific test questions. Skimming and scanning are the best examples of expeditious reading. In *skimming*, the reader avoids details and processes the text in order to obtain the gist of the text. In *scanning*, the reader processes the text selectively by searching for specific words, phrases, dates, etc. and retrieve relevant information for achieving reading purpose or answering a specific test item.

The distinctions drawn by Khalifa and Weir (2009) between careful and expeditious reading and local and global comprehension are very relevant to reading in a test context, where a test taker needs to choose the type of reading required (careful- expeditious) based on the comprehension level required by a test item or task and monitor and adapt speed of reading based on the cognitive demands of the test task and modify the activities if necessary. This component of the model, along with its sub-components, provides a useful framework to examine how different test tasks are processed locally or globally. These distinctions in type of reading and level of processing are highly relevant to assessing reading comprehension, because in a testing context, reading comprehension is assumed to be mostly expeditious, as test takers have a limited amount of time to respond to reading tasks, whereas, in non-test condition such time pressure is typically absent. Depending on the characteristics of the test task or purpose of reading, test takers may also need to read certain parts of the text very carefully. The distinction is also relevant to academic reading, which is assumed to involve largely careful reading, while reading comprehension tests may largely depend on expeditious reading. The literature on

reading models suggest that most current models are premised on careful reading and expeditious reading has been almost totally ignored.

In the section below, cognitive theories of reading and empirical research supporting the metacognitive component of Khalifa and Weir's (2009) model are discussed. Of particular interest are Strategic Competence Theory (Bachman & Palmer, 2010) and Construction-Integration Theory (Kintsch, 1989) and their relevance to the present study.

3.4.1. Metacognitive strategies and reading: Strategic Competence Theory

Khalifa and Weir (2009) modeled reading as a strategic activity where the reader makes use of both cognitive processes and metacognitive strategies to construct text meaning. Reading in general and reading in test-contexts, in particular, involves use of different types of metacognitive strategies. *Strategies*, as “deliberate, planned, intentional, goal-directed and future oriented processing that can be used to accomplish cognitive tasks” (Flavel, 1977, cited in Phakiti, 2003, p. 56), play a crucial role in reading comprehension. They define one of the outstanding features of what successful readers/test takers do in their reading. They use different strategies to develop text representation and mental representation (Phakiti, 2003; Purpura, 1999) and adjust to the reading purpose and the resources available (Mannesa & Hoyerb, 1991; Phakiti, 2003; Purpura, 1999). The strategic component of reading processes and test performance has been linked to the strategic competence component of communicative language ability (Alderson, 2000; Bachman & Palmer, 1996; Douglas, 2000; Purpura, 1999), which is defined as the conscious and deliberate capacity to use appropriate strategies to respond to and accomplish a language task (Phakiti, 2003). It is conceptualized as the capacity to check, plan, select, regulate, and monitor cognitive processes (Bachman and Palmer, 1996; Phakiti, 2003). As a higher order, executive, conscious process, it monitors and regulates cognitive processes and provides a cognitive management function to complete a language task. Khalifa and Weir (2009) share the same perspective on metacognition in their model.

Metacognitive strategy use is not limited to non-test reading where readers read a text linearly to develop the gist. It is perhaps most relevant to reading in testing contexts where test takers face many challenges in doing a test task and need to use metacognitive reading strategies to process the text and the test tasks, often under pressure of time. Test takers need to decide how to proceed and may choose to read the text first or focus on the test items before processing the text. They need to know how to manage uncertain answers or the limited time that is available to

them. Metacognitive strategies help them exert more control over different aspects of their test performance. However, not all metacognitive strategies used in test performance are related to the test construct. Cohen and Upton (2007) make a distinction between strategies that are related to the test construct and those that are not, i.e., test management strategies and test wiseness strategies. *Test wiseness* strategies, unlike test management strategies, are techniques that test takers use to get an answer without necessarily processing the text for an answer. They provide shortcuts to test takers to get to the answer. Examining the type of strategies used across test tasks can reveal valuable information about how test takers interact with the text and the task and reveal the influence of the strategy they choose on their comprehension and task performance. In fact, as research has indicated, strategy use influences text comprehension and test performance (Cohen & Upton, 2007; Rupp, Ferne & Choi, 2006). Based on their individual capacities test takers may use different strategies for the same text or test task. Use of strategies in reading comprehension and reading test performance has a long history and it has been extensively studied in test validation of different language tests (Cohen 2006; Cohen & Upton, 2007). Strategy use can explain some of the variation in test takers' performance (Bachman & Palmer, 1996; Douglas, 2000; Purpura, 1999). Results of test takers' performance on reading tests have indicated that more successful test takers make more frequent use of metacognitive strategies in text processing (Alderson, 2000; Douglas, 2000).

3.4.2. Metacognitive strategies and reading: Construction-Integration Theory

One of the reading comprehension theories that has discussed the role of metacognitive strategies and lends support to the metacognitive component of the Khalifa and Weir's (2009) model is Construction-Integration (hereafter, CI) theory which is briefly discussed below.

CI is basically a text comprehension model which adopts a *constructivist perspective* towards reading as a process of social construction of meaning. It holds that the reader interacts with a text and with the author to construct the meaning (Stanovich, 1994; Donaldson, 2008). Reading is seen as a socio-cognitive process (Bloome, 1985) and the outcome of the interaction between cognitive processes and social processes. It assumes reading as a cultural activity that occurs in a social context. Cognitive constructionists highlight the cognitive and individual construction of knowledge (Piaget, 1970) while the social constructionists pay more attention to the role of environment and social interactions in skill development and learning (Vygotsky, 1978). This constructivist stance to reading assumes that meaning making goes beyond simply

knowing the meanings of words and combining these words in grammatical sentences. Rather, meaning construction of texts is assumed to include background knowledge that readers bring to the text that is both internally formulated and socially constructed (Kaufman, 2004; Donaldson, 2008; Stanovich, 1994). Viewed from such a social process perspective, reading is a cultural activity that happens in a social context. Such a constructivist stance to reading can help understand how cognitive and social processes interact when a reader attempts to construct the meaning from a text in a non-test situation, or complete a test task in a test situation.

CI theory was suggested by Kintsch (1988, 1998) and van Dijk and Kintsch, 1983) and can be assessed as a general theory of knowledge which explains how knowledge is constructed and represented in mind. The keyword to the theory is “knowledge” which includes linguistic knowledge of words, grammar, etc. and non-linguistic world knowledge. It provides an account of how different types of knowledge can influence the reading processes. It emerged from empirical research of reading comprehension which closely studied the role of human memory and recall in the cognitive processes of text comprehension (Kintsch, 1988, 1998, 2004; van Dijk and Kintsch, 1983; Albrecht and O’Brien, 1993; Gernsbacher, 1995; and Rumelhart and McClelland, 1986). The theory, as its name implies, consists of two phases; 1) construction of the text representation and 2) integration of the text representation into the knowledge structure of the reader.

More specifically, the theory suggests several steps in the comprehension process. First, micro-level propositional meaning of the linguistic units of words and phrases are constructed. In the next step, these propositions activate similar or associate propositions in the knowledge net to form coherent representations. Then, the semantic net formed is revised and elaborated by inference processes. In the construction phase the reader constructs the propositional meaning from the linguistic units of words, phrases, and sentences and develops a textbase. The textbase is then integrated with the general knowledge of the reader which provides another context for comprehension of the text. This step is influenced and constrained by the knowledge base of the reader. In the final step, the textbase is organized by assigning values to the different concepts and propositions it contains. The micro-level propositions are interconnected to each other at micro level to form a network of local meaning relationships and at macro level to form a network of propositions at macro level of text structure. Kintsch (1988, 1998, 2004) and van Dijk and Kintsch (1983) have also elaborated on the integration component of the model

characterizing it as a fine-tuning process that automatically occurs at all levels of comprehension process. It occurs in short iterative cycles where a network of text relations is constructed, processed, and integrated with what the reader has retained from the previous cycle. The steps start construction of micro-level text representation and move up to macro-level (integrated) mental representation. They emphasize the automatic nature of the integration process and its role in reading of a text. Failure in integration process forces the reader to engage in strategic problem-solving reading.

CI theory has informed the work of many reading researchers and has proved to be relevant in explaining some of the reading processes involved in comprehension. Research findings have lent support to the theory and have shown that reading involves textbase creation and knowledge-based interpretation of the text (e.g., Carrell, 1992; Horiba et al., 1993) which depend on background knowledge, readers' language knowledge (proficiency), and topical knowledge (e.g., Barry & Lazarte, 1995, 1998; Hammadou, 1991).

Construction-Integration Theory is relevant to the current study in several ways. First, it can explain the complex processes and strategies that test takers use in comprehending texts and developing text and/or mental representation. Second, as one component of the cognitive process of reading is assumed to be linguistic and topical knowledge, the theory can describe the influence of linguistic knowledge and background knowledge on comprehension processes used by the test takers. Third, the theory might also explain some of the possible differences observed across test takers who are at different levels of language knowledge and proficiency. Fourth, in light of the theory, possible differences in background knowledge of the test takers and its influence on their comprehension processes can be better explained. Fifth, Construction-Integration theory is suggested for reading under non-test conditions, but the theory can be applied to reading in testing context. Most important of all, the theory can help in discussing the role of the knowledge sources, linguistic and non-linguistic, in test takers' performance when they search for answers to the IELTS RCM test items. In brief, CI theory can be also seen as part of the constructivist approach, which views reading as a complex process of social construction of meaning from print and provides an underpinning for the understanding of the reading process. It informed this study and helped to explain how readers interact with texts to construct meaning.

In conclusion, the metacognitive component of Khalifa and Weir's (2009) model, Strategic Competence Theory (Bachman and Palmer, 2010), and Construction-Integration Theory (Kintsch, 1989) were relevant to the current study as they provided a framework for consideration of the construct operationalized by the IELTS RCM's test tasks in test takers' responses to the test (e.g., reading, searching, answering).

As the IELTS RCM consists of three different passages and nine different test tasks, it was expected that test takers used different strategies to: 1) engage with the text and each test task; 2) understand the text and the expectation of the test tasks and test items, and 3) find the answer or select an answer from the options given. Overall, examining the strategies used in the test process revealed useful information about the construct of the test, and strategic competence served as a reference in explaining strategy use in test performance.

3.5. Knowledge base component: Knowledge sources

The third component of Khalifa and Weir's (2009) model is the *Knowledge base* component, which is assumed to supply whatever knowledge is required to facilitate formal processes and the discourse processes of comprehension. Each cognitive process calls for certain types of linguistic or non-linguistic knowledge. For example, lexical access requires formal, orthographical, phonological and morphological knowledge of lexicon which are stored in the long-term memory. These components of lexicon are recalled and accessed during the lexical access process. Likewise, for the syntactic parsing process, which helps readers connect words and form a syntactic unit, grammatical knowledge provides a source of syntactic categories to the reader to successfully achieve meaning through the application of this process. As cognitive processes proceed other types of knowledge sources are summoned. For example, higher levels of processing such as inferencing require topical knowledge, while building a mental model requires both topical knowledge and textual knowledge. As stated in the research questions, one key question regarding the validity of the inferences drawn from the IELTS RCM was about the knowledge sources used by test takers during test performance. Khalifa and Weir's (2009) cognitive model of comprehension informed consideration of this dimension of the test construct.

Part of the knowledge base component of the Khalifa and Weir model could be discussed in relation to Schema theory, which explains how reading and topical knowledge interact. In the section below, Schema theory is briefly discussed.

3.5.1. The Knowledge base component and Schema Theory

Discussions of reading comprehension have consistently included knowledge of topic and how this might influence comprehension processes. Research findings have also shown that there are indications from response times of readers to associated concepts that support the role of background knowledge in comprehension (McVee, Dunsmore, & Gavelek, 2005). To describe and explain the role of background knowledge in reading processes, reading researchers have appealed to Schema Theory which has been used as a shorthand reference to generalized knowledge representation in memory. It views readers' knowledge as an integral part of the comprehension process (Swales, 1990; and Wallace, 1992). In the 1980s and 1990s, it was basically accepted as a model of background knowledge in comprehension (Anderson & Pearson, 1984). The theory has been used to explain how readers use knowledge frameworks from memory during comprehension processes. The theory assumes that new information and experiences are learned quickly by activating relevant information from memory (Bartlett, 1932; Cook, 1997), which can explain how knowledge of topic and general knowledge interact with reading comprehension processes. The theory also assumes that recall of information takes place in the context of past experience and relevant information that is stored in long term memory (Nassaji, 2002). Grabe and Stroller (2011) argued that since reading involves topic and text type identification, readers activate and recall relevant information in the memory and use it in processing and comprehending the text. The theory can explain reading texts and test tasks that might involve recall of such information from memory to integrate with the information presented in the text. Schema Theory is also related to inferencing as one basic reading process where readers use their general knowledge structure to draw knowledge-based inferences and use inferencing as a process to categorize and conceptualize text information.

However, in contemporary reading research, Schema Theory is not seen as the best explanation for reading process because it provides explanation only for background knowledge in reading, but not for other types of knowledge (e.g., Pressley, 2006; Snowling & Humle, 2005; Stanovich, 2000). The theory has also been criticized for having multiple definitions and interpretations and being ambiguous and vague. In fact, it seems that recently there has been no serious research applying Schema Theory. Rather, many serious questions have been raised about it: How are schema organized? How large are they? How many are there? Do they change or develop? (e.g., Kintsch, 1998, Nassaji, 2002; Paivio, 2007; Sadoski & Paivio, 2007).

Another criticism, and perhaps the main limitation of the theory, has to do with the fixed structure of Schema Theory (i.e., used in the same manner in all reading situations). In contrast, reading has been better accounted for by dynamic, ever-changing mental representations as hypothesized by Kintsch (1998) and is supported by more recent theories and research. Schema Theory might have had relevance to the current study by showing the degree to which test task performance relied on topical knowledge. It could also have been used to explore evidence of topical knowledge required by the RCM and assumed as part of general knowledge by the test (Alderson, 2000). In sum, Schema Theory can be assessed as a general theory of knowledge that describes how knowledge is constructed and represented. It can provide a useful ground for understanding the role of topical knowledge in reading. However, to better understand how readers process and comprehend texts used in reading assessments, Construction-Integration Theory offers a more encompassing account of the role of knowledge and knowledge-based processes in L2 reading comprehension. Contrary to Schema Theory, which focuses on topical knowledge, Khalifa and Weir's (2009) model addresses the role of all types of knowledge in reading processes, both linguistic and non-linguistic,

To conclude, this first part of Chapter Three has focused on Khalifa and Weir's (2009) reading comprehension model, I have explained that the model covers all the cognitive processes of reading which include word recognition, syntactic parsing, developing semantic and text representation, as well as the metacognitive strategies used for managing and monitoring comprehension. These components can address the main processes of reading comprehension. Other reading theories have tended to focus on one single dimension of reading. For instance, Schema Theory focuses on the role of background knowledge in reading comprehension, whereas Construction-Integration Theory emphasizes the role of world knowledge in the comprehension process. In my view, Khalifa and Weir's (2009) model better integrates all the main components of reading comprehension, i.e., knowledge sources, metacognitive strategies, and cognitive processes, in one model and accounts for how these components dynamically interact and influence reading comprehension. Based on the details of the model and the distinctions made in each of the three components, the model had more explanatory power to account for reading comprehension in test and non-test situations. Details of the model are further discussed in Section 3.6 below.

In brief, the Khalifa and Weir (2009) model was best suited to the research purpose of the current study. It provided a framework to study individual differences in the processing of more proficient (more successful) versus less proficient (less successful) test takers, across first language (L1) English and second language English (L2) readers. The model helped frame assumptions about the test content that were of interest to the validity argument (Embretson & Gorin, 2001; Gorin, 2006; Yang & Embretson, 2007).

3.6. Validation model of IELTS RCM construct

Building on Khalifa and Weir's (2009) model and Gorin's (2006) process model, a construct model for the IELTS RCM was adopted. As shown in Figure 3.3, two strands of theories were used in the validation model adopted for the study. They provided a comprehensive framework to examine the cognitive dimensions of the reading comprehension construct operationalized by the IELTS RCM. Gorin's (2006) Process-Oriented Model of Validation, which is itself informed by Messick's (1989) unified theory of validity, requires multiple sources of validity evidence. Gorin's (2006) model of validation operationalized validity as alignment between Skills, Knowledge, Processes, and Strategies (SKPSs) of the intended and measured construct. The SKPSs provided a practical organizational support for the exploration of what each test task in the IELTS RCM was measuring. Gorin's model was also used in examining the experts' accounts and judgements of the intended SKPSs. Khalifa and Weir's (2009), model with its tripartite components, on the other hand, captures the complexity and multidimensionality of the cognitive processes and metacognitive strategies engaged in reading comprehension.

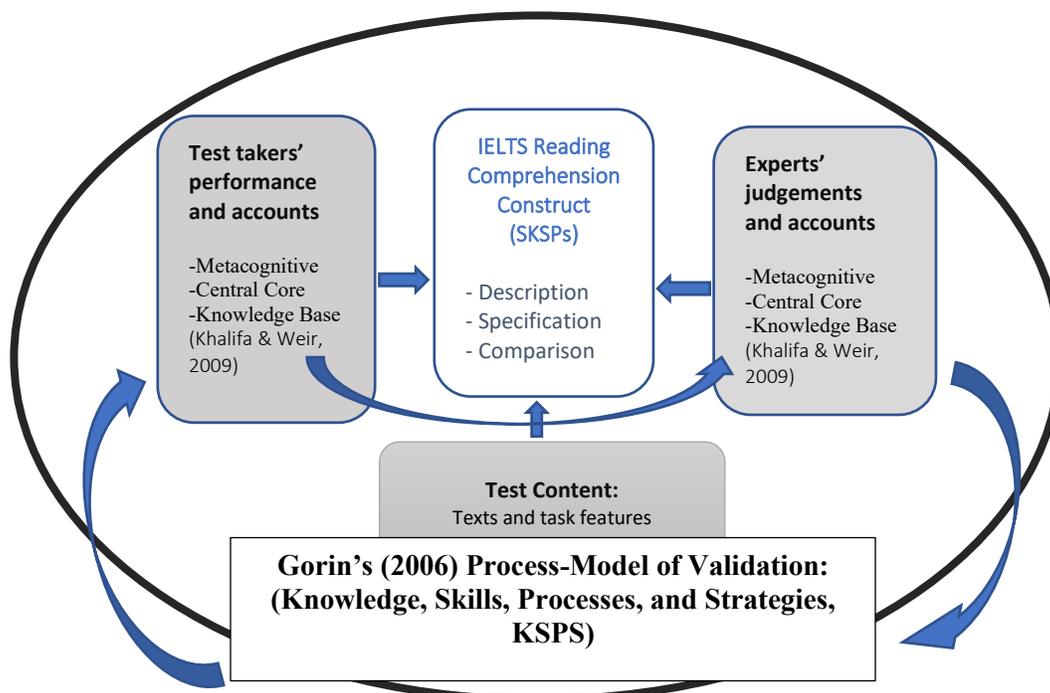


Figure 3.3. Validation model of IELTS RCM (Adapted from Khalifa and Weir, 2009, and Gorin, 2006) used in the present study

It is worth mentioning that the KSPSs component of validity (Ferrara et al.'s 192004) argued by Gorin's (2006) cognitive model of construct validity shares some of the key features included in Khalifa and Weir's (2009) model of reading comprehension. Both models share the cognitive, metacognitive, and knowledge source components as viable sources for examining comprehension processes and reading test performance. In the theoretical framework adopted for this study (Figure 3.3), Khalifa and Weir's (2009) model of reading comprehension was situated within Gorin's validation model, and provided, in my view, a solid cognitive framework for studying the construct of the IELTS RCM. Gorin's (2006) model also justifies use of verbal reports and expert's judgements for closer examination of the KSPSs that defined the construct of the reading comprehension test and the test tasks and items.

Part II: Reading Comprehension Construct: Empirical findings

3.7. Khalifa and Weir's (2009) Model and Assessing Reading

Khalifa and Weir's (2009) reading model has informed many validation studies of large scale, high stakes reading tests which examined the type of reading and level of processing in test task performance, but only a few studies have examined the type and level of processing elicited by the IELTS RCM. At this juncture I wish to review some validation studies that adopted the model in studying construct validity of reading comprehension tests including the IELTS RCM.

Weir and Khalifa (2008b) published an early version of their model in their discussion of the reading section of the Cambridge ESOL (English for Speakers of Other Languages), identifying the variety and complexity of reading comprehension required as a result of the cognitive demands of the test tasks. This helped them decide if the cognitive processes engaged by test takers in responding to the test tasks were representative of the cognitive processes used in reading in the target non-test contexts. Their findings indicated that test tasks were cognitively demanding for less skilled readers who struggled with answering the textually implicit questions, while more skilled test takers could answer them easily. Based on the type of reading used and the cognitive demand of the test tasks they concluded that the test (Main Suite Reading papers) demonstrated construct validity. They argued that task complexity and task demands contributed to the construct validity of the reading test. Their study highlighted the role of the level of processing and the type of processing in considerations of task complexity.

Sarajani and Krishnan (2011) studied the type of reading involved in different item types in the IELTS RCM by cross-examining them in relation to actual academic reading tasks. In their study, two informed test takers analyzed a sample of 14 IELTS reading tests (560 items) for type and level of processing. They found 77% of the items in the IELTS RCM tests involved careful reading, i.e., for meaning (See Section 3.4 for a discussion of careful reading) and only 23% of the items involved expeditious reading, e.g., skimming, scanning (See Section 3.4). which indicated the items are imbalanced in terms of the types of reading involved. Results of item analysis also showed that 70% of the items focused on basic comprehension, 71% of the items tapped local sentence level comprehension, while only less than 1% (7 items) tapped text model representation, which required test takers' comprehension of main ideas and supporting details to

form a macro-structure of the text. In other words, local level items required the test takers to focus on micro-propositions and a quick search of some sections of the text (Weir, et al., 2008; & Sarojani & Krishnan, 2011). Global level items, on the other hand, required the test takers to focus on text macro-propositions by incorporating their topical knowledge and text information. No items tested comprehension at the global or *situation model level*, that is, to build a representation of the content in a text and relate it to mental models of real-life situations through the use of background knowledge. They concluded that the IELTS RCM tests included very few items testing at the higher level of cognitive processing. Their finding raised some serious questions as to whether the IELTS RCM provided a valid representation of academic reading comprehension.

Weir, et al. (2008a) also examined the academic claim of the IELTS RCM test. A sample of 352 students took one of six IELTS test tasks and were asked to fill in a retrospective questionnaire immediately after answering each one. The questionnaire included questions about the type of reading (careful versus expeditious) and level of processing (local versus global). They used descriptive statistics for analysis of the data. Results of data analysis indicated; (a) the major focus of the IELTS was on careful reading; (b) for most participants across the different task types, expeditious reading was the type of reading chosen to answer the questions. These two findings suggest that both expeditious reading and careful reading were used in test performance. Other findings were: (c) most test takers reported finding the information necessary to respond to questions by putting information together across sentences; and (d) there was some evidence of positive relationship between the choice of certain strategies and success on various items. On the basis of these findings, Weir et al. (2008a) concluded that the IELTS RCM demonstrated construct validity as there was a match (to a degree) between the test and general academic reading in terms of type of reading and strategies used.

The same questions were asked about the academic claims of the IELTS RCM in another study conducted by Weir, et al., (2008b), this time with focus on the reading experiences of first-year undergraduate students. Results of this phase of the study was then compared and contrasted with views and judgements of a sample of content instructors who were experts from diverse university disciplines. They applied a descriptive framework of expeditious and careful reading strategies to each item in each testlet. Expert judges reported that the major focus of the IELTS test was on careful reading, while the students' survey data indicated that for academic study,

they needed expeditious skills and strategies in their academic reading. They also found that individual characteristics of IELTS texts did not match those typically identified with academic texts. They also raised questions about the validity of the IELTS RCM as a representation of the academic reading construct.

Other studies of high stakes reading tests used the Khalifa and Weir's (2009) model to examine the type of reading and level of processing in test task performance. Katalayi (2018) adopted the model and studied the construct validity of the reading section of the English State Examination which served as a certificate for university admission in the Democratic Republic of the Congo. A large sample of 496 Grade 12 students took a sample test which consisted of 50 multiple choice items. They also recorded their strategy use during reading test performance by filling in a strategy questionnaire. The strategies covered type of reading (careful versus expeditious) and level of processing (global versus local). Results showed most test items required higher level processing rather than low level processing, and were, as a result, more difficult than lower level items. Further findings indicated lower level items required reading at sentence and/or paragraph level while higher level items required reading at text level. He concluded that the findings did not provide evidence for context validity of the test, and that the test had weak construct validity because the test tasks were not indicative of reading in the actual academic context.

These studies have indicated differences in the type of reading (careful-expeditious) and level of processing (global-local) required by test items. These studies suggested that an important component of establishing construct validity of reading tests is examining the types of reading and the process levels involved in test performance. These two features of processing also contributed to the complexity and difficulty of test tasks, and inspired examination of type of reading and level of processing across different test tasks and test takers in the present study.

3.8. Construct Validity of the IELTS RCM: Evidence from Empirical Research

Based on the available literature, few serious studies have examined the contextual parameters and cognitive processing involved in academic reading (Weir 2005). Among the two most popular high-stakes academic language proficiency tests of English, IELTS has not been as extensively studied as, for example, the TOFEL. Further, unlike the Speaking Module and the Writing Module of the academic IELTS, which has been extensively studied (Barkaoui, 2016;

Belahouel, 2018; Chan, Bax, and Weir, 2017; Knoch, May, Macqueen, Pill, and Storch, 2016; Nakatsuhara, Inoue, and Taylor, 2017; Roothoof and Breeze, 2019; Seedhouse and Morales, 2017; Yu, He, and Isaacs, 2017; Yu, Rea-Dickins, and Kiely, 2012; Wray & Pegg, 2009; to cite only a few), there are very few studies that have explicitly focused on the construct validity of the Reading Comprehension Module. Those few studies, to my knowledge, were all sponsored by the IELTS research program which started in 1995 and mainly focused on the comparison of the task requirements of the actual academic domain in comparison with its operational representation in the IELTS RCM test tasks (Weir, Hawkey, Green, Ünal, and Devi, 2009; Moore, Morton, and Price, 2012; Weir, Hawkey, Green, and Devi, 2012, & Green and Hawkey, 2009). These studies consisted of two or more phases but only one phase of the research focused on the cognitive processes test takers used in their performance. There is no study, it seems, that has focused on the construct validity of each of the nine different test tasks, nor is there any research that has focused on the construct validity of the test as a whole.

In this section, the relevant studies available in the literature on the construct validity of the IELTS RCM, which motivated the current study are briefly reviewed, but an important point needs to be made about the distinction between reading skill and reading construct. Most research studies on the construct of reading have relied on taxonomies that divide reading into a variety of skills and sub-skills (Alderson, 2000, Grabe, 2012; Moor, Morton, and Price 2009). For example, Munby (1978) distinguished 266 skills, sub-categorized under 54 main groups. These taxonomies were then synthesized by other researchers into a more manageable set of categories. For instance, Craver (1997) suggested five basic skills; scanning, skimming, reading, learning and memorizing. He defined reading as *normal* reading of something that is easy to comprehend. Grabe and Stoller (2011) captured reading skills under seven main categories which included reading to search for simple information, reading to skim quickly, reading to learn from the text, reading to integrate information, reading to write, reading to critique the text, and reading to comprehend the general idea. In another taxonomy used in the study of the TOEFL reading test, four main reading skills were examined, namely, reading to: find information, have general comprehension, learn, and integrate information across texts (Enright, et al. 2000). These four categories seem to be more relevant to reading in academic contexts, capturing the basic reading activities student learners are engaged with while the discrete skills model seems to be more appropriate for teaching reading skills at school level. During the last

four decades, there has been a shift or gradual change from discrete skills conceptualizations towards a more abstract and functional conceptualization of reading comprehension. However, most reading tests, including high stakes language proficiency tests such as IELTS, are still adhering to the skills model.

One of the studies relevant to the construct of reading comprehension was conducted by Weir, Hawkey, Green, and Devi (2009) who compared test performance of three groups of test takers at three levels of reading proficiency to identify: 1) the type of reading, and 2) the strategy used in IELTS reading test tasks. The type of reading was defined as careful reading of the text versus selective expeditious reading, and the response strategy was defined as the level or unit in which the test takers found the necessary information for answering each item. These units consisted of sentence, several sentences, and text. A sample of 325 students at three levels of English language proficiency participated in the study. Each of the participants took only one reading section of the tests. After doing the test items, they filled in a questionnaire to identify the type of reading they used (careful reading versus expeditious or selective reading) and the level at which they found the necessary information for answering the item (Strategy). Findings of the study showed that selective expeditious reading played a more important role in answering the test items and most participants used selective expeditious reading across different task types. As to the strategy use, results showed that participant test takers put information together across sentences, suggesting that sentence level comprehension dominated the test. However, no straightforward relationship was found between task type and strategy type which indicated that task type was not a reliable predictor of patterns of strategy use across groups. The study questioned the validity of the inferences drawn from the test because it provided only partial evidence of academic processes involved in academic reading comprehension.

This study was based on Weir and Urquhart's (1998) earlier study, which had conceptualized reading processes in terms of type of reading and level of processing. Instead of examining a long list of discrete skills, they took one step further and focused on two dimensions of difference: reading level (global reading versus local reading) and reading type (careful reading versus expeditious reading). The taxonomy they adopted enjoyed the advantage of being more dynamic and capable of generating a range of reading modes (Moor, Morton, and Price, 2009). Findings of this causal-comparative study have relevance in discussing construct validity of the IELTS RCM, but they do not provide a thorough picture of the construct of the whole test.

The study suffers from some drawbacks. First, the IELTS test was used both as a measure of reading proficiency and as the research instrument, which is not the best way to choose three levels of reading proficiency. Second, the test was not administered in its entirety and each participant took only one section of the test. This hurts the contextual validity of test performance and test results. Breaking down the test content into smaller segments of texts and tasks does not reflect the test construct in its entirety. The methodology used was more practical, but it added some extraneous variables into test performance processes. Third, each test taker answered one item and then filled in the questionnaire to describe his performance, which disrupted natural processing of the test task and calls into question the authentic validity of data collection and the test results. Test performance under such unnatural circumstances cannot reflect the normal test performance and endangers the validity of the data. Overall, the selective nature of their study, focused on few variables, and the unnatural and under-representative context of data collection restricts its potential for exploring the reading construct of the test as a whole. Nor does it allow the researchers to explore all of all the skills, knowledge sources, processes, and strategies that test takers may use across test tasks. Developing a broader picture of the construct of the test requires adopting an exploratory holistic design that is more comprehensive for examination of the test construct.

Another highly relevant study to the construct validity of the IELTS RCM was conducted by Weir, Hawkey, Green, Unaldi, and Devi (2009) who focused on a cross-comparison of the RCM test tasks with actual academic reading activities. They examined the cognitive processing requirements and problems experienced by students in their academic reading and reading for assignments. The study consisted of four different phases. Only the third phase was directly relevant to some cognitive aspects of the RCM IELTS construct. They examined the type of expeditious reading (skimming, scanning, and search reading) and strategy used in answering questions in terms of the level at which they found the answer. Two EAL (English as an alternative language) test taker analysts as informed participants took 14 IELTS RCM samples consisting of 42 texts, total of 560 items. They did each test task and filled in the same questionnaire used in Weir et al. (2009) with the possibility of adding some comments if they felt relevant. The third participant was an English native-speaking test taker (EL1) who did the same test task and recorded self-reported comments of the process he adopted to complete the reading test tasks. Results of data analysis showed a preponderance of careful reading (77%) over

selective expeditious reading and sentence level answer over other levels. They compared results of this phase of their study with results of the first phase. In Phase 1 of their study they had found that academic reading tasks involved more expeditious reading. Based on the findings, the researchers advised some modifications in the test for better representation of the academic reading construct. They recommended adding test tasks and items that involve processing text beyond sentence level.

The study provided a context to compare reading comprehension in two inter-related domains; the academic domain and test domain which operationalizes it. However, the study focused on only two aspects of test performance, i.e., type of reading and types of strategy used. Type of strategy was very narrowly defined as level of reading and was not used in its general cognitive sense. Furthermore, the context in which the IELTS RCM was administered for data collection defies the authenticity of natural test performance. Instead of taking one single sample within the time limits of the test, the three participants took a large sample of 42 texts and they had to stop test performance for each item they answered and fill in a questionnaire and comment on their test performance. Such data collection hinders normal processing of test tasks and items and limits the meaningfulness of the findings. Further, the study focused on few features of test performance and did not address all cognitive dimensions involved in test performance.

Moore, Morton, and Price (2009) investigated the construct validity of the academic IELTS RCM by comparing the reading requirements of IELTS test items with requirements of reading in university courses. Building on McNamara's (1999) three features for the validity evaluation of reading comprehension tests, i.e., task stimulus, task processes, and task demands, they addressed the limitations of validation work on task demands in reading tests for different disciplines. The qualitative study used a mix of content analysis of the IELTS RCM and interviews with academic staff to survey their perception of academic reading comprehension. The data for content analysis of the test consisted of seven IELTS RCM test task types compiled from 13 complete tests which were analyzed based on a framework derived from Urquhart and Weir's (1998) reading model. The framework consisted of two levels of analysis; 1) level of text with which a reader needs to engage to respond to a task (local versus global) and 2) type of engagement which referred to the way(s) a reader needs to engage with texts to answer the test task which involved literal and interpretative inferential meanings. In addition, interview with academic staff was used to analyze the types of reading tasks required in studying academic

courses in 12 disciplines. Results of the study showed that the majority of the IELTS RCM test tasks involved local-literal engagement while a sizeable proportion of the academic tasks required more global-interpretive engagement. Overall, results showed partial agreement of the requirement for actual academic domain tasks and test task domain. Comparison of the two domains demonstrated fair degree of difference.

This study was outstanding for moving beyond literal comprehension and examining both literal and inferential comprehension involved in the texts. The study is unique in adopting content analysis of the test as a source of insights and ideas in the study of the IELTS RCM construct. The study is also unique for incorporating two sources of validity evidence, i.e., analysis of test content, and experts' views and judgements in examining reading construct but the analysis narrowly focused on just two specific aspects of test performance, i.e., type of engagement and level of engagement which do not reflect the full range of activities, skills, processes, strategies, and knowledge sources involved in test performance.

On quite a different plane of research, Green and Hawkey (2012) investigated the validity of the test by examining the process of developing the IELTS; item specification, text selection, item writing, and editing processes. The main objective of the study was to examine the steps item writers take in selecting texts and generating items based on test specifications and their conception of reading construct. Informed by Weir's (2005) socio-cognitive test validation framework and Salisbury's (2005) three-phase framework of item writing (i.e., exploratory phase, concerted phase, and refining phase), they used retrospective reports and direct observation to compare the similarities and differences of experienced and non-experienced item writers in constructing the IELTS RCM test items on the basis of the test specifications. They sampled four trained IELTS item writers and three untrained item writers who took an open-ended questionnaire followed by two interviews; one individual interview and one focus group interview. The interviews focused on issues such as text selection, text adaptation for the test, and item generation. Focus group interviews were concerned with how different item writers constructed test items in line with the requirements of the test specifications. Findings indicated that experienced item writers generated test materials of higher quality and reported their experience of item writing more explicitly than inexperienced writers, though both groups seemed to pass through similar steps in item construction. These findings highlight the importance of item writer training and guidelines for item writing. The study also identified

some of the reading skills that item writers needed to consider in writing the test items, but they were no more than general labels such as understanding the gist, the main idea/themes, and specific information and provided no information regarding what the construct of the IELTS RCM is. This clearly indicates that IELTS RCM adopts a skill-based approach in defining test construct.

Recently, eye tracking has been used as a new addition to the methodology toolbox of validity research. It can provide more precise information on test takers' behaviour by capturing their eye movements during reading and test performance. Bax (2013) examined successful and less successful test takers' cognitive processes during onscreen IELTS reading test items. Using eye tracking data, he collected data from 38 EFL readers' eye movements to examine the cognitive validity of reading test items. Participants' took the test along with a questionnaire and stimulated recall interview data to comment on and describe their eye movements' recordings. After they took the test, they explained what they were doing when they were looking at different parts of the test task. However, the study reported results for only 11 items of the test but not the whole test. The findings indicated that, for some items, there were significant differences between successful and unsuccessful test-takers on their ability to read expeditiously, their focus on particular aspects of the test items, and reading the texts. For other items no significant difference was observed between the two groups of test takers. Bax recommended use of eye tracking, in combination with post-hoc interview and questionnaire data which can provide new insights into the cognitive processes relating to the reading construct.

His study provided a novel and precise perspective to the study of reading comprehension construct. However, in terms of validity evidence there were some limitations to his study. First, the study provided some details about on-screen reading behavior of the participants. However, it had limited and narrow focus on a few aspects of processing such as visits and fixations and did not provide a full picture of the processes to evaluate the construct of the IELTS RCM. Second, the study promised to integrate eye-tracking data with retrospection stimulated recall to explain some of the differences observed, but there was little in the study to show how the two strands of data were integrated. Third, the data reported in the study were limited to few test tasks and did not include all the test tasks used in the IELTS RCM. Examining the construct validity of the IELTS RCM needs to cover the full range of test tasks. Finally, and the most importantly, the study provided useful data *about* what the test takers did in their performance in terms of eyes

movements measures, but it does not talk about *the actual processes* they used. Such eye tracking data by themselves cannot provide the kind of data needed to explore construct validity of reading comprehension tests unless they are integrated with test takers' comments and descriptions and analyzed in light of a resourceful theoretical framework that can explain the cognitive processes used. However, eye-tracking data provides accurate and very precise means to collect the reading behavior of test takers that cannot be collected otherwise to inform the construct of the test. In sum, the study does not provide sufficient evidence needed to examine construct validity of the whole test as it was very limited in the kind and amount of data related to the cognitive processing used in test performance.

As mentioned earlier in Chapter One, these studies indicated that the focus of the IELTS RCM was basically on local literal comprehension and the test did not represent global comprehension which is an integral component of academic reading. Nor did it tap into more inferential dimension of comprehension. Findings also indicated that most test tasks tended to involve careful reading and not much expeditious reading. These studies suffered some limitations and presented some gaps that invited further research. Most of the studies relied on one single source of validity evidence and used questionnaire as their main instrument of data collection. Furthermore, none of these studies addressed the full range of cognitive processes used in test performance. Nor did they focus on the sub-constructs of the test which are represented by each of the nine test tasks. The present study, in fact, was an attempt to address some of these gaps and limitations and provide more diverse sources of evidence for the construct of the IELTS RCM.

3.9. Cognitive Processes and Metacognitive Strategies in Reading Comprehension

Many reading scholars have underscored the role of cognitive processes and metacognitive strategies in L2 comprehension (Bernhardt, 2010; Cohen, and Upton, 2007; Grabe, 2005, 2009). Cognitive processes refer to actions that readers directly take towards the text to comprehend the text better. Examples of cognitive processes include word recognition, syntactic parsing, using context clues, developing text representation and so on. Sheorey and Mokhtari (2001) defined metacognitive reading strategies as “intentional, carefully planned techniques by which learners monitor or manage their reading” (p. 436) and include strategies such as using prior knowledge, guessing word meaning from context, evaluating what is read,

noting text characteristics, and predicting and confirming predictions. Reading researchers have identified different types of metacognitive strategies used during reading. Mokhtari and Reichard (2002), for example, suggested three types of metacognitive strategies; 1) global strategies, 2) problem-solving strategies, and 3) support strategies. They conceptualized global strategies help readers prepare for reading by setting a purpose, previewing text characteristics, skimming, predicting, and activating prior knowledge while the problem-solving strategies help readers solve challenges and problems in reading. For instance, re-reading, slowing down, reading aloud, guessing the meaning of a word, and visualizing information in the text can help readers solve some reading problems. Finally, support strategies such as note-taking, or using outside references are used to help readers during their meaning-making attempts.

According to Paris and Winograd (1990) metacognition helps readers in self-appraisal and self-management of their cognition. Through self-appraisal, readers signify their individual images about their own knowledge capacities and their affective conditions in terms of their knowledge, capabilities, and motivation as readers. Self-management, on the other hand, is related to the mental procedures used during the coordinating facets of problem solving. It helps readers in different phases of task performance to plan before the task, modify during the task, and revise after the task. Consistent with the theoretical view that supports the role of cognitive and metacognitive strategies on reading, Khalifa and Weir (2009), and Pang (2010) argued that metacognitive ability helps readers control what they are doing. According to Klein, Peterson and Simington (1991) good readers make use of several metacognitive strategies such as identifying purpose of reading and text type, predicting meaning while reading, and summarizing what they read. Through use of these strategies, readers control the full spectrum of reading processes and achieve reading and comprehension of the text and the task. Readers with higher cognitive and metacognitive capacity are capable of managing these multiple processes of text comprehension, while readers who lack these cognitive capacities do reading as a verbatim task and fail to attend to the multi-layers of text processing (Kim, Crossley, and Skalicky, 2018; Tsai, Ernst, and Tally, 2010).

Reading researchers have extensively researched the use of metacognitive strategies in different contexts (Ghaith and El-Sanyoura, 2019; Sheorey, & Baboczky, 2008; Sheorey and Mokhtari, 200; Hosseini, 2006; Hong-Nam and Leavell, 2006; Malcolm, 2009; & Park, 2010). These studies were conducted in diverse socio-linguistic contexts including the USA, Iran,

China, Korea, Lebanon, and Bahrain which gives more legitimacy to the findings. Results have indicated metacognitive strategies as key to reading comprehension and managing reading task performance. However, results of studies on the influence of metacognitive strategy use on reading comprehension are not conclusive. While Tavakoli (2014) maintained that readers use support strategies the most, followed by global, and then problem-solving strategies, other studies reported more frequent use of problem-solving strategies (Sheorey, & Baboczky, 2008; Hong-Nam & Leavell, 2006). In another study of the role of metacognitive strategies in reading comprehension, Ghaith and El-Sanyoura (2019) reported that problem-solving strategies were more frequently used, while the global and the support strategies were moderately used. They also found that problem-solving strategies were significant predictors of both literal and higher-order comprehension. In another study, Huang et al. (2009) studied the effect of metacognitive strategy use in online reading of a web-based reading program. They sampled 30 Taiwanese college students with varying English proficiency levels. They found that support strategies, such as looking words up in the dictionary and translating words, were most frequently used by both the proficient and less-proficient groups. However, global strategies, including previewing the text and keywords, predicting the text, and noting text characteristics were the most powerful predictors of EFL students' comprehension.

In language testing, research has indicated that analysis of strategy use in test performance can provide a rich source of information and reveal what test takers actually do during test performance (Cohen & Upton, 2007; Weir 2005). In a seminal study of strategy use in a high-stakes reading test, Cohen and Upton (2006) studied the construct validity of the reading section of the new TOFEL test as a test of academic reading ability by examining the variable use of reading strategies, test-management strategies, and test-wiseness strategies in the performance of 32 non-native speakers representing four language groups (Chinese, Japanese, Korean, and Other). They assumed if the test is valid as a test of academic reading ability, test takers should actually use academic reading skills and strategies in responding to items, rather than relying on test-wiseness strategies. Results showed that participants used different types of strategies, but the strategies used by the participants were primarily test-management strategies not test-wiseness strategies. Participants also read the text to gain both a local and general understanding of the test passages to answer the test tasks. This lend partial support to the validity claim of the test. Analysis of the metacognitive strategies used for answering different

types of items showed that examinees approached the TOEFL reading section as a test-taking task. This is to suggest that, they actually read the text and tried to comprehend the main points of different sections and the whole text before they could answer the items. However, as participants treated the task like a test-taking task not as a normal text to learn from it, “the test did not *truly* constitute an academic reading task, but rather a test-taking task with academic-like aspects to it.” (p.117) Based on the findings, they also concluded that choice of strategy influenced comprehension of the text and test task performance.

Cohen and Upton’s (2006) finding is inconsistent with other studies (Sheeham & Ginter, 2001; Rupp, Ferne, & Choi, 2006) which found that some test takers successfully used test-wiseness strategies such as wild guessing or elimination of options in their test performance. Their study, in particular, is relevant to the current study in several ways. First, their study adopted a qualitative approach in data collection and analysis by using participants’ retrospective verbal reports. Second, they examined construct validity of an academic reading comprehension section of a high-stakes test which is consistent with the main purpose of the current study. Third, the study highlighted a basic feature of academic reading skill, that is, to “gain both a local and general understanding of the passage” (p.117) which can be used as a yard stick in evaluating academic claims of reading tests. Finally, the study helped identify some strategies that are irrelevant to academic reading construct. For some practical reasons, the study, was limited to the analysis of only 13 predetermined items across the two completed reading test sets and not the whole reading section.

In another study of high stakes reading tests, Saraswati (2015) studied test taking strategies used in a school leaving reading test and examined subject experts’ views and judgements about the strategies and skills tapped by the test. Analysis of the retrospective data of the student participants showed that the test did not measure all the reading skills and strategies it claimed to measure. Results also showed disagreement among subject experts over the strategies measured by the test.

Another aspect of metacognitive strategy in reading comprehension tests is the appropriate use of strategies across different types of test tasks and items. Salehi (2011) analyzed the test taking strategies of 40 EFL test takers. A checklist of 28 strategies was used to record the test takers’ strategy use while taking some reading comprehension items. Results showed match between the use of the right type of strategy for the right type of items, lending support to the

construct validity of the test. Al-Shaye (2002) investigated the effectiveness of two different types of metacognitive strategies on comprehension strategies of 11th-grade students. Comparing use of two metacognitive strategies; the K-W-L Plus (Know, Want to know, Learned) and the SQ3R (Survey, Question, Read, Recite, and Review) strategies in a 46 multiple choice reading test, they reported that appropriate use of strategy across different test items had positive effect on their test performance.

These studies examined some of the strategies that were used during test performance and opened a window to look at what test takers did in their test performance. Based on the findings of these studies, it can be argued that there is consensus over the effect of strategy use on reading comprehension processes. Findings of these studies have implications in discussing the construct validity of the reading test used. In terms of reading constructs, strategy use is an important component of reading processes and test validation studies need to consider them as a key component of reading processes and of test construct. In summing up this section of the review, I wish to mention that findings of these studies provide insights for investigating strategy use as a key component of construct validity of the IELTS RCM. However, focusing on strategies, which are only one cognitive aspect of reading can inflate that aspect (Lin and Yu, 2015) especially if data is collected through self-reports of the participants who may tend to over-report their use. It is vital to consider multiple sources of evidence, i.e., skills, knowledge sources and processes used, in research examining of the validity of inferences drawn from tests of reading comprehension.

3.10. Influence of Readers' Characteristics on Reading Processes

There exists a plethora of factors and variables that can affect reading comprehension. Readers' variables have been long discussed as a key factor in influencing second language learning (Grabe, 2009, 2012, Grabe and Stroller, 2002). Individual variables such as age, gender, language background and affective factors such as motivation, attitude, and emotional investment also influence different dimensions of second language reading (Grabe, 2012). In language testing, test takers' language proficiency, their motivation and familiarity with the test his/her experience with the test, and their attitude towards the test are assumed to impact their performance on the test (Bachman and Palmer, 2010; Cohen, 2006; Nassaji, 2000). Likewise, in

second language reading, readers' characteristics such as language background (L1 versus L2) and proficiency level influence their reading comprehension (Grabe, 2005).

3.10.1. Influence of language background on reading comprehension processes

Extensive research on second language reading has indicated that successful readers enjoy several characteristics, including use of background knowledge of the topic, monitoring reading processes by use of effective cognitive and metacognitive reading strategies, automaticity and fluency in text processing, good vocabulary knowledge, motivation for reading, and a positive reading self-concept, and higher level of language proficiency, (Ghaith and El-Sanyoura, 2019; Grabe and Stoller; 2011; Koda & Zehler, 2008). In second language reading, it is a common belief that language background (L1), levels of language proficiency, and cognitive ability of the readers influence reading comprehension. Second language reading researchers have also argued that L1 and L2 reading processes are basically the same across L1 and L2 (Grabe, 2009, Grabe and Stroller, 2002) and both L1 and L2 readers go through the same main cognitive processes of word recognition, sentence parsing, developing propositions, and developing text or mental representations. These processes are indispensable processes of reading comprehension and cannot be avoided. However, in L2, readers especially the less successful or poor readers are not equally efficient in implementing these processes. For example, several studies have compared word processing differences between L1 and L2 readers and have reported that L2 learners generally access and process lexical items more slowly and less automatically (Diependaele, Lemhofer, & Brysbaert, 2013; Gollan, Montoya, Sera, & Sandoval, 2008; Izura & Ellis, 2004, to cite a few). Research has also shown that due to the lower linguistic knowledge of L2 readers, there is more variation in L2 word processing (Cop, Keuleers, Drieghe, & Duyck, 2015; Whitford & Titone, 2015, 2017). L1 and L2 readers, however, differ in the use of metacognitive strategies. For example, a widely cited strategy study by Sheorey and Mokhtari (2001) examined cognitive and metacognitive strategies in L1 and L2 readers. They reported that high proficiency readers in both groups used more cognitive and metacognitive strategies; however, differences between the native and non-native groups were found in their use of support strategies. High proficiency native readers valued support strategies more than low proficiency readers, whereas all L2 readers valued support strategies regardless of reading ability. These findings resonated with previous research that noted the important role of

metacognitive strategies in L2 reading comprehension (Huang et al., 2009; Sheorey & Mokhtari, 2001; Tsai et al., 2010; Zhang et al., 2008).

3.10.2. Influence of level of language proficiency on reading comprehension processes

Research has also indicated that in addition to language background, level of language proficiency also plays a key role in reading processes. Take for example, Kim, Crossley, and Skalicky (2018) who used simplified texts and authentic texts to examine textual features on word processing times in more proficient and less proficient Spanish-speaking readers of English. They found that for text-level effects, processing times for L2 words in simplified passages for the beginner readers and intermediate readers were significantly faster than those for words in authentic passages, suggesting that word processing is faster in simplified texts. They also found that for more proficient readers processing time was significantly faster than lower proficiency groups. Results also indicated that L2 word processing time decreased as L2 readers encountered the words in each new passage.

Results of these studies indicated that reading strategies of L1 and L2 readers differ in several ways (Li and Yu, 2015; Pritchard & O'Hara, 2008; Sheorey and Mokhtari; 2001; Stevens, Slavin & Farnish, 1991; Stevenson et al., 2007; and Tsai, Ernst and Talley, 2010). Tsai, Ernst and Talley (2010) reported that skilled readers used similar reading strategies in their L1 and L2, whereas the less skilled readers did not show much similarity of strategy use between L1 and L2 reading. The less-skilled readers used more strategies and evaluated the text more frequently when they read in L1 than in L2. Research has also indicated second language proficiency influences reading comprehension. For example, Nassaji (2000) examined higher-level text processing skills in reading comprehension in L2 readers and found that lower-level processes (word recognition) and higher-level processes (syntax and semantic) could contribute significantly to the distinction between good readers and poor readers. These findings resonated with Zhang (2001) who suggested that L2 readers' limited English proficiency can restrict their use of more effective and higher-level metacognitive strategies in their reading. The same results concurred with Barnett (1986) who reported positive relationship between readers' recall and higher vocabulary and grammar knowledge.

Poor readers are not just slower in word recognition process and automatic syntactic parsing processes. Linguistic knowledge was also found to influence inferencing skill in reading. Hammadou (1991) reported that high language proficiency readers did better than low language

proficiency readers in inference drawing and identifying causal relations in their reading, a finding that was later replicated and confirmed in a study by Lu (1999). In another study, Nassaji (2003) reported that good readers differ from poor readers in their word recognition, syntactic parsing and developing semantic representation. Good readers were more efficient and faster in all these reading processes. Results of his study were consistent with similar studies such as Lu (1999) and Hammadou (1991).

Some reading studies have simultaneously addressed the role of both language background and level of language proficiency in strategy use during reading in one single study. In these studies, influence of language background and level of language proficiency on different dimensions of reading processes were examined. In this sub-section, some of these studies are reviewed.

One line of research focused on the relationship between language background/ level of language proficiency and reading strategy use. For example, Brantmeier (2002) noted that the successful L2 readers tended to use more top-down strategies such as integrating information, recognising text structure, using background knowledge, drawing inferences and predicting content. In contrast, the bottom-up strategies such as rereading, translating sentences, looking up unknown words were used more frequently by poor readers. Results suggest that poor readers struggled with the formal language skills such as vocabulary and grammar whereas good readers used meaning-related comprehension skills such as developing a text representation and mental representation. In another study, Li and Yu (2015) examined the effect of language background and language proficiency on reading processes by focusing on the similarities and differences of college level students' reading performance of a TOFEL like reading test (L2) and L1 Chinese reading test (National Matriculation Test, NMT). Participants included EFL Taiwanese college level students at low and high levels of L2 language proficiency. The use of metacognitive strategies was found to be readily similar between English and Chinese. However, use of particular cognitive and support strategies differed across language background and levels of language proficiency level. Results lend support to the argument that that certain cognitive and support reading strategies differed between L1 and L2, suggesting that they might be dependent on language background and proficiency level. Results also showed different profiles of reading strategy use between the more-proficient and less-proficient L2 readers. The proficient L2 readers used more effective and diverse reading strategies directed towards comprehending the

reading content than the less-proficient ones who relied mostly on language-oriented strategies that focused on linguistic elements and local relations in the text. Furthermore, results demonstrated that paraphrasing, asking questions, using context clues and translating strategies served as indicators that distinguished the more-proficient readers from their less-proficient peers. They concluded that reading proficiency was closely related to reading strategy use, and a reciprocal relation between English reading proficiency and reading strategy. The same results concurred with Malcolm (2009) who found that high proficiency and low proficiency students shared similar patterns in strategy use but they differed in use of certain strategies such as translation and cognitive skills. Low proficiency students used more L1 translation strategy, whereas high-proficiency students adjusted their reading rates and used visualization of the information more frequently.

Furthermore, Kong (2006) reported more frequent and diverse strategy use in the L2 reading process. He used verbal reports of the L1 and L2 reading of Chinese students and provided evidence that Chinese adult readers verbalized more strategy use when reading in English than in Chinese. L1 Chinese reading performance was characterized as coming naturally and automatically, whereas their L2 reading of English required more apparent efforts and longer time. In addition, more strategies were consciously employed and reported. It seemed that in L1 reading, they used a well-practiced routine, recognizing words and word meaning, attending to structural units, and interconnecting sentences. Results suggested more automatic processing in L1 reading whereas L2 processing were more effortful and deliberate. In another linguistic context, Stevenson et al. (2007) compared reading strategy use in Dutch (L1) and English (L2) among Dutch high school students and found that for L2 readers, the processes involved a large number of language-oriented processing strategies to compensate for lack of linguistic knowledge or processing skills. They concluded that adult L2 readers focused more on processing the linguistic features of the texts and used more conscious strategies in L2 reading comprehension than they did in L1 reading.

These studies and their findings were relevant to the current study in several ways. First, they justify studying the use of cognitive processes and metacognitive strategies as indispensable components of reading processes. Second, these studies indicated that language background and level of language proficiency influence reading comprehension processes. Results of these studies indicated that depending on the sample of participants, results of comprehension

processes are expected to vary across language background and level of L2 proficiency. By implication, validation of a reading test with native-speaking readers cannot be expected to produce the same results as with L2 readers. So, it is vital to consider the sample of participants used in validation studies.

3.10.3. Readers' attitudes towards the test

Finally, another learner/reader variable relevant to reading processes and test performance which has to do with the social dimension of test performance is test takers' attitude towards the test, which can influence the cognitive and affective dimensions of processing in positive or negative ways. Attitudes of test takers filter their test performance and engage positive and/or negative dimensions of affective variables and impact test performance. Research into the role of attitude in test performance suggests that positive attitude impacts test performance on high stakes tests (Murray, Riazi & Cross, 2012; Rasti, 2009; Han, Dai & Yang, 2004). Negative experiences, on the other hand, "can result in the development of attitudes that erode confidence and potentially impact negatively on test performance" (Murray, Riazi & Cross, 2012, p.577). Murray, Riazi and Cross (2012) investigated test taker attitudes by exploring the opinions, beliefs, and feelings of overseas trained teachers preparing for a professional gate-keeping test. Results of correlations between attitudes and demographic and experiential factors indicated a slight predominance of negative attitudes, especially among unsuccessful test takers. In another study, Han, Dai and Yang (2004) showed that majority of the research participants had negative attitude towards the Chinese English Test (CET) which impacted their reading test performance. Finally, Cheng et al. (2014) studied test-takers' motivation, test anxiety, and test performance across a range of social and educational contexts in three high-stakes language tests. They reported complex interrelationships of test-takers' motivation and test anxiety in their test performance and influence of social variables such as test importance to stakeholders and test purposes and personal variables such as gender and age on the test takers' motivation and test anxiety.

These findings were relevant to the current study because IELTS, as discussed in Chapter One, plays a key role in the life of test takers and has a great impact on their ambition to further their study abroad. Test takers come to the test with an attitude towards the test. Test difficulty, text content and structure, and the way it is administered, influences their performance.

Part of the test takers' attitude toward a test stems from the social-economic context and factors in which it is used. A stressful socio-economic context in which test takers take the test

can increase the test takers' negative attitude towards it. In the case of the Academic IELTS, the Iranian test takers who participated in this study faced many serious challenges every step of the way for getting a university admission to an English-medium university. Preparing and taking the test successfully was only one challenge in a chain of calamities and hardships. High expectations for achievement from parents and peers, on the one hand, and economic pressure in terms of paying for preparation courses and the test fees which were typically unaffordable for most of the test takers, created a very stressful context for them. And most important of all, hearing about other test takers' unsuccessful experiences taking the test exacerbated their anxiety. Most IELTS test takers were not usually confident if they can have their best performance in the test. They were emotionally stressed out before they took the test and needed a lot of energy and focus to manage their stresses.

3.11. Influence of Text Features on Comprehension Processes and Test Performance

As discussed earlier, text comprehension emerges from the interaction between the reader and the text (Gorin, 2005). Research on text comprehension has also indicated that different textual features potentially affect text comprehension processes and test performance (Kintsch, 1988, 1998, 2004; van Dijk & Kintsch, 1983; Graesser et al., 1991; Graesser et al., 2004). However, Khalifa and Weir's (2009) multi-componential model of reading comprehension also hypothesizes that text variables and features impact different comprehension processes, levels of engagement, and task difficulty. In testing reading comprehension, all test takers are asked to read some texts and answer some questions. Logically, text features have direct effect on the reading processes and test performance of the test takers. So, text features should be a serious consideration in testing reading comprehension. Text features are also highly critical for the reading construct. By definition, a valid reading test should reflect what test takers do with the text and should demonstrate their ability to work through its content and its diverse features. For test developers, identifying text features and task features, then, is a crucial step in defining and operationalizing a valid reading test. They need to provide sufficient evidence that establishes the relationship between text/task features and test takers' comprehension of the text.

At this juncture, I propose to review studies that have investigated the influence of text features and task features on comprehension processes and test performance. This includes linguistic features of the text, text content, the text types (e.g., factual texts, literary texts), the

propositional density of the text, text authenticity and coherence. This review had relevance to the current study, especially the content analysis phase of the study, where the sample IELTS RCM texts and test tasks were content analyzed in terms of lexical density, syntactic complexity, readability of the text materials used, and a number of other relevant features.

3.11.1. Influence of linguistic features of the text

As discussed earlier two basic cognitive processes of reading comprehension are word recognition and syntactic parsing (Khalifa and Weir, 2008, 2009; Kintch, 1988, 2004; Kintsch, 1983). Therefore, lexical and syntactic characteristics of texts used in tests of reading comprehension influence comprehension processes. Language teachers and language testers have generally assumed that text with more unfamiliar vocabulary and more grammatical complexity are more challenging for processing and comprehension (Grabe, 2005, 2009; Barnett, 1986; and Bosser, 1992). We have all frequently heard from L2 readers, especially the less proficient ones, about their struggle with grammar and more importantly vocabulary of the texts. New words in the text hinder comprehension and reduce clarity of the sentence and text meanings.

Furthermore, guessing word meaning from context, which is not always successful, slows down processing. Likewise, more complex sentences are more challenging to parse and take more cognitive capacity which in turn slow down text processing and may also cause confusion. Research findings have also indicated that vocabulary and grammar both contribute to comprehension but there is no conclusive evidence for the relative contribution and significance of each in reading processes and comprehension. Some studies have concluded a subordinate role of syntactic processing in reading (for example, Brisbois, 1995; Haynes and Carr, 1990; Yamachita, 1999) while other studies have reported the overriding role of syntactic processing over vocabulary breadth (Alderson, 1993; Shiotsu & Weir, 2007; Yalin & Wei, 2011). Consistent with these studies, Shiotsu and Weir (2007) and Yalin and Wei (2011) reported that syntactic knowledge is a better predictor of text comprehension than vocabulary knowledge. However, in teaching and testing contexts, both components are viewed as crucial in reading processes. Hence, attention is paid equally to vocabulary and grammar of the texts which is consistent with Barnett (1986) and Bosser (1992) who contend both grammar and vocabulary affect reading in the same way and serve as significant and powerful predictors of higher reading comprehension skills.

Due to the importance of vocabulary and grammatical knowledge in reading comprehension, language teachers, material developers, and test developers have used different tools, techniques and technologies to assess these linguistic features. Understanding these features is essential in choosing appropriate texts and materials that best fit the readers' level of language proficiency. Indices such as lexical density, lexical richness, and lexical variation are being used to assess lexical properties of the texts. In the context of assessing academic reading, general word lists like Coxhead's Academic Word List (2000) are frequently used to assess lexical features of academic texts. As to the grammatical complexity different measures such as mean sentence length, T-units, and clausal subordination have been used to provide some information about the grammar of the texts and materials used.

3.11.2. Text content: Topical/ Background knowledge

Another important aspect of reading comprehension is the use of background or topical knowledge which has shown to play a key role in comprehension process (Nassaji, 2000). As discussed earlier, many reading theories including Schema Theory and Construction-Integration Theory assume that text content influences reading processes and readers' background knowledge can help them integrate text sentences and develop a text representation (Kintsch, 1988, 1998, 2004; van Dijk & Kintsch, 1983). The research literature indicates that unfamiliar text topics hinder comprehension (Best et al. 2006; Snow, 2002) while for texts with familiar topics, readers can activate their schemata and process them more easily. Findings have also supported the role of topical knowledge in reading fluency and have indicated that in academic reading, readers process and comprehend discipline-related texts more easily (Hock, 1990). The issue of topical knowledge is also relevant to inferencing as one key reading comprehension process. Comprehension of meanings that are implied in the text rely on inferencing and topical knowledge. Grabe and Stroller (2002) argued that the role of background knowledge and inferencing becomes more important when comprehension involves synthesizing information from across texts.

Topical knowledge is also relevant in testing and assessing reading comprehension. To avoid test bias and item bias, test developers are generally advised to choose texts and topics which are neutral. Tests of academic reading, in particular, are non-fictional which are generally sampled from different academic resources such as books and magazines. The topics used are all academic but not too technical so that students with different backgrounds can equally relate to

the topic. In the content analysis phase of the study, topics in the current study, of the IELTS RCM texts are discussed in terms of their topic.

3.11.3. Propositional density: Readability

One aspect of text content is propositional density. As discussed earlier, Construction-Integration Theory conceptualized reading comprehension as a cyclical propositional processing (Kintsch, 1988, 1998) where in each cycle of processing different propositions are constructed to from the textbase. Logically, propositional density of each text can influence propositional processing and as Kintsch (1994) suggested more propositionally dense texts are more difficult to process than less propositionally dense texts. Intuitively, it can be assumed that more sentences and paragraphs and longer sentences, and paragraphs are more difficult to process. Texts with higher propositional density are lexically richer and syntactically more complex. Empirical research findings have indicated propositional density of texts can influence text processes. For instance, Embretson and Wetzel (1987) and Just and Carpenter (1971) have shown that performance of test takers on literal and inferential items in propositionally dense texts was more difficult and correlated with item difficulty. However, results are not consistent and some findings (for example, Gorin, 2005) indicated propositional density did not affect text processing and comprehension.

In teaching and testing reading comprehension, propositional density is assumed to be measurable in terms of text readability which considers the number of words and sentences to yield readability levels and grade levels for the texts. However, it is important to note that readability indices are text-based formulas and have nothing to do with the context of reading, topical knowledge involved, text coherence, and the difficulty of the concepts used. Therefore, readability indices can provide an index to propositional density but cannot substitute understandability of the text because it oversimplifies complexity of comprehension processes where readers with different characteristics interact with the text.

Discussion of the role of propositional density in reading processes was relevant to the content analysis phase of the study. As part of content analysis of the IELTS RCM, sentence length and lexical density of the texts were calculated to provide some basis for measuring propositional density and readability of texts. In addition to the readability indices, test takers' verbal report might also include some information on the features of propositional density or topic familiarity that impact reading and comprehension processes.

3.11.4. Text type

In addition to text content, another feature of text that can affect readers'/test takers' performance is text type. Every text belongs to a certain text type and differs from other texts in its organization and structure. For example, the type and number of ideas and the way they are organized and presented in fictional texts differ from non-fictional texts. It is self-evident that argumentative, narrative, expository, descriptive, problem-solution, and cause-effect texts are structurally different. In reading comprehension, it is generally agreed that readers who are more familiar with the structural organization of the text, i.e., text type, can process it more efficiently (Brantmeier, 2005; Yali and Jiliang, 2007). Knowledge of text genre, or textual schemata as Alidib (2007) calls it, facilitated development of macro-structure of the text and serves as a map for putting the small Lego pieces together and building a mega block of text representation. According to Alidib (2004), genre awareness provides the reader with a frame of reference to identify text and the important information needed to process it. Using the shared features of a given genre structure, the reader can process the text more efficiently.

Researchers have examined the effects of text type on reading comprehension test performance and have provided extensive evidence for the differential performance of test takers across different genres. Olson (1985), for example, found that readers encountered more difficulty in processing expository texts than narrative texts. Best et al. (2006) replicated the same study and concluded that test takers did better on narrative texts than expository texts. Interestingly, they also found that text type interacted with the item type where test takers did significantly better for global items in narrative text than in expository text. Their finding is consistent with Snow (2002) who reported that test takers face more challenges in processing global items than local items in expository texts. In another study, Shin (2002) studied test performance of test takers on items that did not follow the story line of a narrative text and found that they performed poorly on such items. They failed to answer items because the items did not fit the story line. The findings lend support to the claim that narrative texts have a hierarchical structure and that readers follow the story line of the macro-structure of the narrative in their reading. Findings of these studies suggest that reading performance of the readers/test takers is influenced by the type of text used and reminds test developers of the importance of selecting appropriate genre for reading comprehension tests.

3.11.5. Text authenticity manipulation

Furthermore, text authenticity in terms of being intact or manipulated and simplified, influences the degree to which the same features of academic texts such as linguistic features of vocabulary and syntactic features, length, organization, topic, etc. are represented. Therefore, it is vitally important to consider text authenticity in reading comprehension processes. It is discussed as an influencing factor in reading processes and test performance. Devitt (1997) maintained that due to their rich linguistic input, so-called *authentic* reading materials provide more linguistic clues and resources for processing and comprehension. In teaching second language reading, it is assumed that simplified texts are easier to process than authentic texts. Simplified texts usually include less sophisticated words and phrases, less complex sentences, and more explicit cohesive devices (Crossley, Allen, & McNamara 2011, 2012), so they have the potential to free up cognitive resources for text comprehension (Crossley et al., 2014; Yano et al., 1994). Simplified texts, in fact help beginner readers or poor readers process the text more efficiently. Some research has also shown that simplified texts are more comprehensible than authentic texts, suggesting that text simplification positively influences comprehension and increases L2 readers' text comprehension (Crossley et al., 2014; Crossley & McNamara 2016; Tweissi, 1998; Yano et al., 1994). Crossley et al., (2014) further found that L2 learners processed beginning-level texts faster than authentic texts. However, results are not conclusive. Atai and Soleimany (2009), for example, reported that test takers who read authentic literary texts outperformed test takers who read literary non-authentic texts.

Use of strategies in processing authentic texts, is then, more representative of real-life language use and lends more support to construct validity of the test than processes and strategies that are used for comprehension of non-authentic simplified texts. Text authenticity also has much relevance to academic reading assessment where test takers process texts and tasks. With non-authentic texts (defined here as simplified, de-contextualized and unrepresentative texts), it is hard to argue for academic validity of the test as a representative of academic texts and activities.

In teaching second language reading, language teachers and reading researchers might simplify texts to address the difference in L2 proficiency levels of the learners/test takers. They manipulate text levels for L2 readers and make texts easier to read (Crossley, Yang, & McNamara 2014; Yano, Long, & Ross, 1994). Some vocabulary items might be manipulated and

substituted, or some sentences might be re-phrased or re-written. Such manipulations and simplifications of lexical, syntactic, and discorsal features of the text can disturb both text authenticity and text coherence, and consequently the text processes used by the readers. Similarly, in the context of developing tests of reading comprehension, test developers and item writers may need to manipulate and tailor some features of a text to achieve test objectives, meet the recommended standards, or make them more appropriate for a certain level of language proficiency. Such modifications raise questions about authenticity of the text material used in the test and validity of test construct.

3.11.6. Text coherence

Another text feature that influences text comprehension is related to text coherence. Coherent texts use different discourse markers to connect two parts of the text. Coherent texts achieve more effective expression of meanings such as problem-solution relationship, cause-consequence, comparison-contrast relationship, and other meaning relationships. Explicit expression of such relationships through cohesive devices facilitates text processing and comprehension and helps readers better process the intended meaning. Discourse markers provide explicit information about the relationship between different parts of the text which helps readers to use them as signposts and to construct text representation. Attending to these discourse relations between different parts of the text are central to text processing and development of textual and mental representation (Degand and Sanders, 2002; Graesser et al., 2004; Ozuru, Dempsey, & McNamara, 2009). Research has shown that simplified text can decrease coherence of the text and hinder comprehension (Bernhardt, 1991). Additionally, Freedle (1997) found that texts that are judged to be very coherent yield main reading comprehension points that are easier to understand. The same finding was reported by Koda (2005) who reported improving text structure led to improve comprehension.

3.11.7. Order of text information and test tasks

In assessing reading comprehension, another aspect of text structure that influences test performance is related to the ordering and sequence of information in both the text and test questions. Under normal circumstances, readers usually read the information in the same order it is presented in the text. So, location and order of the information in the text influences text representation developed by the reader (Urquhart 1984; Sheehan & Ginther, 2001). Order of

information in tests of reading comprehension also influences item difficulty because, as Freedle and Kostin (1993, 1996) discussed, activation of information that is previously stored by the reader is affected by the location of the information in the text. This suggests that the reader develops text representation according to the order in which the information is presented to him. Manipulation of order of information in the text can contribute to item difficulty. In a study, Sheehan and Ginther (2001) examined the difficulty of the multiple-choice items in reading section of the TOFEL test and reported that items that were not related to the expected order of information were more difficult to process while items that addressed information that was part of the expected order of information were easier to process. This suggests that the order of information in the text influenced development of a coherent representation. Likewise, items that addressed information from the coherent representation were much easier to process than items that address information which was not part of the coherent representation. They concluded that proximity of the relevant information influenced both text representation and item difficulty.

As discussed before, validity of reading comprehension tests depends on the extent to which test tasks and items tap the same cognitive processes and strategies used for comprehension of a given genre which in turn depends on the degree test tasks reflect the structural organization of the text. Test tasks and items that contradict the order of data presentation lack authenticity of text representativeness. This issue is relevant to the construct validity of reading tests. A test with tasks that taps into the order in which the information is presented and processed enjoys more representational credit. Tasks and items that contradict the order of information presented in the text, on the other hand, defy the natural reading processes readers used in developing text/mental representation, and hence, lack credibility to represent construct of academic reading. Order of information in the text can also explain some of the challenges and difficulties test takers face in doing different test tasks. The match between order of the information presented and the information asked by the test tasks and items reflects part of the process of reading and can help better understand the test processes and find out if the test tasks are sensitive to the structure of the text and the way information is presented in the text. In brief, order of test items must reflect the order in which the information is presented in the text and tap into the same cognitive processes and operations that relate to the specific text genre.

3.11.8. Stem structure and its impact on test performance

In addition to text characteristics, test task features in terms of their form, quality, clarity, focus, and orientation also play a key role in test performance. Stems as important parts of test items present the problem to the test takers and trigger comprehension processes. Test developers and item writers are supposed to do their best to develop quality items and stems to facilitate reading processes. Stems can be presented in different forms; complete sentence or incomplete sentences. It is commonly believed that complete sentence stems appear to be easier to comprehend than incomplete sentence stems, but research findings are inconclusive. Some studies have shown no significant difference in the effect of stem completeness/incompleteness on test performance and item difficulty (Board and Whitney, 1972; Dudycha and Carpenter, 1973; Crehan & Haladyna, 1989a, 1989b; Ascalon et al., 2007) while other studies reported that incomplete stems were found to be significantly easier than complete stems. Phipps and Brackbill (2009) reported that incomplete stems were significantly easier than complete stems. In spite of the inconclusive results, the bottom line is that it is vital to consider the stem format to ensure the quality of the test so that test takers' performance is facilitated by the test format. Another aspect of stem quality has to do with stem orientation. Stems can be positive or negative statements. Again, the common belief is that positive stems are more facilitative in comprehension processes. Dudycha and Carpenter (1973) found that examinees performed better on positively worded stems than on negatively worded stems. They also found an interaction of orientation and completeness. In a similar study, Haladyna, Downing, and Rodriguez (2002) came up with a similar finding and recommended test constructors to develop test items with positively worded stems. They also found that question stems were easier than incomplete sentence stem questions. These findings have relevance to reading construct validation studies. Analysis of item formats and their quality can identify possible sources that may contribute to construct irrelevance variance in the test.

These theoretical discussions and empirical findings related to reading processes and text features that influence reading processes, discussed above, had relevance to the current study in many ways. First, they helped to better conceptualize the research problem and purpose. They also helped conceptualize the reading construct as a multi-componential skill which is influenced by readers' characteristics such as language background and level of language proficiency, and text features such as text type, test content, lexical, grammatical, and discoursal features of the

text, and qualities of test items. These factors and variables interact and impact on the test takers' processes and choice of reading strategies to derive meaning from the text and develop a text representation or situation model to accomplish the test tasks. For example, the importance of metacognitive strategies in reading comprehension, evident in the literature and summarized above, required that this study consider their role in reading processes as they are an integral component of the reading construct.

Second, the literature also informed the choice of methodology and the research design. For instance, discussion of readers' variables such as role of readers' language background and level of language proficiency are influenced the choice of a more comprehensive sample of research participants that included L1 and L2 English readers, as well as more and less successful readers.

Third, the literature provided a bridge between the present study and the body of existing knowledge in this area of research. Furthermore, in light of the available literature, it was clear that drawing on different sources of evidence would allow for better understanding and interpretation of the reading construct operationalized by the IELTS RCM. Such sources of evidence would clarify any construct irrelevant variance (Messick, 1989) and allow the researcher to relate qualities of texts, test tasks, test items, and stems to test takers' performance. Finally, the literature reviewed in this chapter enriched the three strands of evidence adopted in this study, i.e., content analysis of the test, verbal accounts of the SKSPs used by test takers, and experts' judgements and accounts.

3.12. Summary of the Chapter

In this chapter, Khalifa and Weir's (2009) Cognitive Reading Comprehension Model was presented. In addition, the literature relevant to the factors that influence reading comprehension, which include the use of cognitive processes and metacognitive strategies in reading, the influence of readers' variables such as their language background, level of language proficiency, and their motivation for reading was discussed. The chapter also discussed text features that influence reading comprehension, including linguistic features, text content, propositional density of the text, text authenticity and coherence, and order of information and test tasks.

Khalifa and Weir's (2009) Cognitive Model of Reading Comprehension provided a working and testable framework for exploring the reading comprehension construct

operationalized by the IELTS RCM. In addition, some of the other cognitive reading theories, such as Strategic Competence Theory, Schema Theory, and Construction-Integration Theory, and their relationship to Khalifa and Weir's model, were also reviewed. The review also included some validation studies of reading tests that adopted Khalifa and Weir's model. Additionally, results of empirical findings related to the construct of the IELTS RCM were discussed.

The chapter was organized in two separate parts. Part One discussed Khalifa and Weir's (2009) model, other reading theories that support it, and research studies that adopted the model for studying the reading comprehension construct. Part Two was devoted to the empirical findings related to the construct of the IELTS RCM and three factors that influence reading comprehension processes; 1) studies that have established the impact of reader's use of cognitive and metacognitive strategies on text processing and test performance, 2) studies that have addressed the influence of readers' variables such as language background (L1-L2), level of language proficiency (high-low), and motivation on reading processes, and 3) studies that have debated the impact of text features on reading processes and task difficulty. The linguistic features of texts were also discussed, namely: text content and topic, text types and their related structural organizations, propositional density of text, text authenticity and coherence, the order in which information in the text is presented, and role of stem structure in test performance. It also included influence of item features on item difficulty, such as structure, focus, orientation and completeness.

In the following chapter, the methodology and research design of the study are presented.

CHAPTER FOUR: METHODOLOGY

4.1. Introduction

As mentioned in the previous chapters the focus of this study was on exploring the construct validity of the Reading Comprehension Module of the IELTS (IELTS RCM) test. The last three chapters presented the research questions, the theoretical framework of the study, and some of the literature related to the empirical findings of the reading comprehension construct and the IELTS RCM. The study adopted a cognitive approach to construct validity by examining the skills, knowledge sources, processes and strategies (SKPSs) involved in the test. To this end, in addition to test content analysis of a sample IELTS RCM, test taking accounts and judgements of multiple stakeholders were examined. These included 1) undergraduate L1 English speakers, 2) undergraduate EFL L2 learners (more successful and less successful), and 3) language testing experts. The assumption was that a diverse pool of evidence from different stakeholders could enrich the construct evidence of the IELTS RCM. This chapter presents the methodology that was adopted to conduct the study and answer the research questions. It includes details of the research design, participant recruitment procedure, participants, instruments, and procedures for data collection and data analysis.

4.2. Research Design

This study is a qualitative case study, consisting of three phases: phase one, content analysis; phase two, test takers' accounts of the IELTS RCM; and phase three, experts' accounts and judgements of the IELTS RCM. (See Figure 4.1, and Section 4.6 for details by phase of the research procedures). However, more emphasis and analysis were put on phase one and phase two of the study. In phase one, the sample IELTS RCM texts and test tasks were content analyzed in terms of lexical density, syntactic complexity, readability of the text materials used, and a number of other relevant features. In phase two, evidence was collected from three different stakeholder groups in order to explore the construct validity of reading comprehension as it is operationalized by the IELTS RCM. In phase three, experts' accounts and judgements of the IELTS RCM were elicited to extend and triangulate evidence from the other phases.

Overall, the study adopted a *multiple case study* design (Chmiliar, 2010), which is defined below in relation to general case study research. Yin (2002) characterizes case study

research as involving extensive data collection with multiple forms of data to provide rich, detailed, and in-depth information about the phenomenon under investigation. Stake (1995) argues that a case study must be “bounded” so that it can be separately studied within a specific time, place, or context boundary. The bounded system, he indicates, may be as simple as a single individual or group or it may include programs, events, or activities. As an extension of bounded case study design, *multiple-case design*, also known as collective case design, refers to “case study research in which several instrumental bounded cases are selected to develop a more in-depth understanding of the phenomena” (Chmiliar, 2010, p. 582)

Chmiliar (2010) indicates that multiple case study design is appropriate for a research context where an in-depth exploration of a specific bounded system is targeted. The design provides diverse sources of evidence collected through multiple cases to see how the system works and operates. “Multiple-case design allows examination of processes and outcomes across many cases, identification of how individual cases might be affected by different environments and the specific conditions under which a finding might occur” (Chmiliar, 2010, p. 583). The design helps to study how the specific feature(s) of each “bounded system” might change or influence the processes and outcome. Multiple cases each belonging to a certain bounded system provide a more varied range of conditions and circumstances and can contribute to more powerful results and more extensive descriptions and explanations of the phenomena. It may also improve the transferability of the findings. With more diverse cases results can provide rich description and a broader context for interpretation and use. Results obtained from multiple cases can enhance possible transferability by providing more evidence for understanding the phenomena across different contexts and boundaries.

As Chmiliar (2010) points out, two different procedures can be used for multiple-case design processes: 1) parallel design process and 2) sequential design process. This study adopted a parallel design process where cases are selected in advance and are studied at the same time, independent of one another. Unlike sequential design, cases are not selected as a result of the outcome of the previously completed case.

This research sought to explore the knowledge sources, skills, processes and strategies (SKSPs) used by three groups of test takers and provide rich, diverse evidence of the reading construct of the IELTS RCM. These purposes were best served by a parallel multiple case design. In this design, each group of participants acted as an independent “bounded system”

characterized by their first language or their second language proficiency level or expertise. For example, in phase two, the L1 test takers came from a context where they had acquired English as their L1. L2 learners, who had learned English as a foreign language, formed a different bounded system. The context in which they had learnt English was so different from the L1 speakers that it had the potential to create a different dynamic and system of learning which might in turn influence the SKPSs they used in test performance. The same applies to more successful and less successful L2 learners. It can be argued that more successful L2 learners have different skills and abilities to engage with the test tasks and that they operate differently in the context of the test tasks. The phase three experts provided yet another window on the phenomenon of interest.

The use of a multiple case design fit the research focus of the study, which sought to explore accounts of different stakeholders in doing the IELTS RCM. The design also allowed within-case comparison to reveal individual differences within each bounded system of cases. Moreover, once a full account of each case was analyzed and interpreted, cross-case comparisons could be conducted to identify key variables and examine how they were patterned in each set of bounded cases. Cross-case analysis also helped find out how the KSPSs used by the participants varied or remained the same across different systems. Moreover, the design allowed analysis of the accounts of each single participant (case) and the sum of cases in each bounded system, i.e., L1 English speakers, successful L2 learners, and less successful L2 learners. Data collected from this diverse system of cases provided a more comprehensive description of the construct of reading in the IELTS RCM. In fact, the multiplicity of cases provided a unique, more holistic characterization of the construct of the IELTS RCM and could show possible changes or alternation of the construct across different boundary systems. In each case, accounts from the participants (e.g., English L1 speakers, more successful L2 learners, less successful L2 learners, and experts) were examined independently and compared against each other.

In brief, using a multiple case design provided more information and evidence for exploring the construct of reading comprehension in the IELTS RCM. Cross-case examination revealed similarities and differences in terms of the KSPSs used by different test takers and provided a more comprehensive and deeper understanding of the variables and issues relevant to the way the construct operates in different conditions. It also helped enhance the meaningfulness and transferability of the results.

This study was reviewed and approved by the Carleton University Research Ethics Board (CUREB) on March 27, 2019 (Appendix A). Approval was granted to conduct research with human participants both in Canada and Iran.

4.3. Sampling Design

There is consensus among researchers that in quantitative research an ideal sampling design in educational research is probability random sampling because it allows for generalization over population (deMarrais & Lapan, 2003; Scott & Morrison, 2005;). For qualitative research, which aims at understanding a local phenomenon and has no interest in generalizing the findings, researchers usually choose sample participants who represent a range of perspectives (Cresswell, 2007). Purposive sampling may also be used when a researcher wishes to include only those people who meet a very narrow or specific set of criteria. In the context of the current qualitative study, probability random sampling could not be achieved for practical reasons such as the unavailability of all undergraduate participants from the faculty for random selection. Therefore, for selection of the participants, purposive non-probability sampling was used. Purposive sampling is assumed to yield information-rich cases (Patton, 2011) and allows selection of participants who 1) can make themselves available, 2) are willing to participate, and most importantly, 3) are capable of communicating their accounts, experiences, and opinions in a reflective, expressive, and articulate manner (Bernard, 2006; Spradley, 1979). The purposive non-random sampling strategy applied in this study required participants to have certain characteristics that aligned with the objectives set by the researcher. An account of the inclusion criteria is provided below.

4.3.1. Inclusion criteria (Phases Two and Three)

The study consisted of three separate phases of data collection (see Figure 4.1 below for an overview of the data collection phases). In phase one of the study, a content analysis of a sample IELTS REC was undertaken. In phase two, participants were recruited from two main groups: 1) undergraduate L1 English speakers and 2) undergraduate L2 English learners. Two criteria were used for selecting the L1 English speaking research participants: 1) the participants must have acquired English as their first and only language spoken at home, and 2) participants needed to be studying at the undergraduate level or about to graduate. Whether the participant was an L1 English speaker or not was based on the participants' claims about acquiring English

as their first language. In addition to the availability of the participants for data collection, another reason why undergraduate students were sampled for this study was that most Academic IELTS applicants who apply for admission for an undergrad or master's degree in English speaking countries are between the ages of 19-25. Most undergrad students apply as they finish high school. Participants in this study fell within the same age range as the actual Academic IELTS test takers.

For L2 participants, however, the inclusion criteria were slightly different. 1) The participant had to be studying at an undergrad level or completing an undergrad degree, and 2) the participant had to demonstrate motivation and interest in taking Academic IELTS for university admission purposes. Without such an interest, taking the IELTS test would not signify any importance and would be irrelevant to them. Only participants who were planning to take the test or were preparing for the test to gain admission to a university in an English-speaking country were selected for the study. Such participants were familiar and engaged with IELTS and would possibly provide more information about how they had taken or would go about taking the IELTS RCM and perform on its tasks.

In phase three of the study, a sample of testing experts were selected for data collection. The inclusion criteria for this group of participants included the following criteria: 1) A Master's or PhD in TESL/applied linguistics, 2) good knowledge of language testing and assessment, 3) work experience in test development, test administration, test scoring and test use. These criteria were used to ensure that the participants' competency and work experience in second language teaching and assessment qualified them to evaluate the construct of the IELTS RCM and its different test tasks. Of the 10 language testing experts, five had experience in teaching IELTS preparation courses, and five had strong language testing backgrounds and had conducted research in applied linguistics. All participants voluntarily participated in the study and signed the informed consent form that had been approved by the Carleton University Research Ethics Board (CUREB) (See Appendix B for a copy of the Ethics Certificate and a copy of an informed consent form). In sum, three different groups of participants took part in this study: 1) undergraduate L1 English speakers, 2) undergrad L2 learners (more successful and less successful), and 3) language testing experts.

The following section presents detailed information about participants of the current study. This includes ethics clearance, sampling design, recruitment, and demographics of the

participants within each of the three principle groups, namely the L1 English test takers, the more successful L2 test takers, and the less successful L2 test takers.

4.3.2. Participant recruitment

To recruit research participants, emails were sent to the academic email addresses of undergraduate students in the Faculty of Arts and Social Sciences at a mid-sized Canadian University. The participant recruitment email described the purpose and outline of the study and included details of the inclusion criteria. It was sent to the undergraduate students' emails (See Appendix A). Interested students who responded to the email received more details about the study and arrangements were made for their participation.

4.4. Participants

In this section some details of the research participants are presented.

4.4.1. Undergraduate L1 English speakers

As presented in Table 4.1, the first group of research participants included a sample of 10 (4 male and 6 female) undergraduate L1 English speakers studying at a Canadian university. They were studying different undergraduate programs and had different majors (linguistics, philosophy, communication, English literature, social sciences, cognitive science and psychology). They were 2nd year and 3rd year students. None of the L1 English speakers—whose ages ranged between 20-25—had taken IELTS before, nor did they have any intention to do so. They did not know much about the test but expressed interest in participating in the study. Based on the research criteria, all of them were qualified to participate in the study.

4.4.2. Undergraduate L2 English learners

The second group of participants was a sample of 11 undergraduate Iranian students who had learned English as a foreign language either in private language schools or at university. They had at least 7 years of English language learning experience from attending public school in Iran where English is a compulsory course. They had also attended some English language learning programs at private language schools for 4-8 semesters to improve their English after graduating from high school. They were all preparing for the IELTS test in order to gain admission to universities in Canada, Australia, and UK. They had taken some mock IELTS tests.

They majored in TEFL (Teaching English as a Foreign Language), computer sciences, industrial management, arts, and mechanical engineering. Out of 11 English L2 learners, whose ages ranged between 20-25, three had already taken the test once. Others were preparing to take the test and were attending test preparation courses in a language institute.

It is worth noting that all L1 English speakers were 2nd year and 3rd year students whereas for the L2 participants, three had just graduated from their undergraduate programs and were attending a preparation course to take IELTS as a requirement to apply for master's programs in an English-speaking country such as Canada or Australia. Based on the results of the cloze test and the reading essay, the L2 learners were at different levels of language proficiency; six were at lower language proficiency levels and five participants were at higher levels. They were identified as more successful and less successful test takers, respectively, in the study for this reason. To protect the participants' identity, four letter pseudonyms were assigned to each research participant.

4.4.3. Expert Participants

To address the third research question, which examined language testing experts' accounts and judgements of the reading construct of the IELTS RCM reading test, a sample of 10 language testing experts with a wide range of education, skills, and work experience participated in the second phase of the study. They were recruited by email (Appendix B) from institutions where I had colleagues or had studied or worked as a teacher myself, in Canada and Iran. They all took the IELTS RCM sample test and the “*Construct of the IELTS RCM Questionnaire*” to provide their accounts of the test. They also participated in a semi-structured interview to detail their accounts of the test and its construct in terms of the skills, knowledge sources, skills, processes, and strategies (SKPSs) used and measured in each test task. A summary of the demographic information of the participants is presented in Table 4.1.

Table 4.1.

Demographic details of the research participants

Participants		Age range	Male	Female	Total
L1 English speakers	10	19-22	4	6	10
More successful L2 English learners	5	19-25	3	3	5
Less successful L2 English learners	5	19-25	2	3	6
Testing Experts	10	30-65	7	3	10
Total	30		17	13	31

The 10 experts had a wide range of skills and experience in language teaching, language testing, teaching and assessing reading comprehension, and IELTS testing. Their work experience ranged between 8-30 years of teaching, researching and practicing reading comprehension, language testing, and developing and administering language tests such as IELTS. They included 3 female and 7 male participants whose ages ranged between 30-65. Seven participants had academic careers as full professors, associate professors, or assistant professors. The other three experts had extensive experience in teaching IELTS and TOFEL iBT preparation courses. Table 4.2 reports details of the experts' education and years of work experience.

Table 4.2.

Testing experts' characteristics

	Master	PhD Candidates	PHD	Total
	1	2	7	10
Years of work experience	8 yrs.	9 yrs.	7-30 yrs.	

4.5. Instruments

To collect the required data for addressing the research questions, different instruments were used. In this section, some details of the instruments used are presented.

In order to have a homogeneous group of participants in terms of their language proficiency and reading ability and collect the required data to answer the research questions, seven different instruments were used, some for measuring English language proficiency of the participants (instruments 1-3) and some for collecting the required data to address the research questions (instruments 4-7); 1) a cloze test developed for this study (See Appendix D), 2) a reading essay with five essay-type questions (See Appendix E), 3) a written summary of the essay (See Appendix F), 4) a pilot IELTS RCM, 5) the main IELTS RCM sample (See Appendix G), 6) Test Performance Observation Scheme (See Appendix H), and 7) Construct of the IELTS RCM Questionnaire (See Appendix I) .

4.5.1. Cloze test

To measure the language proficiency of the L2 participants, a cloze test was administered. The cloze test was specifically developed for this study. The standard cloze procedure recommends removing words from a reading passage at regular intervals (usually every seventh word), leaving blanks of standard length (Mousavi, 2009). Standards also

recommend choosing topics and texts that fit the standards of cloze test characteristics. These standards were followed and a general reading passage on the topic of food and culture entitled “*What food tells us about culture*” was selected from an online magazine (freelymagazine.com). Since the main purpose of the cloze test was to assess the L2 learners’ language proficiency and reading ability, the difficulty level of the cloze text had to be appropriate to the L2 participants. Text that was too difficult, or a passage that was too easy would not be helpful in measuring the language proficiency of the L2 participants. A text of average difficulty was more practical for measuring L2 participants’ language proficiency and reading ability.

As is standard practice in cloze tests, the first two sentences of the text were left intact (Mousavi, 2009). Then, every seventh word was systematically deleted from the texts, resulting in a total of 40 open ended deletions. Some unaltered lead-out was also left at the end of the passage. Next, the test was piloted with a small sample of five undergraduate L1 English speakers to check for possible unnoticed problems. Their responses, especially the content words which could be more subject to variation, were analyzed. Such a clozentropy scoring procedure is advised for improving the reliability of scoring (Mousavi, 2009). Results of the pilot study showed no serious challenges in terms of the text, the deletions, scoring, or administration. A list of acceptable answers was developed from the responses of the five pilot participants. Acceptable response scoring was adopted for scoring the responses. Minor misspellings were ignored.

4.5.2. Reading essay

In addition to the cloze test, attempts were made to gather further evidence of the participants’ reading ability in a non-test context. In academic reading classes, texts are generally read in the context of a real task such as class discussion, class presentation, or in writing summaries and/or reflections. Therefore, more realistic test tasks were included in the assessment of the participants’ language proficiency. Among different options, essay-type questions were thought to be more relevant to the kind of academic reading activities students usually do.

Since all the participants were undergraduate students—and were used to reading different academic texts such as research articles or book chapters in their disciplines—authentic academic texts could better represent an estimate of the academic reading ability of the participants than test tasks that were designed for testing general reading purposes. With the assumption that reading an essay could be used as a more authentic measure of reading

proficiency of undergraduate students, attempts were made to choose an essay that was more representative of academic essays.

Another consideration in the choice of the essay was the participants' academic backgrounds. Since they majored in different disciplines in the humanities and social sciences, care was taken not to place greater demand on their knowledge of content area by avoiding topics that were intrinsically disturbing, emotionally charged, politically sensitive, and cognitively demanding. To this end, a more general topic which did not call for specific background knowledge was sought. With all these considerations in mind, a text with a topic of common interest was selected, namely, "Non-verbal communication skills" from the book "*Exploring Language*" by Goshgarian (2004). The selected essay was relatively longer than IELTS texts but not as long as an academic essay. At the end of the essay there were five essay-type questions addressing different topics and sections of the essay which seemed more realistic and closer to the type of academic reading activities than the IELTS RCM test tasks. In practice, academic essays are at least some 6000-8000 words long. However, for practical reasons, a shorter essay had to be selected. Overall, compared with the RCM IELTS test tasks, a reading and essay with some essay-type questions seemed more realistic and authentic for assessment of the academic reading ability of the participants. All participants took this reading essay.

Both groups of participants took the cloze test and the reading essay and provided their oral and written summaries of the essay to help assess their level of English language proficiency

To have a clear idea of the difficulty of the cloze passage and the reading essay, the readability of the instruments used was also calculated. The most common indices of readability, including the Gunning Fog Index, SMOG Index, Flesch-Kincaid Grade Level, and Flesch Reading Ease were used in assessing the readability of these instruments. Table 4.3 presents the results.

Table 4.3.

Readability of the Cloze test and the reading essay

	Cloze test	Reading Essay
# words	308	1162
# sentences	18	60
Lexical density	.36	.45
Average sentence length	17	19.37
Gunning Fog index	11	14.9
Flesch Kincaid Grade level	8.9	12.7
SMOG	11.5	14.3
Flesch Reading Ease	62.2	38.7

As shown in Table 4.3, the cloze test was at the 9th grade level while the reading essay was at the 12th grade level. Other readability indices told the same story. The SMOG reading index of the reading essay was 14.3. The same index for the cloze text was 11.5 which indicated the cloze was much easier than the reading essay. The reason why a lower readability was chosen for the cloze test was that the cloze test was *altered* (because of all of the ellipted words) which could make it harder to read. Processing an altered text is more difficult than processing a normal, intact text. An altered text includes several gaps which interrupt normal reading and processing. It was assumed that the number of deletions would increase text difficulty. Easy texts can include more deletions while for difficult texts the number of deletions should be limited. Additionally, as Bachman (1982, 1985) pointed out, cloze tests have a very high method effect which justified the choice of a less difficult text. Another consideration in choosing a text with lower readability was that for the L2 participants, an altered text with a higher readability index could be too challenging to process. Moreover, a text with higher readability may not easily lend itself to the cloze procedure because longer texts tend to be more technical and include proper names, numbers and specific ideas that defy the cloze procedure. The reading essay, on the other hand, was at 12th grade high school level. As the L2 participants' reading proficiency is lower than L1 English readers, it was important to select an essay at an appropriate difficulty level that would allow the L2 English speakers to process the passage and answer the essay-type questions. In terms of the syntactic complexity of the text, the average sentence length for the cloze text and the essay was 17 and 19 respectively, which indicates the use of complex and compound-complex sentences in both texts.

4.5.3. IELTS Reading Comprehension Module: Pilot test

The sample pilot test was taken from the official web page of IELTS organization (<https://www.ielts.org/about-the-test/sample-test-questions>) which presents a number of samples for each module of IELTS as well as all the information test takers might need to know about IELTS Academic and IELTS General Training. It provides test candidates with all the information needed for registering, preparing and taking the test, as well as training, booking a test, scoring, and so on. The sample IELTS RCM included several reading passages and nine test tasks; the *True, False, Not given Task*, the *Matching Features Task*, the *Diagram Completion*

Task, the Matching Headings Task, the Summary Completion Tasks 1 and 2, the Multiple-Choice Task (two answers), the Multiple Choice Task, and the Yes/No/Not given Task.

Participants were expected to take the IELTS RCM and provide their accounts of the test, but it was not clear if they were all familiar enough with the test. To make sure all participants in phase one of the study were familiar with the IELTS RCM, a sample IELTS RCM was piloted with all the participants. Results showed that participants had different knowledge and background about IELTS. All the L1 English speakers had just heard about the test and had no idea what it looked like. The L2 participants, on the other hand, were more familiar with the whole test, including the Reading Comprehension Module of the test. Some of them had taken the test before and knew what the different components of the RCM are. Others had attended two preparation courses before. As the participants could have had different degrees of familiarity with IELTS, a sample RCM was administered to all participants. This allowed them to familiarize themselves with the test and let them develop a good understanding of the different types of test tasks that are included in the IELTS RCM. This ensured that all participants were on the same page in terms of knowing what the RCM consists of and their experience with the test. In brief, the main purpose for using the sample test was to make sure all participants were familiar enough with the test and they all had experience with the test tasks before they took the main sample that meant to be used for addressing the research questions. Taking the sample test could also provide an opportunity to practice verbal reporting of the test taking experience and the processes and strategies they used. The test was taken from IELTS homepage (www.IELTS.org).

4.5.4. IELTS Reading Comprehension Module: The main test

The instrument that was the main research instrument used for addressing and answering the research questions was an actual sample IELTS RCM published by IELTS and Cambridge University Press. It consisted of three reading texts, each of which was followed by three different and distinct test tasks. The first reading passage (*Raising the Mary Rose*) was followed by three different tasks: 1) the *True/False/Not given Task*, 2) the *Matching Features Task*, and 3) the *Diagram Completion Task*. The second text (*What destroyed the civilization of Easter Island*) was followed by three other tasks 1) the *Matching Heading Task*, 2) the *Summary Completion Task 1* (selecting words from the text), and 3) the *Multiple-Choice* (two answers) Task. Finally,

the third passage, entitled *Neuroaesthetics*, was similarly followed by three more tasks; 1) the *Multiple-Choice Task*, 2) the *Summary Completion Task 2*, and 3) the *Yes/No/Not given Task*.

It is worth noting that the main sample IELTS RCM was content analyzed. Both the text features and the task features of the test were analyzed in terms of their linguistic, topical, discursal, and cultural dimensions. The analysis was conducted on the assumption that features such as vocabulary, grammar, topic, and other features of the text impact comprehension processes and contribute to task and item difficulty.

4.5.5. Readability of the IELTS RCM and academic texts

Since the focus of the study was on exploring the construct of the academic reading of the IELTS, some characteristics of the IELTS RCM text passages were examined. Examining characteristics such as readability helped in comparing the IELTS RCM texts with actual academic texts. To this end, three real academic texts (an academic book chapter, an academic article, and an online short essay) were selected for comparison. The academic articles were taken from the L1 English speaker participants who were doing their first- or second-year courses. They had read them as part of their academic reading. Such comparison provided some information as to how different or similar the IELTS RCM and actual academic texts are. Table 4.4 presents the results of readability scores for the IELTS RCM texts and the three academic readings (a short essay, a book chapter, and an academic article).

Table 4.4.

Readability and linguistic features of the research instruments

# words	# words	# sentences	Average sentence length	Flesch grade level	SMOG	Flesch Reading Ease
The Mary Rose	891	39	23	11.89	11.70	50.53
Easter Island	906	49	15	11.65	12.65	44.45
Neuraesthetics	985	42	23.	13.55	13.95	39.68
Psychology short article	1117	47	24	11.68	12.52	53.06
Linguistic book chapter	7632	366	21	13.27	14.28	44.20
Linguistic article	7377	307	24	13.06	13.99	44.20

As shown in Table 4.4, based on the Flesch grade level readability indices, the IELTS text and the academic texts were relatively at the same grade levels. The academic texts were not much more difficult than the IELTS RCM texts. The first two texts of IELTS were at a grade 12 level and the third text was at university level (13th). The psychology short article was at the 12th

grade level while the book chapter and the academic article were at university levels (13th). The table also shows that the readability of the reading essay and the three reading passages in the IELTS RCM were within the range of (37-44.4). The same results were found in other reading indices. SMOG reading indices for the three reading passages of IELTS were 11.70, 12.65, and 13.95, respectively. IELTS texts were at the 12th grade level and university level. The academic readings, on the other hand, were a bit more difficult and were at a university level (13th).

4.5.6. Test Performance Observation Scheme (TPOS)

One more research instrument used for data collection was the “*Test Performance Observation Scheme*”. It was designed to record test takers’ behavior while performing on the test tasks. The scheme was used for recording the amount of time spent on each test task by each test taker and the number of times a test taker looked back and forth between the text and the test tasks. Moreover, test takers’ patterns of task performance in terms of the order in which they answered the task items, delays in answering an item, and changes in an answer were also part of the test performance observation scheme. These observations were recorded in the observation scheme.

4.5.7. Construct of the IELTS RCM Questionnaire

For collecting testing experts’ accounts and judgements, a top down approach was adopted. To this end, a questionnaire was developed to capture some of the main SKSPs that could be involved in each test task. Prior to data collection, a set of classifications, features, reading skills, knowledge sources, strategies, and processes were identified and organized. These were based on the existing literature on reading research, assessment of reading, and the theoretical model of the study (Khalifa and Weir, 2009; Gorin, 2006). The set of classifications formed the core codes and categories of the questionnaire which was named “Construct of the IELTS RCM Questionnaire”. To collect experts’ judgements and accounts, Saldaña’s (2009) protocol coding was applied. Protocol coding is defined as “the collection and, in particular, the coding of qualitative data according to a pre-established, recommended, standardized, or prescribed system” (p.151). Testing experts received the Construct of IELTS RCM Questionnaire, which they then used for coding each test task.

The questions that were included in the questionnaire addressed some of the basic classifications of the reading construct, which is extensively discussed in reading research and

reading assessment literature. For instance, the three levels of comprehension adopted in the questionnaire were based on Kintsch's (1988; 1998) theory of comprehension, which provides an elaborate description of how meaning is produced through constructing different representations during reading. Based on this view of reading, Van Dijk and Kintsch (1983) developed a discourse model of comprehension which proposes text propositions are linked via semantic relations. The model also suggests that representations occur at three different levels: 1) the linguistic level (representation of words/sentence), 2) the text level (representation of meaning), and 3) the situation level (representation of the text integrated with reader's prior knowledge). In the current research, these three levels of representations were termed literal meaning, inferential meaning, and evaluative meaning.

Unlike other models of reading (for instance, Hoover & Tunmer, 1993; Rayner & Pollatsek, 1989), Khalifa and Weir's (2009) model suggested careful reading and expeditious reading. In careful reading, all sections of the text are carefully read to understand the precise meaning expressed, while expeditious reading is characterized by quick, non-linear, and selective reading to access the needed information. Such a classification is relevant to reading in testing contexts.

The reading skills included in the questionnaire were aimed at identifying the reading skills measured by each type of test task. In spite of the shortcomings of skill approaches to reading instruction and reading assessment, it is still the dominant approach in many types of reading tests—including high stakes tests. The skill approach to reading has proved useful and practical for diagnostic and pedagogical purposes. It has helped test developers to define and operationalize reading construct. The IELTS RCM, just like many high-stakes tests, is a skill-based test which claims to be a measure of several reading skills such as reading for gist, reading for main ideas, reading for details, skimming, understanding logical argument reading for detail, skimming, understanding logical argument and recognizing writer's opinions, attitudes and purpose. To examine what reading skills are possibly tapped by each task type, a list of several reading skills was included in the questionnaire. These skills were adopted from the literature on reading and were frequently discussed and referred to in reading literature (Alderson, 1990, 2000; Khalifa & Weir, 2009).

In addition to some demographic information of the experts such as gender, age, level of education, and years of work experience, the questionnaire consisted of several classifications

that are discussed in literature on reading construct and include the levels of reading required for each task (word, phrase, sentence, paragraph, and text), the knowledge needed to answer each type of test task (grammatical, vocabulary, textual, background), the type of comprehension involved in each task (literal, inferential, and evaluative), the speed of reading involved (careful, expeditious), the difficulty of the text (easy-difficult), the difficulty of the task (easy-difficult), the level of processing (low level-high level), the time needed to do each type of item (less than 1 minute-more than 2 minutes), and the reading skill(s) measured by each test task. These features were all based on the existing literature on reading research and assessing reading. Table 4.5. presents key questions and features that were included in the “*Construct of the IELTS RCM Questionnaire*”.

Table 4.5.

Construct of IELTS RCM Questionnaire

Level of reading	Word	Phrase	Sentence	Inter-sentence	Paragraph	Text
Knowledge needed	Lexical	Grammatical (sentential/ inter-sentential)		Textual (cohesion, coherence)	World knowledge	Others
Type of comprehension	Literal	Inferential		Evaluative		
Type of Reading	Careful ... 1.....	10.....			Expeditious	
Difficulty of the text	Easy 1.....	10.....			Difficult	
Difficulty of the task	Easy 1.....	10...			Difficult	
Level of task processing	Low level ... 1.....	10...			High level	
Time needed	<input type="checkbox"/> -1m.	<input type="checkbox"/> 1m	<input type="checkbox"/> 90s	<input type="checkbox"/> 2ms	<input type="checkbox"/> more	
Reading skills measured	<input type="checkbox"/> Understanding the main idea <input type="checkbox"/> Paragraph structure <input type="checkbox"/> Vocabulary knowledge <input type="checkbox"/> Sentence comprehension <input type="checkbox"/> Lexical Inferencing <input type="checkbox"/> World knowledge inferencing <input type="checkbox"/> Scanning <input type="checkbox"/> Text type knowledge Other skills:		<input type="checkbox"/> Understanding details <input type="checkbox"/> Text structure <input type="checkbox"/> Grammar knowledge <input type="checkbox"/> Inter-sentential comprehension <input type="checkbox"/> Text based inferencing <input type="checkbox"/> Skimming <input type="checkbox"/> Writer’s Attitude <input type="checkbox"/> Reading speed			

The questionnaire employed a set of different types of questions to address these features of the reading construct in the IELTS RCM. For some items, such as the level of reading and knowledge required, the questions were in closed-ended formats. For other items, an ordinal rating scale format (1-10) was used to explore experts’ judgements on 1) speed of reading (careful-expeditious), 2) difficulty of the text (easy-difficult), 3) difficulty of the task (easy-difficult), and 4) level of processing (high-low). These items were weighed on a scale of 1-10. For skills measured, based on the existing literature, a list of 16 reading skills was listed. Finally, two open-ended questions addressing further skills that might be tapped by the test task and

further comments were also included in the questionnaire. The questionnaire was completed by all the testing experts after they took the sample IELTS RCM.

4.6. Procedures (Data Collection)

As shown in Figure 4.1, data collection procedures included three main phases; 1) selecting the sample IELTS RCM; 2) administering English language proficiency tests and the main instruments of the research (IELTS RCM test) to the sample of L1 and L2 test takers; and, 3) administering the sample IELTS RCM test and the *IELTS RCM test construct questionnaire* to the sample of testing experts. In phase two, each test taker attended three separate sessions for data collection while the experts required only one session. Data were collected during spring 2019 in Canada and summer 2019 in Iran. Details of data collection procedures are presented below.

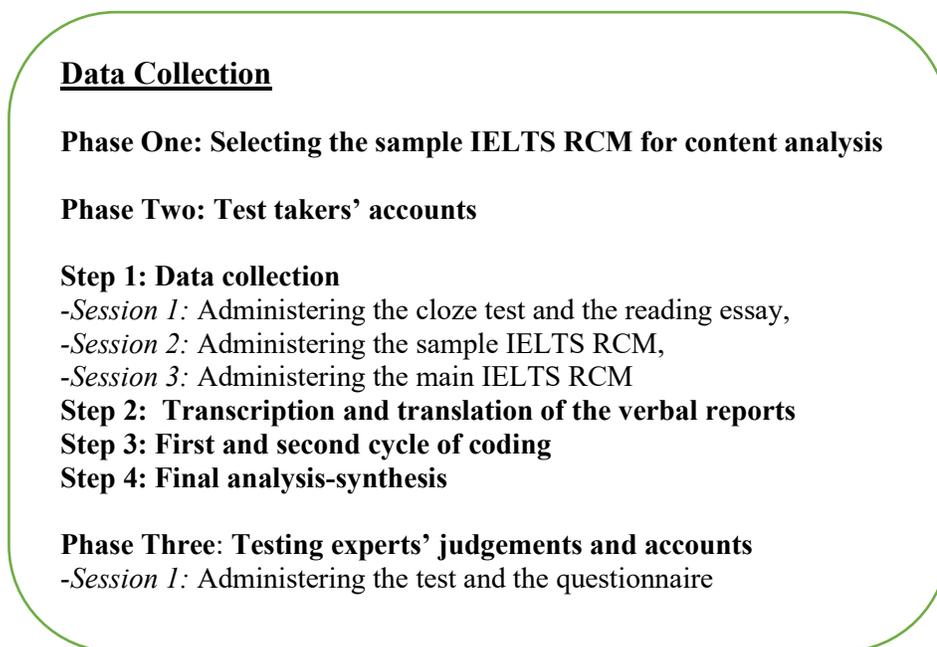


Figure 4.1. Data collection and data analysis procedures

4.6.1. Phase one: The sample IELTS RCM

In the first phase of data collection, an actual but retired sample IELTS RCM published by Cambridge English Language Assessment, IELTS Academic 11, 2016) was selected for content analysis and administration to the research participants.

4.6.2. Phase two: Test takers' accounts

4.6.2.1. Step one: Data collection

4.6.2.1.1. Session one: Measures of language proficiency and reading ability

In the first step, data was collected from test takers participants who participated in three sessions of data collection. Each session took 70-90 minutes. The sessions were held separately for each individual participant. The first session aimed at assessing the language proficiency and reading ability of the participants. They were introduced to the main research purpose and procedures of data collection so that they know what they are expected to do during data collection procedures. Each participant took the 40-item cloze test and the reading essay, which included five essay-type questions. Participants were also asked to write a brief summary of the essay and provide an oral summary of the essay. These written summaries were used in the assessment of the language proficiency of the EFL test takers. However, it is worth mentioning that all the data collected in the first phase of data collection was more relevant to the assessment of language proficiency and reading ability of the L2 participants and not the L1 English speakers. As they were all doing their major of interest at undergraduate level, L1 English speakers were assumed to be a relatively homogenous group in terms of their level of language proficiency; it made no sense to measure their language proficiency. However, they took the same instruments and tests and provided some index for comparison with the L2 participants. In brief, the first session focused on collecting data for assessing participants' language proficiency and reading ability.

4.6.2.1.2 Session two: Orientation with IELTS and verbal reporting

In the second session, which was held at the convenience of each participant, a sample IELTS RCM was administered to familiarize them with the test and its different test tasks. This session also helped test takers practice how to present their retrospective verbal accounts of what they did during their test performance. The participants took the sample test and did the test within a time limit of 60 minutes. After reading each passage and doing the test tasks, they were asked to immediately report what they did in terms of the activities, strategies, processes, and thoughts they had in mind, as well as the steps they took in doing each test task. Their immediate retrospection provided fresh access to details of the cognitive processes involved in test performance and opened a window to examine the skills, knowledge sources, processes, strategies (SKPSs) the participants used.

In brief, this phase of data collection was meant to help each participant in two significant ways: 1) get them fully familiar with the IELTS RCM test tasks and 2) practice the use of immediate retrospection and provide a clear account of what they did while performing on the test. This orientation session helped participants be more aware of what they were expected to do, which in turn contributed to the reliability and validity of the retrospective data collected in the third session of data collection.

4.6.2.1.3. Session three: The main instrument (IELTS RCM)

The third session was the main and final session of data collection. During this session the main sample of IELTS RCM was administered. The data collected in this session was used for data analysis and addressing the research questions. In this session, participants took the sample IELTS RCM test within a time limit of 60 minutes. Each of the three reading passages of the test was followed by three different test tasks. Each participant was carefully observed for their test behavior in terms of the amount of time spent reading each text, the amount of time spent on each type of test task, and the frequency of back and forth movement between the test task and the text. The Test Performance Observation Scheme (TPOS). The TPOS was used for more reliable recording of the participants' reading behavior. These observations were carefully recorded in the observation scheme. Later these observations were cross-checked with the test takers' accounts of the processes and strategies they used during test performance

After they completed each test task, each participant provided their account of test performance in their L1, which was Farsi. Their accounts were audio recorded for later transcription and data analysis. While recording, care was taken to have a high acoustic quality of the verbal reports which could secure and improve data transcription reliability. Zoom H1n which is a portable, hand-held voice recorder was used for audio recordings. In their immediate retrospection, participants talked about what they actually did, the steps they took, what they thought, the certainties and uncertainties they had, and the challenges they faced while doing the test tasks. They described how they did the test task, what strategies used, what specific reading activities they did and what they were thinking and doing while answering the test task. They were also asked to add any further points and comments to their performance when they finished their retrospection account. They answered some questions about different features of the IELTS texts and the test tasks, such as the perceived difficulty of the texts and the tasks, and if they

found the texts and the test tasks similar to, or different from, their academic texts and academic reading activities.

Four steps were taken in the codification and analysis of the data; 1) transcription of the retrospective verbal accounts, 2) translation of the Farsi transcripts into English, 3) first cycle of coding, and 4) second cycle of coding. Next section presents details of these steps.

4.6.2.2. Step one: Transcription of the retrospective verbal reports

There is consensus among research methodologists that transcription of retrospective data from oral to written mode is a form of closer analysis and in itself is a form of initial analysis (Rapley, 2007). It involves a kind of abstraction of the live face to face oral interaction to a more formal literal text. Therefore, accurate transcription of the data can play a crucial role in reliability of the data and validity of the analysis. With this in mind, the first step in data analysis started with transcription of the retrospective verbal reports provided by the test takers.

Transcription of the audio file was a tedious and time-consuming process. Each participant had produced some 40- 75 minutes of audio. The transcriber carefully listened to each audio file and translated it. Whenever necessary certain sections of the audio files were replayed for accurate transcription. All the audio files were transcribed based on the current transcription conventions. For ease of codification and better readability of the transcripts, detail such as false starts, hesitations, and repetitions were all excluded. This could facilitate communication of meanings to the readers and coders in later stages of data coding and data analysis. Transcription of the data was manually done by the researcher. He had collected the data and could better preserve the contextual and emotional aspects of the participants' retrospective verbal reports. The data were transcribed word by word and captured ideas and accounts expressed by the participants. During transcription, the semi-formal accounts of the participants and their statements were transformed into a more formal style. This could provide a clear idea what each participant was talking about. For example, Mart, one of the participants said, "No process of elimination", she meant "(I did not use) process of elimination.

To make sure the transcripts were reliable and representative of the accounts given by the test takers and no significant part was missed from the original audio files, the transcripts were also double checked with the audio files. No major differences were observed in the transcripts and only few minor changes were made in the first transcript.

4.6.2.3. Step two: Translation of the Farsi transcripts into English

The next step was translation of the Farsi transcripts into English. L2 participants presented their accounts of the test processes in their L1 (Farsi) because they could express their ideas in their L1 more efficiently and add as much details and descriptions to their accounts. Their accounts, then, had to be translated to English for consistent coding. The rationale was that two sets of transcripts, one in English (by the L1 English speakers) and one in Farsi (by the EFL participants) might be a source of variability and unreliability of coding. Therefore, it was decided to have all the transcripts in English. A Farsi native speaker, who was an official Farsi translator, translated all the Farsi transcripts into English. Translation of the transcripts was done with full awareness that it might contaminate the data and change some minor features of the original Farsi transcript. A sample of the translations were then back translated for accuracy to check that the original ideas and details of the transcripts had been rendered accurately. Further, because the translator was fully aware of what the Farsi speakers said and how they said it and considered the intricacies involved in translation of the transcripts. He reread the transcripts and replayed the recordings to make sure the English translations are accuracy.

4.6.2.4. Step three: Coding the data (First cycle of coding)

Coding as one of the most common forms of qualitative data analysis can organize the data and put them into different categories. As shown in Figure 4.2. Saldaña (2009) argued that coding is a multi-cycle process. As coding proceeds more codes emerge from sub-codes which are then grouped into “categories” and “themes”. In fact, in the first cycle of coding a set of sub-codes and codes that best describe and summarize the data emerge. In second and third cycles of coding, interrelated codes are grouped into a category. In the final stages of coding, themes may emerge from a set of categories. Finally, themes generate the final theory that explains the phenomenon under study. In brief, his model of data coding suggests moving from real data to theoretical concepts by grouping sub-codes into codes, sub-categories into categories and categories into themes and themes to a theory.

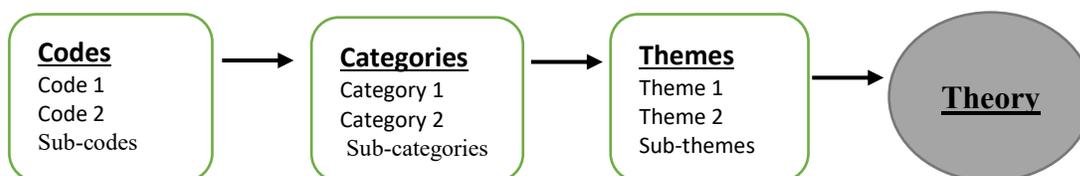


Figure 4.2. Coding procedures (Adopted from Saldaña, 2009)

4.6.2.5. Step Four: Data coding (first and second cycles)

In the fourth step of data analysis, the transcribed data were coded for capturing the main features of test performance. Saldaña (2009) suggests adopting a multi-cycle approach for data coding. The first cycle wrestles with the data while the second cycle wrestles with the codes and their organization into categories. In fact, in the first cycle of coding one or more coding methods are used to capture the basic features, ideas, and actions observed in the data while in the second cycle these codes are integrated into some *meta codes* to capture the similarities among the codes for finding possible pattern(s) in the data. According to Saldaña (2009), the first cycle of coding mirrors the actual data and develops an inventory of codes while the second cycle focuses on summarizing these codes by applying some theoretical concepts or meta codes to retrieve categories and explore any possible pattern in the data. The first and the second cycles of coding are further discussed below.

The first cycle of coding attends to the message of the data by coding the actions and ideas that are talked about by the participants. Most codes of this cycle address observable or conceptual action in the data. In *descriptive coding*, descriptive words and phrases can be used to capture the substance of what is talked about. Another method that can be used in the first cycle is *pattern coding* which is basically characterized by using gerunds (-ing) that capture both behavioral and conceptual action. *Initial coding* also searches for processes, properties, dimensions of categories, conceptual ideas in the data. It is hard to draw a line where each of these coding methods begin or end because there is so much overlap between them and each is capable capturing some features of the data. However, the bottom line is that they all end up capturing the main actions, concepts, and messages in the data. In the context of this research for the first cycle of coding, descriptive coding, process coding, initial coding, and exploratory coding methods were used to capture the specific features observed in the data. Such a flexible approach to coding could guarantee validity of the codes and capturing as much in the data as possible.

In the second cycle of coding, codes are related to each other to provide specificity that is needed to interpret the data. As Saldaña (2009) argues,

The primary goal during second cycle coding is to develop a sense of categorical, thematic, conceptual, and/or theoretical organization from your array of first cycle codes. (p.149).

The focus in this cycle was on organizing codes into categories and coherent themes. This was achieved by integrating codes into categories (*meta codes*) which are abstract concepts that identify similarly coded data and help coders organize the corpus by attributing meaning to that organization.

In the second cycle of coding, the coded passages were retrieved and inspected for recoding and integration into categories. The first and second cycles of coding revealed the core skills, knowledge sources processes, and strategies (SKSPs) used in taking IELTS RCM.

Similar to the first cycle of coding, the coding methods used in second cycle coding applied different approaches to coding, including: focused coding – which identifies the most recurrent or salient codes; axial coding, which extends or expands a category to sub-categories and specific features and details; and, theoretical coding – which relates codes to umbrella, core categories, or meta codes (Saldaña, 2009). In other words, codes that belonged to the same or similar categories and subcategories were merged into a more holistic and conceptual code. Based on this conceptualization of coding, the verbal data were analyzed. Next section provides details of data analysis of the test takers' accounts.

4.6.3. Phase three: testing experts' judgements and accounts

In the third phase of data collection, 10 testing experts with diverse backgrounds in their level of education and language teaching and language testing experience participated in the second phase of the study. First, the main categories of the *Construct of IELTS RCM Questionnaire* (See instruments, section 4.5.7 above) were introduced, explained, and discussed. Then, they took the sample IELTS RCM test and did it within the time limit set by the test for each section of the test (20 minutes) and then provided their accounts and judgements of different test tasks by filling in *The Construct of the IELTS Questionnaire*. They also provided further elaborations on their judgements of the test tasks. The idea was that they would be in a much better position to judge the test tasks if they actually took the test. Actual test performance could provide them with a fresh and authentic context to judge each test task. However, (2/10) experts preferred to study each passage of the test for some 10-15 minutes instead of taking the test. They provided their judgements and evaluation of the test tasks without taking the test. I so that they could provide further details, comments, and explanations, if they had any. They also commented on the relevance of reading construct of IELTS RCM to the construct of academic

reading. They talked about if and how the IELTS text and test tasks could be compared to academic texts and typical academic reading practices and activities.

4.7. Data Analysis

Data analysis included three phases, 1) content analysis of the test, 2) analysis of the participants verbal reports and 3) analysis of the experts' judgements and accounts. The coding of the L1 and L2 test takers accounts involved *bottom up*, first and second cycle coding approaches: "When you apply and reapply codes" to verbatim data, in "a process that permits data to be divided, grouped, reorganized and linked in order to consolidate meaning and develop explanation" (Saldaña, 2009, p.9). However, content analysis of the test and the data elicited from the testing experts was coded top-down protocol coding (Saldaña, 2009) because both involved coding according to some specifics, pre-established classifications (e.g., Construct of the IELTS RCM Questionnaire).

Analysis of the Data:

Phase 1: Content analysis of the test

Phase 2:

-Analysis of language proficiency tests

-Analysis of the test takers' verbal reports

-First cycle of coding: Processes and strategies (categories)

-Second cycle of coding: Themes

Phase 3: Analysis of the experts' judgements and accounts

Figure 4.3. Phases of data analysis Data

4.7.1. Phase one: analysis the IELTS RCM

For content analysis of the sample IELTS RCM which consisted of three texts and nine test tasks, the texts and the test tasks were separately analyzed. The purpose was to describe those features of the texts and the test tasks that were deemed relevant to the construct validity of the test. Each text was analyzed in terms of topical features, textual features, and linguistic features, such as readability, text length, average sentence length, and lexical profile features including, general words, academic words, and off-list words. The test tasks were also analyzed in terms of their format, language features, the specific reading skill targeted by the test task, the

level at which the task could be answered, lexical clues and features used, the type of comprehension involved (literal, inferential, evaluative), and how the text and the test tasks were related to each other. Chapter Five presents the results of the Analysis.

4.7.2. Phase two: Analysis of test takers' accounts

4.7.2.1. Analysis of language Proficiency and reading proficiency tests

As explained previously in Section 4.5., to assess the language proficiency of the participants, three measures of language proficiency and reading proficiency were used: 1) a cloze test, 2) a reading essay with some open-ended questions, and 3) a written summary. Using different measures of language proficiency improved the reliability of the assessment and was a means of dividing the L2 research participants into different proficiency groups (low – high) (Henning, 1987).

First, cloze test and reading essay were scored. The cloze test was scored using accepted scoring methods. Participants' responses to the essay questions were judged for 1) language and 2) content by experienced writing instructors on a scale of 5 band scores (1-5), with each level of the scale representing one level of language proficiency (advanced, upper intermediate, intermediate, lower intermediate, and beginner). As shown in Table 4.6, criteria for *language* included appropriate vocabulary, grammar and organization. As for content, developing an argument, task achievement, and academic tone were set as rating criteria.

Table 4. 6.
Rating scale for essay questions

Criteria		Levels
Language	Vocabulary	1.....5
	Grammar	1.....5
	Organization	1.....5
Content	Development of an argument	1.....5
	Task achievement	1.....5
	Academic tone	1.....5

For the written summaries two experienced writing instructor raters rated the written summaries for five criteria: 1) accuracy, 2) grammatical range, 3) lexical resourcefulness, 4) task achievement, and 5) coherence and cohesion. These criteria were assessed based on a scale of a 5-band score. These are the same criteria used to assess the L2 test takers' performance on the Writing Module of IELTS. (See Appendix J for rating scale). Task achievement which was defined as the extent to which the essay has been appropriately

summarized with the main ideas highlighted and supported by content presented in the essay. Coherence and cohesion were defined as use of logical meaningful sequencing of ideas in the summary and appropriate use of cohesive devices, respectively. For lexical resourcefulness, it was operationalized as the use a wide range of lexical items. Finally, for grammatical range and accuracy it was operationalized as accurate use of various grammatical structures that help clear and fluent development of the summary. The two raters were asked to rate and report the results of the written summary on a scale of 1-5 (advanced learner, upper intermediate, intermediate, lower intermediate, and beginner). Inter-rater reliability of the ratings enjoyed 84% agreement.

4.7.2.2. Results of Language Proficiency and Reading Proficiency of the test takers

Based on the results of test takers' score on the cloze test, reading essay, and written summary, the participants were placed at different levels of language proficiency (advanced, upper intermediate, intermediate, lower intermediate, beginner). Table 4.7 presents the results of these measures of language proficiency and reading proficiency for the three groups of participants.

Table 4.7.

Results of language proficiency testing of L1 and L2 participants)

Instruments	L1 test takers	L2: Successful test takers	L2: Less successful test takers
Cloze test: Mean score	37/40	32/40	23/40
Reading essay levels	Advanced (5)	Upper intermediate (4-5)	Lower intermediate (2-3)
Written summary levels	5	4	2

As shown in Table 4.7, participants were at different levels of language proficiency and reading proficiency. The cloze mean score for the L1 English participants was 37. There was a big score gap after the 5th top score for the L2 participants. Therefore, the top 5 L2 participants were considered at upper intermediate levels of language proficiency, and the next 6 L2 participants were considered to be at lower intermediate level of language proficiency. The mean score for each group was also calculated. For the upper intermediate (more proficient) group and the lower intermediate (less proficient) group, the mean scores were 32 and 23, respectively.

Results of the reading essay showed that all the L1 participants were placed at advanced levels of the band scores of 5. For the L2 participants, 5 were placed at band score 4 and the other 6 were placed at band score 2. For the written summary, the L1 participants were all rated as advanced, while the L2 participants were placed at two levels of language proficiency: upper intermediate and lower intermediate. Based on their ratings, five L2 participants were scored at

level 4 and six participants were scored at level 2. Overall, results of the reading essay and the written summary matched the results of the cloze test scores, i.e., the participants who were placed at higher or lower levels of language proficiency were consistently the same across cloze test, the reading essay, and the written summary. It is worth mentioning that due to the small sample size of the participants, the groups could not be tested for significant statistical differences.

4.7.2.3. Results of IELTS RCM of the test takers

As mentioned in Chapter Two, the IELTS RCM includes nine band scores. Each band score corresponds to a range of actual (raw point) scores. Based on the number of correct answers in the 40-item test, a band score is assigned to the test score. Table 4.8 presents the band scores and corresponding test scores for the Academic IELTS RCM as specified in (w.w.w.IELTS.org).

Table 4.8.

IELTS Academic Reading marking scheme

Band Score	9	8.5	8	7.5	7	6.5	6	5.5	5	4.5	4	3.5	3	2.5
Score/40	39-40	37-38	35-36	33-34	30-32	27-29	23-26	19-22	15-18	13-14	10-12	8-9	6-7	4-5

Based on the test takers' scores on the IELTS RCM sample test, and the marking scheme for IELTS Academic Reading (Table 6.3.) test takers were placed in one of the band scores that corresponded with their actual test scores. Test takers at band score 6.5 and above were termed as more successful L2 test takers and test takers who scored lower than 6.5 were labelled as less successful L2 test takers. Level of success was defined on the basis of admission criteria applied by most admission offices of English-speaking universities including Canadian and American universities. Scores below 6.5 are rejected and scores above 6.5 meet the minimum requirements for admission. Table 4.9 presents results of the test takers' band scores and their corresponding test scores.

Table 4.9.
Results of test takers' scores on the IELTS RCM and their IELTS band scores

Participants	1	2	3	4	5	6	7	8	9	10	11
L1 Test takers	32	34	34	35	36	37	37	38	38	39	-
Band score	7	7.5	7.5	8	8	8.5	8.5	8.5	8.5	9	-
L2 Test takers	20	20	22	24	25	26	33	33	38	38	39
Band scores	<u>5.5</u>	<u>5.5</u>	<u>5.5</u>	<u>6</u>	<u>6</u>	<u>6</u>	<u>7.5</u>	<u>7.5</u>	<u>8.5</u>	<u>8.5</u>	<u>9</u>

As shown in Table 4.9, all the L1 speakers met the 6.5 standard for admission. For English L1 speakers, one participant scored at band score 7, two of them scored at band score 7.5, two scored at band score 8, four participants made it to band score 8.5, and one received the highest band score, which is 9. As for the L2 test takers, two participants were at band (7-7.5) and three participants were at the highest level of the IELTS reading scale (8.5-9). These five participants were deemed to be more successful L2 test takers. The less successful L2 test takers included six participants, three of whom scored at band score 5.5 and three at band score 6.

It is worth noting that results of the IELTS RCM for the L2 test takers were consistent with the results of measures of language proficiency and reading proficiency as reported in the previous section. In both sets of tests, they were positioned at the same level of language proficiency and reading proficiency. Another point that should be highlighted here relates to the rationale for using three categories of test takers (English L1 speakers, more successful L2 test takers and less successful L2 test takers). Arguably, different levels of language proficiency and reading ability would enrich evidence regarding the construct(s) operationalized by different test tasks and the IELTS RCM as a whole. It was anticipated that L1 and L2 test takers would use different skills, knowledge sources, processes and strategies (SKSPs) during test performance, and that high-low proficiency L2 test takers would also differ.

4.7.2.4. NVivo12

For coding and analysis of the transcript, NVivo12 was used. NVivo is a qualitative data analysis tool and coding application that is extensively used when researchers work with qualitative data such as discourse studies, media data, interview data and text data. It is commonly used for the analysis of a wide range of data such as audio, video, image, unstructured text, interview, focus group, and social media. It is a useful tool for data analysis.

The software is produced by QSR International and supports deep levels of data analysis. It includes different options for codification and different queries such as memo writing, coding, searching for key words, doing word counts, making graphic displays, screen coding, cross-tabulation, comparison of the data, text searching, matrix coding, and word cloud presentation of the data. The node function in particular serves as a container that collects all the material related to one code together so that the researcher can easily access all the references in the data to the that code.

In the context of the current research, immediate retrospective data and interview data were collected for addressing the research questions. The retrospective verbal reports provided participants' accounts of what they did while performing on the test. The collected data were first transcribed for codification. Next, they were uploaded for codification in NVivo. The application was used for coding all the data collected for phase one of the study and facilitated the creation, emergence, and integration of several codes that were used in the test taking processes. NVivo, just like any data analysis software, was a strong aid for thought but as Weitzman and Miles (1995) indicated it does not substitute thought. NVivo facilitated data analysis by structuring the interview material for further analysis but as Kvale (2011) indicated, the task and the responsibility of interpretation remained with the researcher.

4.7.2.5. First cycle of coding

In practice, for accurate and reliable coding, first a sample of three transcripts were randomly selected for initial coding, one from the L1 English test takers, one from the more successful L2 test takers, and one from less successful L2 test takers were. The transcripts were carefully read and reread and whatever codes that the data seemed to suggest were used for coding. Some of the codes that emerged from the data were directly mentioned by the test takers in their retrospective verbal reports. For example, for the code "moving back and forth between the text and the test task" Abed, one of the L2 test takers, said, "*I went back to the passage to see where I can locate the relevant info.*" In another example one of the test takers said, "*I went back to the text to make sure that I did not accidentally miss this point.*" This suggested the code "moving back and forth between the text and the task" and coded as such. Another test taker said,

Then for the next item, what they used as a food source. I thought it was birds, I double checked and the same sentence in the text which says, they ate birds. (Angl)

the codes “guessing an answer”, and “double checking the answer” seemed to best capture the statement made in the transcript. Or for the code “reading the hole text”, Broo said, *“Then I read through the whole text word by word from the very beginning to the end.”* Another example is provided by Angle who said, *“I think now that I should have crossed out the headings that I chose for each paragraph which could have helped me focus more on fewer options. I did it later in the middle of doing the test.”*

This statement formed the code “eliminating some options”.

As the codes emerged from the data, they were defined and supplemented with one or two examples. More examples were then compared to the original examples and definition of the code to make sure it is consistently applied to more instances. Wherever necessary, changes were made in the definition of the codes to accommodate more instances that were coded in the data. However, to make sure the coding was on the right track and was reliable, a PhD candidate was asked to code the same transcripts using whatever first cycle coding methods that seemed appropriate to the data, reminding her that descriptive coding, pattern coding, and In Vivo coding seem more relevant to the data. She coded the sample transcripts. Her codes were then compared with the codes that had already emerged from the trial coding. There was 88% agreement between the number and types of codes that emerged from the two sets of codes, supporting the validity of the approach adopted and reliability of the results produced.

After more reflection and modification of the emerging codes and getting some feedback from the second coder, the initial coding procedure was applied to coding all the data. This phase of coding was conducted by the researcher. It is worth mentioning that during coding the transcripts some infrequent new codes emerged which were added to the list of codes. For example, the code “tagging the paragraph with a word or phrase”, was observed just in one transcript. As suggested by Saldaña (2009), personal debriefing, memo-writing, and reality checks were used during this phase of data analysis. They proved very instrumental in different stages of coding. For instance, memo writing helped define each code for later use and more control over codification.

Data coding produced a list of codes (code book), each describing and summarizing part of the test takers' accounts and the processes and strategies they used in their test performance. Some of the codes that emerged from the first cycle of coding are listed in Table 4.10. (For the full list of the codes and actual examples from which they emerged, see appendices K, L, and M.) Results of the first cycle of coding and second cycle of coding for the test takers are presented in Chapter Six.

Table 4.10.

Some codes emerging from first cycle of coding

Answering the test task
 Delay answering the test item
 Eliminating some of the choices,
 Highlighting key words/phrases in text
 Highlighting key words and phrases in the test task
 Highlighting the answer in the text,
 Moving back and forth between the text and the task
 Reading the text or part of the text
 Reading the test task
 Double checking the answer
 Use previous reading
 Focus on doing the test task
 Read the paragraph carefully
 Use of lexical clues
 Reading the last sentence in the paragraph
 Reading the whole text
 Guessing an answer
 Eliminating the options
 Re-reading part of the paragraph,
 Scanning specific words/phrases in the text
 Skimming the whole paragraph

Next, to improve reliability of coding, a second coder was invited to code a sample of three transcripts. She was briefed about how the codes had emerged for the transcripts and received instruction about the coding processes. She also received the same transcripts and the code book for coding the sample transcripts. Before coding she practiced coding a sample as part of her training, too. Results of her coding was then used for estimating inter-coder reliability. Inter-coder percentage agreement was calculated at .92 agreement, which indicates a high index of intercoder reliability.

4.7.3. Second cycle of coding

To exploring the cognitive evidence for the construct of the IELTS RCM in terms of the skills, knowledge sources, strategies, and processes (SKSPs) used in test performance, the second cycle of coding aimed at integration of the codes that emerged from the first cycle of coding into categories and then into themes. Results of the first cycle of coding were not more than a set of discrete isolated codes that accounted for different parts of the test takers' accounts. They needed to be integrated into meaningful categories and themes. In this phase of data analysis, the categories were integrated and themed to identify the main components of the IELTS RCM construct. In the final analysis, the categories that emerged from the codes were integrated and themed to sketch a model of the reading construct for the IELTS RCM. Theming the codes resulted in three main themes

Results of data analysis for each research question are fully reported and described in the next three chapters. Chapter Five presents result of test content analysis. Chapter Six presents result of test takers' accounts and Chapter Seven presents results of testing experts' accounts and judgements.

4.7.4. Phase three: Analysis of testing experts' judgements and accounts

Phase two: Analysis of the experts' judgements and accounts

In the final phase of data analysis, testing experts' judgements were coded on the basis of their responses to the IELTS RCM validity questionnaire and their accounts of the sample test.

Results are reported in Chapter Seven.

4.8. Summary of the Chapter

This chapter presented details of the methodology and procedures employed for data collection and data analysis which included details of the multiple case design adopted in the study including research participants recruitment, the test instruments used, the retrospective verbal report method used for data collection, and the two-cycle processes of data coding.

The study aimed at collecting validity evidence from three different sources including, test content, test takers, and testing experts and adopted a multiple case study design to provide more comprehensive evidence for the validity of the IELTS RCM. A sample of IELTS RCM that was used as the main research instrument was analyzed in terms of linguistic, discoursal, topical, and cultural features. For data collection, three groups of participants from different language

backgrounds and levels of language ability and a group of testing experts participated in the study. In addition to a test of general language proficiency, test takers also took a sample of IELTS RCM and provided their retrospective accounts of the processes and strategies they used during test performance. Testing experts took the IELTS RCM sample test, provided retrospective verbal accounts of their test taking experience, and responded to a validity questionnaire to record their judgements about each test task. Following Saldaña's (2009) coding guidelines, two cycles of coding were used for coding the verbal reports of the test takers. The first cycle of coding summarized the main activities observed in the verbal reports while the second cycle integrated the codes into categories and themes.

Results are presented and discussed in the following three chapters, by phase. Chapter Five reports on the phase one results of the content analysis; Chapter Six on phase two, the test takers' accounts; and Chapter Seven, on the experts' judgements and accounts of the IELTS RCM. In Chapter Eight the results of all three chapters are merged and discussed in relation to the research questions guiding the study. In the final chapter, Chapter Nine, the implications are discussed and directions for future research are identified.

CHAPTER FIVE: RESULTS: CONTENT ANALYSIS

5.1. Introduction

The overarching question of the study was exploring cognitive evidence of the construct of reading comprehension of the Reading Comprehension Module of IELTS. (IELTS RCM) in terms of the knowledge sources, skills, processes, and strategies (KSPSSs) used in test performance of test takers. For practicality and methodological reasons, the question was broken down into four research questions. The first question was

1. What does content analysis of the IELTS RCM test tasks in terms of linguistic, textual, and topical features of the texts and the test tasks reveal about the construct of the test?

The focus of the first research question was on collecting some construct validity evidence from the test content analysis, namely, analysis of the texts used in the test and the nine test tasks which comprised of 3-6 items each. Analysis of test content can provide direct evidence as to what the test purpose is and what each specific test task and test items measure. It can also provide more context for understanding the processes and strategies used by the test takers and allow comparison of the test tasks with the real-life domain of academic reading as practiced by university students and academicians. Content analysis of the IELTS RCM was carried out to explore some of the main features of the texts and the test tasks and individual test items and to assess some of the main linguistic and textual properties of the text such as what reading skills are tapped by each test task, what level of processing is required, and how features of texts and test tasks operationalize construct of the test task. In addition to the analysis of test items in terms of format, purpose, and linguistic characteristics, Green et al. (2010) framework was adopted for content analysis of the texts. More specifically, content analysis focused on text readability, lexical density, lexical properties of the text, text length, topical knowledge, cultural knowledge, and degree of text abstractness. This chapter reports on the results related to the content analysis of the sample IELTS RCM sample test used in the study.

5.2. Results of Test Task Analysis of the Sample IELTS RCM

As mentioned earlier in Chapter Three, the sample IELTS RCM consisted of three texts each followed by three different test tasks. Content analysis of the IELTS RCM involved two phases. In the first phase, linguistic, textual, and topical features of each text and focused on the analysis of the following features; length, syntactic complexity, lexical density, readability, genre, topic, cultural knowledge, and degree of abstractness of the texts used in the test were assessed and analyzed. In the second phase task features of the 9 test tasks and the test items were assessed in terms of 1) the intended skills to be measured, 2) the lexical clues and synonyms used, 3) the relevant information in the text that contained the answer to the test items, and 4) the type of comprehension (literal, inferential, evaluative) involved,.

5.2.1. Passage One- “Raising the Mary Rose”

The first text, i.e., “*Raising the Mary Rose*”, was a 7-paragraph text which described the story and history of a ship “The Mary Rose” and details of its sinking, how it was discovered and how it was raised. It looked like a very short report of a historical event. The text contained many dates and proper names referring to the timeline of the whole story and the people involved. The last three paragraphs provided some technical description of how it was raised. The text was straightforward and factual and did not seem to require much background knowledge or cultural knowledge, but the last section of the text was somehow more technical and readers with background knowledge in engineering or readers familiar with historic texts may do better in reading and processing the text. One of the test tasks, i.e., the *Diagram Completion Task* was devoted to this technical part and dealt with the processes related to the raising of the Mary Rose.

The text was analyzed for its textual features. As shown in Table 5.1, compared with academic text which are usually thousands of words, the text was 891 words long. Average sentence length was 23 words which clearly indicated the sentences used were syntactically complex and/or compound complex sentences, meaning the sentences consisted of two or more independent and dependent clauses. Lexical density of the text was .56 which is relatively dense. Based on the readability indices calculated

the “*Raising the Mary Rose*”, text was a 12th grade level, meaning it is what 12 grade high school readers read. In terms of genre or text type, the text was descriptive and looked like a short report article published in a history magazine that reports historical discoveries.

As to the cultural knowledge required for reading the text, the text was not culturally sensitive or culturally loaded. It was related to warships and ship battles which is more relevant to readers who come from a geographical or historical background where warships, ship battles, and navy vessels are part of their country and history. For readers who come from a geography surrounded by lands, a ship is just a word and not part of their history and public knowledge. Finally, in terms of text abstractness, the text was basically factual and looked like a reportage of what had happened to the ship and how it was raised.

Table 5.1.

Textual features of the “Raising the Mary Rose” text

Text features	Raising the Mary Rose
Text length	891
Sentence length	23
Lexical density	0.56
Readability	12 th grade
Genre	Magazine short report
Topical Knowledge	To some extent
Cultural knowledge	Moderate
Text abstractness	Very concrete

Further, to examine lexical properties of the IELTS RCM texts and assess the proportion of the General Service List (GSL), Academic Word Lists (AWL) and off-list words used in the text, an online vocabulary profile calculator (www.lextutor.ca) was used. Table 5. 2 shows results of lexical profile for the Mary Rose text.

Table 5.2.

Lexical profile of the “Mary Rose text (IELTS RCM)

The Mary Rose	Families	Types	Tokens	Percent	Cumulative frequency
K1 words	233	271	729	72.75	72.75
K2 words	54	61	75	7.49%	80.24
AWL	48	53	61	6.09%	86.33
Sub list 1: approach area available factor period require research structure vary					
Sub list 2: design distinct final institute maintain site survey text transfer					
Sub list 3: framework layer locate max technique technology					
Sub list 4: approximate errors option project					
Sub list 5: academic expose network precise unaware					
Sub list 6: attach initiated input recover reveal					
Sub list 7: classic equipment files submit survive					
Sub list 8: minimise via					
Off-list words	-	85	137	13.67%	100.00

As shown in Table 5. 2, a great proportion of the words used in the text 72.75% (N= 233) were K1 words (the first 1000 words of the GSL) and only 7.49% (N=54) words were K2 words (the second 1000 words of the GSL). Results also produced lists of word from each sub-list.

Moreover, words from the AWL made 6% (N=48) of the words and off-list words covered 13.67% of the words. The off-list words seem to be a considerable number of words in the IELTS RCM text. Size of off-list words are very important in reading comprehension test because they can directly influence reading performance of the test takers. It is worth noting that in these indices, N refers to the numbers of word families used in the text.

Having analyzed linguistic and textual features of the first text used in the IELTS RCM, now I turn to the analysis of the three tasks that followed the text; 1) the *True, False, Not given Task*, 2) the *Matching Features Task*, and 3) the *Diagram Completion Task*. These tasks and the items included in each were analyzed for their content and format.

5.2.1.1. Task one: The True, False, Not given Task

As shown in Table 5.3, the *True, False, Not given Task* included 4 items. Each item was a statement whose truth was expected to be judged against the text. The task items were ordered based on the order of paragraphs in the text. The ordering of the task items could help test takers in following the order of the items in their search for the answer.

Table 5.3.

The True, False, Not given Task

-
- 1 There is some doubt about what caused the Mary Rose to sink.
 - 2 The Mary Rose was the only ship to sink in the battle of 19 July 1545.
 - 3 Most of one side of the Mary Rose lay undamaged under the sea.
 - 4 Alexander McKee knew that the wreck would contain many valuable historical objects.
-

Some of the statements included a word or a phrase that was not explicitly mentioned in the paragraph which complicate finding the answer and demand more careful reading of the relevant information to decide if the statement was true, false, or not given. Such items involved inferencing or lexical inferencing. For instance, in item 1 “*There is some doubt about what caused the Mary Rose to sink*”, the task focused on when and how the ship sank. The item included the word “doubt” which was not explicitly mentioned in the text. Therefore, the item needed more scrutiny before it could be answered. It involved knowledge of synonym and some lexical inferencing skills. The answer to this item lied in one single sentence that said it all; “*Accounts of what happened to the ship vary*” the word “vary” connoted uncertainty and doubt which could be lexically inferred by the test taker. So, the first item could not be answered unless one would read parts of the text to infer if there was some doubt about the cause of

sinking. Other examples of synonyms and lexical clues that connected the test item to the text are listed in Table 5.4.

In item 2, “*The Mary Rose was the only ship to sink in the battle of 19 July 1545.*”, the word “only” was not mentioned in the text and there was no explicit mention of The Mary Rose being the only ship that sank in the battle, so one needed to infer it from the text. The text just reported the battle when the Mary Rose sank and did not mention other ships at all. So, the correct answer was “*Not given*”. Intuitively, it was hard to imagine that the Mary Rose was the only ship that sank. One could simply rely on common sense and choose “False” as an answer. In fact, world knowledge and common sense suggested the correct answer was “False”. Therefore, the item was counter intuitive and confusing. In fact, it seems that test takers need to forget common knowledge, that suggests a battle logically involves many ships and probably the Mary Rose was not the only ship that sank. This item was problematic in the sense that the test takers needed to forget about their common sense and just look at what the text says or suggests. The item involved inferential comprehension from ideas discussed in two paragraphs in the text. So, item 2 involved inferential comprehension.

Unlike the first two items which asked for implicit inferential meaning, items 3 and 4 asked for some explicit literal meaning stated in the text. Test takers could simply locate the relevant information in the text and get the answer. The same ideas were mentioned in the text, but the language used in these items differed from the text and some synonyms were used. In fact, items 3 and 4 were basically paraphrases. The lexical clues in the test items could guide the test takers locate the relevant information in the text. Most often these clues were lexical where a synonym or negated antonym was included in the statement. For example, in item 3, *nearly all* and *most of* were synonyms. Also, in item 4, *housed* and *contained* were also synonyms. These features of the item indicated implicit assessment of vocabulary because test takers should either know these synonyms or guess them from the context by lexical inferencing. These lexical clues are also presented in Table 5.4.

As indicated in Table .4, the relevant information that contained the answer to items 1,3, and 4 in *True, False, Not given* items were at sentence level. For item 2, it was at paragraph level. So, except for item 2, the task basically measured sentence level comprehension and test takers could simply read few sentences to get to the answer. However, it goes without saying that

a test taker who could attend to the flow and interconnectedness of ideas in the whole text can be in a better position to locate the relevant information and answer these items.

Table 5. 4.
Summary of the True, False, Not given Task features

	Intended skill	Lexical clues	Relevant information	Type of comprehension
Item 1	Sentence level comprehension of specific details	<i>Doubt-vary</i>	sentence	Inferential
Item 2	Paragraph level comprehension	-	paragraph	Inferential
Item 3	Sentence level comprehension of specific details	<i>Nearly all of- most of Intact-undamaged The starboard half-one side of)</i>	sentence	Literal
Item 4	Sentence level comprehension of specific details	<i>- (housed-contained) - (treasure trove of beautifully preserved artefacts-valuable historical objects)</i>	sentence	Literal

5.2.1.2. Task two: The Matching Features Task

As shown in Table 5.5, the second task was the *Matching Features Task* which essentially asked test takers to match four events stated in four statements with a date from the table of dates. The focus of the statements was on how the Mary Rose was found and the attempts made by different people to raise it from the sea. The task was like asking when the following events happened.

Table 5.5.
The Matching Features Task

-
- 5 A search for the Mary Rose was launched.
6 One person's exploration of the Mary Rose site stopped.
7 It was agreed that the hull of the Mary Rose should be raised.
8 The site of the Mary Rose was found by chance.

List of Dates

A-1836 B- 1840. C- 1965 D-1967

E-1971	F-1979	G- 1982
--------	--------	---------

The task statements measured literal comprehension of specific details which also involved lexical inferencing of some of the words and phrases used in the items. So, answering them involved lexical inferencing. This task also included some lexical clues in the form of synonyms. In each statement 5-8, a synonym was used. Again, these clues might complicate the search for an answer if the test taker does not know their meaning. Such clues could be seen as an implicit way of measuring vocabulary knowledge.

All of the relevant information that contained the answers to these items could be found in few sentences in a paragraph, but not the whole paragraph. It appears that test takers can answer these items if they read and comprehend a paragraph or part of a paragraph. Again, test takers who can attend to the flow and interconnection between ideas in the whole text can be in a better position to locate the relevant information and answer the items. Table 5.6 presents the summary of the main features of the *Matching Features Task*.

Table 5. 6.
Summary of the main features of the Matching Features Task

	Intended skill	Lexical clues	Relevant information	Type of comprehension
Item 5	Comprehension of specific details at Inter-sentence level	<i>Initiated- launched</i>	Few sentences	Literal and lexical inferencing
Item 6	Comprehension of specific details at Inter-sentence level	<i>Faded into obscurity- stopped</i>	Few sentences	Literal and lexical inferencing
Item 7	Comprehension of specific details at Inter-sentence level	<i>Given the go ahead- agreed</i>	Few sentences	Literal and lexical inferencing
Item 8	Comprehension of specific details at Inter-sentence level	<i>Turned out to be- found by chance</i>	Few sentences	Literal and lexical inferencing

5.2.1.3. Task three: The *Diagram Completion Task*

The third task was the *Diagram Completion Task*. As presented in Figure 5.1, the *Diagram Completion Task* raised some questions about stage 1 and stage 2 of the “*Raising the Mary Rose*” which was explicitly stated in the title of the diagram. The diagram itself consisted of two parts. The first part addressed stage one and the second part addressed stage 2. Stage one was presented in a single graph while stage two is presented in two graphs. The task presents a two-piece diagram with five text boxes that contain a blank. The blanks should be filled in with

“no more than two words”. Answers to these blanks are all related to the last two paragraphs which describe three stages of “*Raising the Mary Rose*”. Stage one includes two items while stage two includes three items.

Studying and understanding the diagram does not seem to be straightforward and might depend on the topical knowledge of the test taker and the degree they are visual in their learning. Therefore, it is not really clear if the graphs can be used and understood by all the test takers. They can be interpreted and understood differently by different test takers, depending on their learning styles and their topical knowledge. There seems to be an element of subjectivity in processing and understanding the diagram which might add to the complexity and difficulty of the task. Another feature worth noting is that unlike the two previous tasks which are very explicit in terms of the linguistic medium used to ask the questions, the *Diagram Completion Task* makes use of a visual medium. The task mixes the linguistic and visual mediums of processing and comprehension and calls for two different mediums of comprehension (as well as back and forth movement between them). The main text described stages of “*Raising the Mary Rose*” linguistically while the diagram was essentially visual. However, one question that remained open was the extent to which actual test takers make use of the diagram in their processing. It is possible to process the test task by relying just on the text box and ignore the diagram altogether. The text boxes are more straightforward and provide enough context for understanding what was missing.

Raising the hull of the *Mary Rose*: Stages one and two

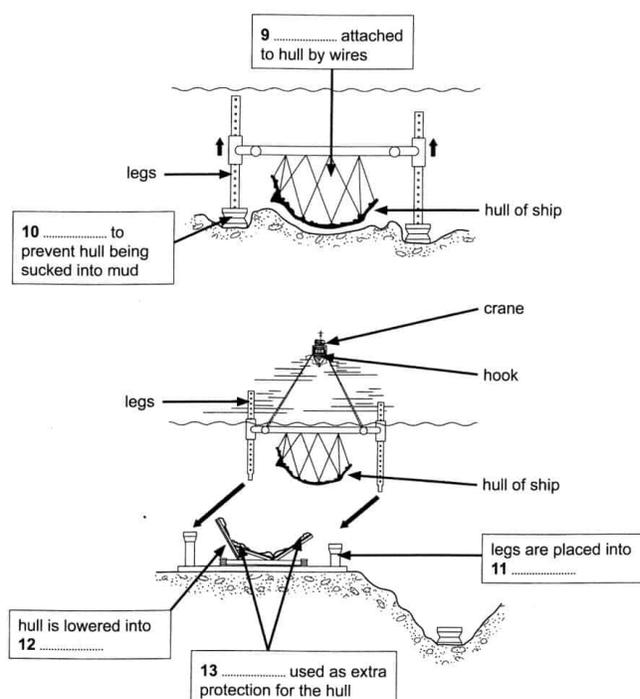


Figure 5.1. The *Diagram Completion Task*

Another feature of the *Diagram Completion Task* has to do with the missing lexical items which are by nature technical. They are not general words and readers may not exactly know what they are. Words and phrases such as “*hydraulic jacks, stabbing guide, lifting frame, lifting cradle, airbag*” are not typical, frequent words and phrases and involve certain amount of technical knowledge of engineering, or mechanics. Use of such technical words can influence processing and comprehending the task either in positive or negative ways, depending on the background knowledge of the test taker. Overall, the *Diagram Completion Task* assumes some visual literacy and background knowledge on the part of the test takers. So, visual learners might be at an advantage doing this task. It might be an asset for test takers who can better understand the processes and operations used in engineering projects and raising a ship and readers who are more visual.

One feature of the textbox used in the *Diagram Completion Task* was the use of some lexical clues such as synonyms to help test takers in their response to these items, provided that they knew these lexical items and attended to them. As shown in Table 4.7, in most of these items, paraphrasing and lexical clues such as synonyms were used. For example, in item 9 the phrase “*by-via*”, in item 10 “*preventing hull being sucked into the mud- overcome the hull being*

sucked into the mud” are synonyms. In item 11, “*legs are placed into- locate legs into*” and in item 13, “*used as extra protection for the hull- provide additional cushioning for the hull*” make use of synonyms. To get the answer, these synonyms can guide the test takers in the text and help in locating the specific relevant information. Some paraphrasing was also used to push the test takers read the relevant sentences more carefully. So, these items implicitly measured some vocabulary items in the context of paraphrases sentences. The only exception was item 12 where some lexical inferencing was needed. The word “*transferred into*” should be inferred as “*lowered into*”.

The relevant information that contained the answer to the *Diagram Completion Task* were all included in a short phrase that was part of a sentence.

Inclusion of the *Diagram Completion Task* in the academic module of the IELTS RCM can be justified on the ground that academic texts include different kinds of visual tools such as tables, diagrams, graphs, etc., but in academic textbooks, diagrams are basically used as learning aids rather than as tools for assessing readers’ vocabulary knowledge or comprehension. Even if they are used for assessment or practice purposes, they tend to be used as a context for further elaboration and explanation rather than filling them with one or more words from the text. It is hard to imagine that in academic context diagrams are used for vocabulary learning or assessment.

Table 5. 7.
Summary of the Diagram Completion Task features

	Intended skill	Lexical clues	Relevant information	Type of comprehension
Item 9	Vocabulary knowledge in (inter)-sentence context	<i>Via-by</i>	Sentence	Literal
Item 10	Vocabulary knowledge in (inter)-sentence context	<i>Overcome-to prevent</i>	Sentence	Literal
Item 11	Vocabulary knowledge in (inter)-sentence context	<i>Locate- placed into</i>	Sentence	Literal
Item 12	Vocabulary knowledge in (inter)-sentence context	<i>Transferred under water-lowered into</i>	Few sentences	Literal and lexical inferencing

Item 13	Vocabulary knowledge in (inter)-sentence context	<i>Additional cushioning-extra protection</i>	Sentence	Literal
----------------	--	---	----------	---------

5.2.2. Passage Two- “What destroyed Easter Island?”

The second text, i.e., “What destroyed Easter Island”, was a 7-paragraphs text which presented two opposing views about how *Easter Island* was destroyed along with some evidence to support each view. The text included description of some real events and some speculations and abstract thoughts and ideas about these events. One outstanding feature of the text was frequent use of numbers, dates, and several proper names of scholars and places which made it more concrete. Another outstanding feature of the text was use of many different names and words with the same reference which could be very confusing for someone who is not familiar with these names and he is reading these names for the first time. This included names of places, scholars, and different people. Table 5.8 presents examples of these phrases and words with the same reference Table 5.8.

Words and phrases with the same reference used in the passage two

Human statues, moai, statues
moai builders, Polynesians, the Rapa Nui, settlers
Rapa Nui, the island,
Rpa Nui is used both to refer to the island and the people of Rapa Nui

With regard to textual features, as shown in Table 5.9, the text was very different from an academic text in terms of length. It was 891 words long. Average sentence length was 18 words which clearly indicated the sentences were not short simple sentences, rather, they were syntactically complex or compound sentences. Lexical density of the text was .59 which is relatively dense. Based on the readability indices calculated, this text was at 12th grade level which means the text matched readability of texts read by 12th graders. In terms of genre or text type, the text was relatively argumentative and provided two opposing views on the destruction of the island. Description of different pieces of evidence for each point of view contributed both to the descriptive and argumentative nature of the text. However, the arguments and the evidence were not complex or abstract. They were real and concrete. Overall, the text looked like a short introduction or background to a serious academic text on the topic. The topic was a general history topic that one may find in a general magazine or history magazine published for the public. It did not look like an academic paper in any sense. The topic seemed to be more familiar to the western reader where archeology has a long history and most archeological

discoveries are reported and presented to the public in the form of reports, interviews, documentaries, movies, etc. So, it is more likely that L1 readers are familiar with the topic or know more about it.

Table 5.9.

Textual features of the “Easter Island” text

Text features	Easter Island
Text length	906
Sentence length	18
Lexical density	0.59
Readability	12 th grade
Genre	Magazine short article
Topical Knowledge	To some extent
Cultural knowledge	Moderate
Text abstractness	Mostly concrete ideas

The text was also analyzed for its lexical features. Table 5.10 presents the main results of the lexical profile of the Easter Island text.

Table 5.10.

Lexical profile of the “Easter Island” text (IELTS RCM)

	Families	Types	Tokens	Percent	Cumulative frequency
Easter Island					
K1 words	193	238	612	68.30%	68.30
K2 words	52	54	85	9.49%	77.79
AWL	29	31	41	4.58%	82.37
Sub list 1: assume create environment evidence identity indicate interprets issue require researcher					
Sub list 2: construct culture maintain resource					
Sub list 3: dominance					
Sub list 4: civil					
Sub list 5: energy stable sustainable					
Sub list 6: author display transport					
Sub list 7: isolate					
Sub list 8: thereby					

Sub list 9: erosion scenario vision					
Sub list 10: collapse convince					
Off-list words	-	101	158	17.63%	100.00

As shown in Table 5. 2, a great proportion of the words used in the text 68.30 % (N=193) were K1 words and only 9.49% (N=52) words were K2 words. Results also produced lists of word from each sub-list. Moreover, words from the AWL made 4.58% (N=29) of the words and off-list words covered 17.63% of the words. Again, the off-list words covered a sizeable proportion of the words in the text. In these indices N refers to the numbers of word families used in the text.

Having analyzed linguistic and textual features of the first text used in the IELTS RCM, now I turn to the analysis of the three tasks that followed the text; 1) the *Matching Headings Task*, 2) the *Summary Completion Task 1* and 3) the *Multiple Choice (two answers) Task*.

5.2.2.1. The Matching Headings Task

The *Matching Heading Task* asked for choosing a topic for each paragraph from a list of headings. There were 9 heading options for the 7 paragraphs (See Table 5.11). So, there were two extra options. One feature of the headings was that the language of the heading was abstract and distant from the paragraphs which left them open to inferencing by the test takers. Use of exact words and phrases from the paragraph was intentionally avoided in the headings. More specifically, in most of the headings the wordings were different from the wording of the passage. All the headings substituted synonyms for key words and phrases used in the paragraphs. Some of the lexical clues and synonym words/phrases are listed in Table 5.11.

One very important point about inferential items such as the items in the *Matching Headings Task* is that drawing inferences is different from choosing among inferences made by the test developer. Therefore, a distinction needs to be made between “drawing inference” and “choosing from inferences made by test developers”. Drawing an inference involves different processes than deciding which inference is right. In the *Matching Heading Task*, the test takers do not need to draw inferences, rather they just choose from a list of inferences made by the test developer. Therefore, inferential items in this test format do not necessarily measure inferencing process in its active productive form.

Another feature of the *Matching Heading Task* is related to “authenticity”. The way headings were used in the test task differed from the natural way headings are assigned to

paragraphs or texts in academic texts. In academic context headings are very close to the text and do not involve much abstraction. They are basically summaries that best represent the gist or the main theme of the paragraph or text. They introduce the main focus of the piece and reinforce the main message of the piece. There is no twist between what the passage says and what the headings suggest. Writers do their best to help the reader get as much information from the headings. The *Matching Headings Task*, on the other hand, seemed to be in contrast with usual practice of academic texts because they are not straightforward, and the test taker need.

Closer examination of the headings also showed that some headings covered almost all the paragraph while other headings summarize just few sentences from the paragraph and did not fully address the whole paragraph. For instance, heading i covered just the second half of paragraph D and left the first half out of its coverage. This inconsistency in defining and/or operationalizing the *Matching Headings Task* is one shortcoming of the task which can confuse the test takers and function a source of construct irrelevance in the test score.

Table 5.11.

List of headings for the Matching Headings Task

List of Headings:

- i Evidence of innovative environment management practices
 - ii An undisputed answer to a question about the moai
 - iii The future of the moai statues
 - iv A theory which supports a local belief
 - v The future of Easter Island
 - vi Two opposing views about the Rapanui people
 - vii Destruction outside the inhabitants' control
 - viii How the statues made a situation worse
 - ix Diminishing food resource
-

However, The *Matching Headings Task* seems relevant to academic reading in that academic texts are full of headings and sub-headings which signal transition and function as signposts that guide the reader throughout the text. This is why readers of academic texts tend to pay attention to the headings and sub-headings in the text. Based on personal experience. Additionally, I also argue that headings are relevant to academic reading where readers usually tend to summarize some sections or paragraphs of the text as they read and headings which helps them in getting the gist of the section to develop text/mental representation. However, the way the *Matching Headings Task* was operationalized in the test, makes it hard to see how it authentically represents the same practice in academic reading. When readers read for academic purpose, they are not asked to assign headings for texts or parts of a text. Nor do they take the

heading as a summary of the text. Summaries tend to be more elaborate while headings are always short phrases. Therefore, in spite of frequent use of headings in academic texts, the *Matching Headings Task* does not seem to authentically represent it as practiced in academic context.

Most of the *Matching Heading items* measured literal and inferential comprehension at paragraph level. attention to both the gist (the main idea) of each paragraph and details seem to be necessary for task performance. All test items included some synonyms and lexical clues for matching the headings to the paragraphs. For example, in item 15, *Decrease crop yield* and *diminishing food resources* are very close in meaning which can help in choosing the right heading. Table 5.12 presents summary of the intended skills, lexical clues, relevant information and type of comprehension involved in the *Matching Headings Task*.

Table 5.12.
Summary of the Matching Headings Task features

	Intended skill	Synonyms & Lexical clues	Relevant information	Type of comprehension
Item 14	Paragraph level comprehension of literal and inferential meaning	<i>Proved/in doubt- an undisputed answer</i>	Paragraph	Literal and inferential
Item 15	Paragraph level comprehension of literal and inferential meaning	<i>Decrease crop yield- diminishing food resources</i>	Paragraph	Literal and inferential
Item 16	Paragraph level comprehension of literal and inferential meaning	<i>Accelerated self-destruction- made a situation worse</i>	Paragraph	Literal and inferential
Item 17	Paragraph level comprehension of literal and inferential meaning	<i>Sustainable farming- innovative environment management</i>	Paragraph	Literal and inferential

Item 18	Paragraph level comprehension of literal and inferential meaning	<i>Backs up folklore-support a local belief</i>	Paragraph	Literal and inferential
Item 19	Paragraph level comprehension of literal and inferential meaning	<i>Not wholly responsible-outside the inhabitants' control</i>	Paragraph	Literal and inferential
Item 20	Paragraph level comprehension of literal and inferential meaning	<i>Vision-view</i>	Paragraph	Literal and inferential

5.2.2.2. The Summary Completion Task 1

As shown in Table 5.13, the second task was the *Summary Completion Task 1* which consisted of a short text with 4 blanks. It had to be filled in with one single word from the text. The summary was a 119 words short paragraph consisting of 4 randomly selected sentences from the text, each with a blank. They were paraphrased sentences from two consecutive paragraphs (B and C). The paraphrased sentences incorporated a number of exact words and phrases from the text. Contrary to what the task claims, the summary text was not really a summary in its general sense of the word where the main points of a text or paragraph are briefly re-written in few sentences. Summaries usually cover the main and most important ideas of the text rather than rephrasing or paraphrasing a certain part of the text selectively.

Similar to other test tasks, the relevant information and the *Summary Completion Task* included a number of synonyms and paraphrases. The relevant information and the test tasks include a number of synonyms.

Table 5.13
The *Summary Completion Task 1*

Jared Diamond's View	
Diamond believes that the Polynesian settlers on Rapa Nui destroyed its forests, cutting down its trees for fuel and clearing land for 21 -----.	Twentieth-century discoveries of pollen prove that Rapa Nui had once been covered in palm forests, which had turned into grassland by the time the Europeans arrived on the island. When the islanders were no longer able to build the 22 -----they needed to go fishing, they began using the island's 23 ----- as a food source, according to Diamond. Diamond also claims that the moai were built to show the power of the island's chieftains, and that the methods of transporting the statues needed not only a great number of people, but also a great deal of 24 -----.

As can be seen in Table 5.14, examples of such paraphrasing included use of synonyms such as (*clear- cut down*), (*ate- use as food source*), (*required-needed*) (*hailed-transferred*), and

(*a lot of-a great deal of*). The relevant information that contained the answers to the summary task were words embedded in few clauses and sentences in the text and test takers could search for them in few sentences of two paragraphs. Locating the relevant information in the text, the reader can simply read one or two sentences and fill in the blanks with appropriate words

Table 5.14.

Relevant information in the text

	<i>The Summary Completion item</i>	Text
21	cutting down its trees for fuel and clearing land forcleared the forest for firewood and <u>farming</u>
22	build ... for fishing	construct wooden <u>canoes for fishing</u>
23	began using the island's as a food source	they ate <u>birds</u>
24	.. needed not only a great number of people, but also a great deal of	..but they required both a lot of <u>wood</u> and a lot of people
	...	

The task seemed to measure vocabulary knowledge as sentence level. The type of comprehension involved was all literal comprehension. Table 5.15 presents a summary of the *Summary Completion Task* features.

Table 5.15.

Summary of the Summary Completion Task features

	Intended skill	Lexical clues	Relevant information	Type of comprehension
Item 21	Vocabulary at sentence level	<i>Cleared-cut down</i>	Sentence	Literal
Item 22	Vocabulary at sentence level	<i>Build-construct</i>	Sentence	Literal
Item 23	Vocabulary at sentence level	<i>Ate birds-birds as food source</i>	Sentence	Literal
Item 24	Vocabulary at sentence level	<i>Required-needed A lot of- a great deal of</i>	Sentence	Literal

5.2.2.3. The Multiple Choice (Two answers) Task

The last task for Easter Island text was the *Multiple Choice (two answers) Task*. As shown in Table 5.7, the task asks for two points of disagreement between the two researchers

mentioned in the passage, Hunt and Lipo. The options cover some of the ideas that are discussed in different paragraphs. All the options are inferences drawn by the test developers. So, they are abstract statements. Test takers need to decide if these inferences make sense when examined against the information given in the paragraphs. The text, on the other hand, presents several points of disagreement throughout the passage either explicitly or implicitly. All paragraphs except for the first and last paragraphs describe one point of disagreement or another. Therefore, test takers need to read through these paragraphs and find two points of disagreement that best match the options.

The task basically required reading almost all the text and sieving through it to find out the correct answers. This task differed from other tasks in that it is asking for two main ideas that summarize the information presented in the whole text. Unlike most of the 40 items on the test that asked for some specific details this item was one of the few items that addressed comprehension of larger chunks of the passage. The item could not be answered without reading 4-5 consecutive paragraphs. It requires careful reading or skimming through the text to identify the two points of disagreement. Specific details are placed in one or few sentences embedded in a paragraph but the two main points of disagreement that asked for in this test task were discussed throughout the passage in five different paragraphs. These points were not explicitly mentioned in the text and needed to be inferred from the text.

Table 5.16.

The Multiple Choice (two answers) Task

On what points do Hunt and Lipo disagree with Diamond?

- A the period when the moai were created
 - B how the moai were transported
 - C the impact of the moai on Rapanui society
 - D how the moai were carved
 - E the origins of the people who made the moai
-

Text level comprehension of the main idea is the main purpose of the *Multiple Choice (two answers) Task*. Therefore, reading and comprehending the whole text is required for correct response. Unlike other test tasks, the *Multiple-Choice (two answers) Task* did not involve much lexical clues and they do not seem to be as important as in other test tasks. Table 5.17 presents a summary of the main features of the *Multiple Choice (two answers) Task*.

Table 5.17.

Summary of the Multiple Choice (two answers) Task features

	Intended skill	Lexical clues	Relevant information	Type of comprehension
Item 25	Text level comprehension	<i>Moving the moai-transporting the moai</i>	Several paragraphs	Literal and inferential
Item 26	Text level comprehension	-	Several paragraphs	Literal and inferential

5.2.3. Passage Three- “Neuraesthetics”

The third passage- “Neuraesthetics”- was a college level reading passage consisting of 9 paragraphs. It discussed how people perceive art and factors that might affect their perception of art. The title of the text could be scary for test takers who did not know the word “*Neuroaesthetics*”. It seemed strange and required some morphological knowledge of English to break it down to its component parts, “neuro” and “aesthetics”. Still one needs to know these two words to guess what the whole word may meant. The text began with introducing how art was perceived and reported on four research studies conducted by different scholars on this topic and ended with some abstract ideas such as the complexity of art appreciation and its subjective nature.

A point worth mentioning at this juncture is that in reading comprehension, titles are expected to be easy gateways for readers by activating their background knowledge to facilitate comprehension and they are usually easier than the text while for this text the title works differently. If the reader does not know the words, he should read part of the text to guess what the title really means.

With regard to textual features, as shown in Table 5. 18, the text was much shorter than a typical academic text which is at least few thousands of words long. The text was 985 words long. Average sentence length was 23 words which clearly indicated the sentences were syntactically complex and/or compound sentences. Lexical density of the text was .55 which is relatively dense. Based on the readability indices calculated, this text was at 12th grade level which means the text matched readability of texts read by 12th graders. In terms of genre or text type, the text sounded more academic because the topic was more abstract substantiated with some experiments that were more tangible and real. With such content, the text seemed to succeed striking a balance between the abstract and the concrete. Text content elaborated on an abstract topic by providing some scientific concrete evidence from research studies. The evidence was relatively technical and needed some background. Someone with no interest or

background in the arts could not easily read and process the text. Compared with the other two texts, i.e., “Raising the Mary Rose” and “Easter Island”, “Neuraesthetics” text seemed more academic. The text is not neutral to background knowledge Readers with background in humanities and social science most probably can better read this text.

Again, readers coming from a cultural context with more art education, art criticism, and art appreciation. Also, western readers seem to come from a cultural background that better fits reading and processing this text. The topic seems to be more familiar to the western reader where painting, architecture, archeology, sculpture of great artists and painters make a huge contribution to culture and has a long history reporting and presenting art to the public in the form of reports, interviews, documentaries, movies, etc. So, it is more likely that L1 English speaking readers are familiar with the topic or know more about it. Ccomprehension of the text seems to demand some cultural and background knowledge.

With regard to textual features, as shown in Table 5.18, the text was differed from an academic text in terms of length. It was 985 words long. Average sentence length was 18 words which clearly indicated the sentences were not short simple sentences, rather, they were syntactically complex or compound sentences. Lexical density of the text was .55 which is relatively dense. Based on the readability indices calculated, this text was at 12th grade level which means the text matched readability of texts read by 12th graders. In terms of genre or text type, the text was a descriptive-argumentative which provided different pieces of evidence related to art appreciation. The evidence provided in the text were empirical findings which gives the text some academic tone. The text looked like a short introduction. The topic was a not technical, yet it required some background for better comprehension. Overall, the text looked more academic than the other two texts. It did not look like an academic paper in any sense. The topic seemed to be more familiar to the western reader where archeology has a long history and most archeological discoveries are reported and presented to the public in the form of reports, interviews, documentaries, movies, etc. So, it is more likely that L1 readers are familiar with the topic or know more about it.

Table 5.18.

Textual features of the “Easter Island” text

Text features	Neuraesthetics
----------------------	-----------------------

Text length	985
Sentence length	23
Lexical density	0.55
Readability	College level
Genre	Short academic introduction
Topical Knowledge	High
Cultural knowledge	Moderate
Text abstractness	Mostly concrete ideas

For further analysis of the “*Neuroaesthetics*” text, the lexical profile of the text was assessed. Results are reported in Table 5.19

Table 5.19.
Lexical profile of the “Neuroaesthetics” text (IELTS RCM)

	Families	Types	Tokens	Percent	Cumulative frequency
Neuroaesthetics					
K1 words	214	276	759	76.82%	76.82
K2 words	32	33	47	4.76%	81.58
AWL	53	59	81	8.20%	89.78
Sub list 1: analyse approach area concept consist create environment interpretation interpret occur process respond role similar underestimate					
Sub list 2: complexity impact previous reconstruct seek					
Sub list 3: constant demonstrate exclusive instance link					
Sub list 4: debate emerge hypothesis label					
Sub list 5: alter challenge energetic evolve generation image mentality objective precise style trend versions					
Sub list 6: abstract					
Sub list 7: adapt confirm definite dynamic volunteer					
Sub list 8: ambiguous appreciate crucial radical visual					
Sub-list 9: vision					
Sub list 10: inclination					
Off-list words	-	82	101	10.22%	100.00

Lexical profile of the text, as shown in Table 5.19, also showed that a great proportion of the words used in the text 66.82 % (N= 214) were K1 words and only 4.76% (N=32) words were K2 words. Results also produced lists of word from each sub-list. Moreover, words from the AWL made 8.20% (N=53) of the words and off-list words covered 10.22% of the words. Again, the off-list words covered a sizeable proportion of the words in the text. In these indices N refers to the numbers of word families used in the text.

Having analyzed linguistic and textual features of the text, now I turn to the analysis of the three tasks that followed the text; 1) the *Multiple-Choice Task*, 2) the *Summary Completion Task 2*, and 3) the *Yes, No, Not given Task*. They are analyzed in the following section.

5.2.3.1. The Multiple-Choice Task (one answer)

The Multiple-Choice Task consisted of five items. Except for the last item which asks for an appropriate subtitle for the whole text, items 1-4 ask for some specific information. As shown in Table 5.20, item 1 asks why shape-matching test is mentioned in the text. Items 2, 3, and 4 ask about results of the studies reported at the end of the relevant paragraphs. Item 1 and 4 make direct reference to the paragraph where the answer lies. This helps test takers locate the relevant section and info in the text. Unlike items 1-4 that require comprehension of specific information, item 30 asks for a sub-title of the whole article and requires global comprehension of the whole text.

Table 5.20.

The Multiple Choice Task

26. In the second paragraph, the writer refers to a shape-matching test in order to illustrate

- A the subjective nature of art appreciation.
- B the reliance of modern art on abstract forms.
- C our tendency to be influenced by the opinions of others.
- D a common problem encountered when processing visual data.

27. Angelina Hawley-Dolan's findings indicate that people

- A mostly favour works of art which they know well.
- B hold fixed ideas about what makes a good work of art.
- C are often misled by their initial expectations of a work of art.
- D have the ability to perceive the intention behind works of art.

28. Results of studies involving Robert Pepperell's pieces suggest that people

- A can appreciate a painting without fully understanding it.
- B find it satisfying to work out what a painting represents.
- C vary widely in the time they spend looking at paintings.
- D generally prefer representational art to abstract art.

29. What do the experiments described in the fifth paragraph suggest about the paintings of Mondrian?

- A They are more carefully put together than they appear.
 - B They can be interpreted in a number of different ways.
 - C They challenge our assumptions about shape and colour.
 - D They are easier to appreciate than many other abstract works.
-

30. What would be the most appropriate subtitle for the article?

- A Some scientific insights into how the brain responds to abstract art
 - B Recent studies focusing on the neural activity of abstract artists
 - C A comparison of the neurological bases of abstract and representational art
 - D How brain research has altered public opinion about abstract art
-

In terms of type of comprehension, item 1 involved literal comprehension of two paraphrase sentences.

Item 2 used the word “indicate” which tends toward literal dimension of comprehension while items 3 and 4 included the word “suggest” which suggests inferencing side of comprehension. Item 5 was asking for global comprehension of the text which is clearly asking for inferential and evaluating meaning. As presented in Table 5.21 all the multiple-choice items make use of some lexical clues such as synonyms in their stems. They included “*If they see others doing the same- our tendency to be influenced by others*”, “*perceive-sense*”, “*intention-vision*”, “*satisfying-rewarding*” “*workout-decipher*”, “*carefully put together- meticulously composed*”, “*paintings-works of art*”, and “*insights-shed light on*”.

For the *Multiple-Choice Task*, the relevant information in the text was also identified. Table 5 13 presents the relevant information for answering the Multiple-Choice items.

Table 5.21.

The relevant information that provides the answer to Multiple Choice items

Items	Relevant information
1	We certainly have an inclination to follow crowd. When asked to make simple perceptual decisions such as matching a shape to its rotated image, for example, People often choose a wrong answer if they see others doing the same thing.
2	It seems that the viewer can sense the artist’s vision in paintings, even if they can’t explain why.
3	It seems that the harder it is to decipher the meaning, the more rewarding is the moment of recognition.
4	Eye-tracking studies confirm that they are meticulously composed, and that simply rotating a piece radically changes the way we view it.

As shown in Table 4. 13, for items 1-4, the target information that helped test takers answer the question could be found in one or two sentences. Readers did not need to read more than few sentences to get the answer. Test takers might need to search larger units to locate those few sentences at the end of each paragraph, but their search did not need to go beyond paragraph

level, except for item 5 which addressed the whole text. So, if the answers lied in few sentences, comprehension cannot go beyond paragraph level. Just like all other test tasks in the IELTS RCM, almost all the MC items, used paraphrased sentences and synonym. Either stems of each item or the options included words and phrases that are synonyms with words and phrases from the text. These lexical clues could directly help test takers in choosing the right answer. Table 5.22 presents the summary of the *Multiple-Choice Task* features.

Table 5.22.

Summary of the Multiple-Choice Task features

	Intended skill	Synonyms & Lexical clues	Relevant information	Type of comprehension
Item 27	Inter-sentence/Paragraph comprehension	<i>if they see others doing the same- our tendency to be influenced by others</i>	Several sentences	Literal
Item 28	Inter-sentence/ Paragraph comprehension	<i>(Perceive-sense) (intention-vision)</i>	Several sentences	Inferential
Item 29	Inter-sentence/Paragraph comprehension	<i>(Satisfying- rewarding) (Work out-decipher)</i>	Several sentence	Inferential
Item 30	Inter-sentence/ Paragraph comprehension	<i>(carefully put together- meticulously composed) (Paintings-works of art)</i>	Several sentences	Inferential
Item 31	Text level comprehension	No specific lexical clues	The whole text	Literal and inferential

5.2.3.2. The Summary Completion Task 2

The *Summary Completion Task 2* was a short passage of 74 words, three sentences each with a blank that needed to be filled in with an appropriate word from the 8 options given in the table of words. All the options were nouns. The first sentence of the summary was a paraphrase while the other two sentences are exact copies of the same sentences in the text. Such copied sentences and phrases were definitely much easier to answer. Just like the *Summary Completion Task 1*, the *Summary Completion Task 2* was not a summary of the text or a particular paragraph. It was basically a mix of paraphrase and copy of three separate sentences located in two different paragraphs in the text, paragraph 1 and 7. Table 5.23 presents the *Summary Completion Task 2*, the list of options given.

Table 5.23.

*The Summary Completion Task 2***Art and the Brain**

The discipline of neuroaesthetics aims to bring scientific objectivity to the study of art. Neurological studies of the brain, for example, demonstrate the impact which Impressionist paintings have on **31**----- Alex Forsythe of the University of Liverpool believes many artists give their works the precise degree of **32**----- which most appeals to the viewer's brain. She also observes that pleasing works of art often contain certain repeated **33**----- which occur frequently in the natural world.

Options

A- interpretation	B- complexity	C- emotions	D. movements
E- Skill	F- layout	G. Concern	H. images

The relevant information for the *Summary Completion Task* items and the correct answer (highlighted word) is presented in Table 4.25.

Table 5. 24.

Relevant information for the Summary Completion Task 2

Relevant information in the text

- 1) Neurological studies of the brain, for example, demonstrate the impact which Impressionist painting have on our **emotions**. (Paragraph 1)
- 2) Alex Forsythe of the University of Liverpool believes many artists give their works the precise degree of **complexity** which most appeals to the viewer's brain. (Paragraph 7)
- 3) She also observes that pleasing works of art often contain certain repeated **images** which occur frequently in the natural world. (Paragraph 7)

The items tap vocabulary knowledge in the context of literal comprehension at sentence level. paraphrasing skill is also a skill that can help answering the test items. of test task. Just like other test tasks, most items involve synonyms and lexical clues that connect the item to the text. Table 5.25 presents the Summary of the features for the *Summary Completion Task 2*.

Table 5.25.

Summary of the features for the Summary Completion Task 2

	Intended skill	Lexical clues	Relevant information	Type of comprehension
Item 32	Vocabulary at sentence level context	Impact of emotions-impact of feelings	Sentence	Literal
Item 33	Vocabulary at sentence level context	Precise degree of complexity-precise degree of ...	Sentence	Literal
Item 34	Vocabulary at sentence level context	Contain certain repeated images-contain certain repeated...	Sentence	Literal

5.2.3.3. The Yes, No, Not given Task

As shown in Table 5.26, the *Yes, No, Not given Task* consisted of 6 statements that had to be checked against the ideas discussed in the last three paragraphs of the text. Some of the statements were literal paraphrases and some were inferred statements, Unlike the *True, False, Not given Task*, which was more factual and could be more directly understood from the text, the *Yes, No, Not given* items were more implicit and abstract. The relevant information, as shown in Table 5.26, included a sentence embedded in a single paragraph.

Table 5.26.

The Yes, No, Not give Task and the relevant information in the text

Task items	Relevant info in the text
35. (NO) Forsythe’s findings contradicted previous beliefs on the function of ‘fractals’ in art.	What is more, appealing pieces both abstract and representational, show signs of “fractals” repeated motifs recurring in different scales.
36. (YES) Certain ideas regarding the link between ‘mirror neurons’ and art appreciation require further verification.	This may be down to our brain’s “mirror neurons”, which are known to mimic others’ actions. The hypothesis will need to be thoroughly tested.
37 (NG) People’s taste in paintings depends entirely on the current artistic trends of the period.	While the fashions of the time might shape what is currently popular, works that are best adopted to our visual system may be most likely to linger once the trends of previous generations have been forgotten.
38 (YES) Scientists should seek to define the precise rules which govern people’s reactions to works of art.	It would, however, be foolish to reduce art appreciation to a set of scientific laws.

39. (NG) Art appreciation should always involve taking into consideration the cultural context in which an artist worked.	We shouldn't underestimate the importance of the style of a particular artist, their place in history and the artistic environment of their time.
40. (NG) It is easier to find meaning in the field of science than in that of art.	In some ways, it's not so different to science, where we are constantly looking for systems and decoding meaning so that we can view and appreciate the world in a new way.

Similar to most of the IELTS RCM test tasks, most statements included some clues and key words that could be easily identified in the text. For example, “*fractals*”, “*mirror neurons*”, “*people’s taste in painting*”, “*precise rules*”, and “*art appreciation*” in items 1-6 are directly mentioned in the text and test takers can easily scan and locate them in the text. Some of these statements were not explicitly stated in the text. These statements were either paraphrases of some inferences drawn from the text and test takers needed to examine them against the ideas discussed in the text. As shown in Table 5.22, in each statement some lexical clues were used that match the text. For example, (*further verification- need to be thoroughly tested*) in item 36, (*trends of the time-trends of previous generations*) for item 37, (*precise rules- scientific laws and reaction to works of art- art appreciation*) in item 38, (*find meaning- encode meaning and cultural context- artistic environment of the time*) in item 39 were near synonyms that can help test takers locate the relevant information, though they were not enough to provide the answer and some lexical inferencing need to be made to get to the answer. Except for item 36 which was explicitly stated in the text, for other items it is hard to imagine one could examine the inferences made in the statement without careful reading of the text. Moreover, in item 37, the word “entirely” was very indicative what the answer could be especially test-wise test takers. As it is known, such absolute terms tend to make the ideas false. Moreover, one feature of the *Yes, No, Not given Task* was that it was the last task and test takers who were aware of the order of the test tasks or text structure could, probably, recognize that the last three paragraphs of the text were not addressed by the previous test tasks (i.e., the *Multiple Choice Task* and the *Summary Completion Task*). This awareness could help them narrow down their search for the relevant information and focus on the last three paragraphs for finding the answers. Table 5. 27 presents a summary of the main features of the *Yes, No, Not given Task* including the intended skill, the synonyms and lexical clues, the relevant information and type of comprehension.

Table 5.27.

Summary of the Yes, No, no given Task features

	Intended skill	Synonyms & Lexical clues	Relevant information	Type of comprehension
Item 35	inferencing at inter-sentential/paragraph level	-	Sentence	Inferential & evaluative
Item 36	Literal comprehension at sentence level	<i>need to be thoroughly tested- further verification</i>	Sentence	Literal
Item 37	Inferencing at sentence level	<i>trends of the time- trends of previous generations</i>	Sentence	Inferential
Item 38	Inferencing at inter-sentence level	<i>precise rules reaction to works of art- scientific laws art appreciation</i>	Sentence	Inferential
Item 39	Literal meaning at inter-sentence level	<i>cultural context- artistic environment of the time</i>	Sentence	Literal
Item 40	Inferencing at inter-sentence level	<i>find meaning- decode meaning (should involve- should not underestimate)</i>	Sentence	Inferential

5.3. Comparison of text features of the IELTS RCM texts and academic texts

Another step taken in the analysis of the IELTS RCM content was comparison of some linguistic and textual features of the IELTS RCM texts with two samples of academic texts; 1) an academic book chapter and 2) an empirical research article chapter. These two sample academic materials were provided by the undergraduate students who were using them as part of their course material. The comparison could provide some evidence if the IELTS RCM texts represent actual academic texts. Table 5.28 presents result of these linguistic and textual features of the IELTS RCM texts and the sample of academic texts.

Table 5.28.

Lexical profile of academic text "book chapter"

	Families	Types	Tokens	Percent	Cumulative frequency
Book chapter (academic source)					
K1 words	435	673	5293	69.99%	69.99
K2 words	97	120	188	2.49%	72.48
AWL	157	264	966	12.77%	85.25
Off-list words	-	427	1115	14.74%	100.00

As results in Table 5.28 show, the words used in the book chapter sample included 70% (N=435) K1 words, 2.5% (N=97) K2 words, 13% (N=157) AWL words, and 15% of list words.

In these indices N refers to the numbers of word families used in the text. In these indices N refers to the numbers of word families used in the text.

The same procedure was used for the empirical research article sample. Results are reported in Table 5.29.

Table 5.29
Lexical profile of the sample academic text “Linguistics Article”

	Families	Types	Tokens	Percent	Cumulative frequency
Research article (academic source)					
K1 words	474	757	5280	72.97%	72.97
K2 words	167	215	387	5.35%	78.32
AWL	174	251	481	6.65%	84.97
Off-list words	-	630	1088	15.04%	100.00

As results in Table 5.29 show, the words used in the book chapter sample included 73% (N=757) K1 words, 5% (N=215) K2 words, 6% (N=251) AWL words, and 15% (N=630) of list words. In these indices N refers to the numbers of word families used in the text.

At this juncture, as indices of textual features were separately presented for each text, results of the textual features for the IELTS RCM texts and the two academic samples are presented altogether for ease of comparison. (See Table 5.30)

Table 5.30.
Summary of text features for the IELTS RCM text and the two sample academic texts

Text features	The Mary Rose	Easter Island	Neuraesthetics	Book chapter	Research article
Text length	891	906	985	7632	7377
Sentence length	23	18	23	21	24
Lexical density	0.56	0.59	0.55	0.57	0.57
Readability	12 th grade	12 th grade	College level	University level	University level
Genre	Magazine short article	Magazine short article	Short academic introduction	Academic book chapter	Refereed research article

Background Knowledge	To some extent	To some extent	To some extent	Largely required	Largely required
Cultural knowledge	Moderate	Moderate	Moderate	None	none
Text abstractness	Very concrete	Mostly concrete ideas	Mostly concrete ideas	Mostly abstract	Mostly abstract

As shown in Table 5.30, the IELTS RCM texts were much shorter than academic texts- less than 1000 words in length, while the sample academic texts were more than 7000 words. In other words, the test texts were approximately 1/7 of the length of academic texts. In terms of the average sentence length, except for one text of the IELTS RCM with an average sentence length of 18 words, there was not a difference between the sentence length of the sample IELTS RCM texts and academic texts, which ranged between 21-24 words per sentence. In terms of lexical density, the indices for the IELTS RCM ranged between .55-.59. Academic texts fell within the same density range, suggesting that there was no difference between the IELTS RCM texts and the sample academic texts in terms of lexical density. Texts were also compared for their readability. Based on the readability index of The IELTS RCM two texts were 12th grade, and one text was at college level. Readability of the academic texts, however, was at university level with a SMOG index of 14 which suggest they were higher in readability.

As to the rhetorical organization, the academic texts were totally different from the IELTS RCM texts. While the IELTS RCM texts were short articles that one might find in a general magazine or a very short introduction to a topic. Undergrad academic texts were considerably longer, reflecting disciplinary subject or content areas (Artemeva & Fox, 2010). They included technical jargon and definition of technical terms wrapped in frequent citations that relate topic of the discussion to the body of research within the discipline. They were more complicated in terms of their structure and organization, consisting of several sections and sub-section (i.e., formatted based on the standards of academic writing with many headings and sub-headings and referencing). The IELTS RCM texts, on the other hand, were simple in format, few paragraphs with one following the other with no specific formatting features. Considering background knowledge, for all of the three RCM texts, a certain amount of background knowledge can support processing of the text, while for the academic text, background knowledge plays a significantly more vital role because the topic and ideas that are discussed are more technical and needed some background. In terms of cultural knowledge, it might be safe to

say that the IELTS RCM text are written with the assumption that they are part of public knowledge, an assumption which might not be necessarily true for all readers. The text involve some cultural knowledge in the sense that if they fall within the public knowledge of the test takers, they can be processed more efficiently while for the academic texts cultural knowledge does not play much of a role. For the academic texts technical knowledge of the topic is more influential. Finally, for text abstractness, all the IELTS RCM texts covered concrete ideas, with a thin layer of abstract discussion, while for the academic texts, the main points are basically abstract discussions and arguments with some concrete examples and empirical evidence.

5.4. Comparison of the lexical profile of the IELTS RCM texts and academic texts

Finally, to assess the extent to which the IELTS RCM texts represent lexical features of actual of academic texts, lexical profile of the LIELS RCM texts were calculated and compared with the lexical profile of two academic sources that the undergraduate students provided the researcher; 1) an academic book chapter and 2) an academic research article. The academic sources were actual material the L1 test takers were reading for their course work. Table 5.31 presents results of the lexical profile for the IELTS RCM and the two academic sources.

Table 5.31.
Lexical profile of the main sample of IELTS RCM and the two academic sources

The Mary Rose	Families	Types	Tokens	Percent	Cumulative frequency
K1 words	233	271	729	72.75	72.75
K2 words	54	61	75	7.49%	80.24
AWL	48	53	61	6.09%	86.33
Off-list words	-	85	137	13.67%	100.00
Easter Island					
K1 words	193	238	612	68.30%	68.30
K2 words	52	54	85	9.49%	77.79
AWL	29	31	41	4.58%	82.37
Off-list words	-	101	158	17.63%	100.00
Neuroaesthetics					

K1 words	214	276	759	76.82%	76.82
K2 words	32	33	47	4.76%	81.58
AWL	53	59	81	8.20%	89.78
Off-list words	-	82	101	10.22%	100.00
Linguistics book chapter (academic source)					
K1 words	435	673	5293	69.99%	69.99
K2 words	97	120	188	2.49%	72.48
AWL	157	264	966	12.77%	85.25
Off-list words	-	427	1115	14.74%	100.00
Linguistics article (academic source)					
K1 words	474	757	5280	72.97%	72.97
K2 words	167	215	387	5.35%	78.32
AWL	174	251	481	6.65%	84.97
Off-list words	-	630	1088	15.04%	100.00

As shown in Table 5.31, the IELTS RCM text and the academic texts were similar in terms of the percentage of the first and the second 1000 words used. The coverage for K1 words (the first 1000 words) for both the IELTS RCM and the sample academic texts ranged between (68%-76%). The coverage range of the academic texts fell within the range of the IELTS texts. Coverage of K2 words (the second 1000 words) for the IELTS RCM ranged between 77%-81%. Again, the coverage range of the academic text fell within the range of the IELTS texts. For the academic texts this coverage ranged between 72%-78%. Interestingly the K1 and K2 words had higher coverage in the IELTS RCM which can be indicative of the purposeful manipulation of the original text and tailoring the text to fit the operationalization of test construct. This claim can be further supported by the fact that almost all test tasks of the IELTS RCM included lexical clues such as synonyms. Use of such lexical clues can be much facilitated by inclusion of more words from the K2 words in the text.

A lexical feature that differentiated the IELTS RCM texts from the academic texts was the variability of coverage of academic words which ranged between 4%-8% of the words for the IELTS RCM texts while this index was about 12.77% and 6.66% for the linguistic book chapter and the linguistic article, respectively. Based on the two sample texts, it can be argued that academic words in the IELTS RCM are underrepresented. A more significant difference lies in the type of academic words used in these sample texts. For the IELTS RCM texts, academic word types ranged between 29-59 types while of the two academic texts this index was 157-174, suggesting huge difference in the number of families of AWL used in the texts. This clearly indicate that academic words are highly underrepresented in the IELTS RCM. Use of more types

of academic words contributes a lot to the academic tone and attitude that is dominant in academic texts

5.5. Summary of Answer Level, Skills Measured, and Type of Comprehension

As results of text features and task features were presented separately for each text and test item, at this juncture, for a more global view of the results, they are put together and summary of the findings are presented.

Content analysis of the IELTS RCM helped identify the answer level for each test item which refers to the unit that contained the relevant information for answering an item. As presented before, for some items such as the *True, False, Not given Task* and the *Diagram Completion Task*, the relevant information was a single sentence while for some items such as the *Matching Features Task* and the *Multiple Choice Task* it was a couple of sentences, yet for tasks such as matching heading it was the larger unit of a whole paragraph that contained the relevant information. Only for one test task (the *Multiple Choice (two answers) Task*) the relevant information was contained in several paragraphs. The last item of the *Multiple-Choice Task* also asked for global meaning which was discussed all through the text. Table 5.32 presents results of content analysis in terms of answer level, skill measured, and type of comprehension for each of the 9 test tasks.

Table 5.32.
Summary of answer level, skills measured, and type of comprehension for IELTS RCM test tasks

Test tasks	Answer level	No. of items	Skills measured	Type of comprehension
<i>True, False Not given Task</i>	One sentence	4	- Sentence and paragraph comprehension, understanding specific details	Literal and inferential
<i>Matching Features Task</i>	Multiple sentences	4	-Sentence and paragraph comprehension, understanding specific details	Literal

<i>Diagram Completion Task</i>	One sentence	5	-Sentence comprehension and vocabulary knowledge, understanding specific details - Sentence level paraphrasing	Literal
<i>Matching Headings Task</i>	The whole paragraph	7	-Paragraph comprehension, understanding the main idea and specific details	Literal and inferential
<i>Summary Completion Task 1</i>	Multiple sentences	4	Sentence comprehension- understanding specific details Vocabulary knowledge	Literal
<i>Multiple Choice Task (two answers)</i>	Several paragraphs	3	Text comprehension- understanding main ideas	Literal and inferential
<i>Multiple Choice Task</i>	Multiple sentences	4	Sentence+ comprehension- understanding specific details	Literal and inferential
<i>Summary Completion Task 2</i>	Multiple sentences	3	Sentence+ comprehension- understanding specific details	Literal
<i>Yes, No, Not given Task</i>	Multiple sentences	6	Sentence+ and paragraph comprehension, understanding specific details and main ideas	Literal, inferential, and evaluative

As shown in Table 5.28, it seemed that for 7 tasks out of 9 tasks the relevant information was contained in one or a few sentences. These tasks and items tapped into the sentence level comprehension and did not address macro discourse processes such as cohesion and coherence which require developing text representation. Most test tasks of the IELTS RCM can be processed and answered at lower level of sentence and paragraph comprehension because the tasks demand that level of processing. In answering some of the test tasks such as the *True, False, Not given Task*, the *Diagram Completion Task*, and the *Summary Completion Tasks 1 and 2*, can be answered the by simply attending to sentence level lexico-grammatical features of few sentences. It seems that the lexico-grammatical features at the level of sentence(s) is the main asset to use for answering the task items. As shown in the content analysis, test tasks are loaded with lexical clues and vocabulary items that are synonymous with words used in the text, therefore, test takers can attend to them and get the answer they are looking for. The *Matching Headings Task* (7 items) require for paragraph level comprehension and test takers should read all the paragraphs one by one to get to the answer. Two items in the *Multiple Choice (two answers) Task* required comprehension of several paragraphs. Finally, the last item of the *Multiple-Choice Task*, as discussed which asked for a sub-title for the text required processing and comprehension of almost all the text.

To have a more precise measure of the number of items in the IELTS RCM that tapped into the sentential-textual continuum of comprehension, the number of test items for each level of comprehension was counted. Table 5.33 shows the frequency and percentage of the items for each level of comprehension.

Table 5.33.

Summary of comprehension level of the IELTS RCM test tasks

Answer level	Number of items (percentage)	Test tasks
Sentence	8 (20%)	<i>True, False, Not given Task, Matching Features Task</i>
Several sentences	22 (55%)	<i>Diagram Completion Task, Summary Tasks 1 and 2, Multiple Choice Task, Yes, No, Not given Task</i>
Paragraph	7 (17.5%)	<i>Matching Headings Task</i>
Several paragraphs	2 (5%)	<i>Multiple-Choice (two answers)</i>
Text	1 (2.5%)	<i>Multiple Choice Task</i>
Total	40 (100%)	

As the results in Table 5.29 show, 20% (N=8) of the items required comprehension of one single sentence and (55% (N=22) items required comprehension of few sentences. This accounts for 30(75%) of comprehension. Two items (5%) tapped multiple-paragraphs level of comprehension and text level comprehension was the least frequent target of test tasks with one single item (2.5%).

Based on these numbers and percentages, it can be concluded that the IELTS RCM basically taps the low-level comprehension of specific details at sentence and inter-sentence levels while the high-level discourse comprehension skills are poorly measured by the test tasks. Based on these frequencies and percentages it can be concluded that high-level textual comprehension is severely underrepresented by the IELTS RCM.

5.6. Reading dimensions in the IELTS RCM (Brown and Abeywickrama, 2004)

One final step taken in the content analysis of the test was applying Brown and Abeywickrama's (2004) model of reading dimensions to examine the scope of the reading measured in the IELTS RCM. In this section, results of this phase of content analysis are reported.

Brown and Abeywickrama (2004) suggested four dimensions of reading comprehension, each representing a certain scope at which reading is practiced and the specific focus on reading skills (See the Glossary for brief definitions; Chapter Three, Section 3.2.2 for a detailed discussion); 1) perceptive reading, 2) selective reading, 3) interactive reading and 4) extensive reading. Some of these reading dimensions were observed in the IELTS RCM texts. Based on Brown and Abeywickrama's (2004) taxonomy of reading dimensions, results of content analysis of the IELTS RCM represented some features of selective and interactive reading but neither perceptive reading dimension nor the extensive dimension were represented in the test. For example, in the *Summary Completion Tasks 1 and 2*, and the *Diagram Completion Task* the focus was on the use of few lexical items in few paraphrased sentences which correspond with characteristics of selective reading. Receptive reading dimension characterizes reading skills and activities of beginner readers who are basically concerned with encoding graphemes at lowest level of phrase and words. Extensive reading, on the other hand, characterizes high level of processing with focus on discorsal and topical aspects of the text in professional setting.

The texts were much shorter than typical academic texts. Nor were they academic in terms of their topic, tone, and genre features. For instance, in almost any academic source, there is an argument which is being discussed in the light of some evidence for or against it while in the IELTS RCM texts there is no such argument made. Or in terms of genre features, none of the IELTS RCM texts shared main features of academic genres such as research articles or book chapters. Additionally, academic genres are diverse and include a wide variety of reading material including technical report (e. g., lab reports), professional journal articles, reference material, textbooks, theses and dissertations, applications, forms and directions, questionnaires, announcements, emails, etc. The sample texts used in the IELTS RCM sample test fell short of representing any of these main genres of academic reading. Overall, none of the texts fell within perceptive and/or intensive reading dimensions and all test tasks represented either selective reading or interactive reading dimensions. Table 5.34 presents result of reading dimensions tapped by the IELTS RCM test tasks.

Table 5.34.

Reading dimensions of the IELTS RCM test tasks

Dimension of reading	Frequency and percentage	Test tasks
-----------------------------	---------------------------------	-------------------

Selective reading	30 (75%)	<i>True, False, Not given, Matching Features, Diagram Completion Summary Tasks 1 and 2, Multiple Choice, Yes, No, Not given</i>
Interactive reading	7 (17.5%)	<i>Matching Headings</i>
Interactive	2 (5%)	<i>MC (two answers)</i>
interactive	1 (2.5%)	<i>Multiple Choice (last item)</i>
40 (100%)		

As shown in Table 5.34, most test tasks and test items represented the selective dimension of reading. Some 30 (75%) items represented selective dimension and only 10 (25%) items represented interactive dimension. Perceptive dimension and more importantly extensive dimension which characterized the main features of academic reading are totally absent in the test. Based, on these results, it can be argued that the IELTS RCM under-represents the main academic features of reading comprehension.

5.7. Summary of the Chapter

This chapter focused on the content analysis of the test as one of the main sources of validity evidence. It is arguably a useful source of validity evidence as it allows researchers to examine the linguistic, textual, topical and cultural contexts that shape test performance. The texts used in the sample IELTS RCM consisted of three independent reading passages each followed by three different and distinct test tasks. Different linguistic features of the text such as lexical density and grammatical complexity and readability of the texts were analyzed. Moreover, textual features, text topics, and cultural aspects of the text were also discussed. Furthermore, features of test tasks and test items were analyzed in terms of their language and their relationship to the text. The relevant information for each item was identified and the lexico-grammatical and textual clues that can help answering them were also highlighted. Results indicated that the test focus is on literal comprehension at sentence and inter-sentence levels. Inferential comprehension and textual comprehension were marginally tapped by the test tasks. Findings of content analysis were integrated with other sources of evidence to discuss the construct of the IELTS RCM. In the following chapter, another source of validity evidence is considered, namely, the responses of L1 and L2 test takers to the IELTS RCM sample tes

CHAPTER SIX: RESULTS OF TEST TAKERS' ACCOUNTS

6.1. Introduction

Chapter Five presented findings from the content analysis of the IELTS RCM, which allowed for an initial consideration of the construct of reading comprehension operationalized by the test. This chapter examines another source of validity evidence, namely, test takers' accounts of the skills, knowledge sources, processes, and strategies (SKSPs) used during their performance on the test. The results addressed the second research question raised in this study, by exploring the SKSPs that test takers report using during test performance. To this end, as discussed in Chapter Four, a sample of 21 test takers from different language backgrounds (L1 and L2) and levels of language proficiency (low L2 proficiency and high L2 proficiency) participated in the study. They took an IELTS RCM sample test and provided immediate retrospective verbal reports of the processes and strategies they used in taking the test. Their verbal reports were transcribed, coded, and analyzed to address the second research question and provide another empirically derived window on the test's operational definition of the reading comprehension construct.

In the sections which follow below, I begin by presenting the results of test takers' test performance as reported in their verbatim accounts of each of the nine test tasks and the patterns that emerged from their performance. This is followed by findings from the first cycle of coding which were organized the codes into three main themes (Reading, Searching, and Answering). Next, results of the data collected through "Test Performance Observation Scheme" (TPOS) will be reported. Finally, results of task difficulty that emerged from integrating test takers' IELTS RCM test score, data from TPOS, and testing experts' judgements of task difficulty will be discussed.

However, before reporting the results of the first and the second cycles of coding a few points need to be made about the terms used in naming and describing the codes and the cognitive processes they represent. Terms such as processes, strategies, skills, metacognition, careful reading, expeditious reading, etc. are subjective terms and open to different interpretations. Therefore, they need to be clearly and precisely defined. (See the Glossary which provides an operational definition for these terms). Three key terms that are relevant to the

findings reported in the next section include *process*, *strategy*, *category*, and *theme* which are defined here.

- **Category: Process** was defined as a tacit automatic operation that is used to respond to a test task. Based on Khalifa and Weir's (2009) model, reading processes include word recognition, syntactic parsing, forming semantic proposition, inferencing, building text representation and creating a mental representation.
- **Category: Strategy** was defined as a deliberate goal-directed action used to accomplish reading, searching or answering a test task.

Although for purposes of this research strategy and process are teased apart, in reality they often overlap as test takers engage in responding to a test task.

- **Themes: Reading, Searching, and Answering** were identified as recurrent patterns of response that conceptually incorporate processes and strategies used in test taking. For example, readers read the text, read the test item, highlighted key words in the test item, went back to the text, searched for relevant information in the text, read the relevant information, and answered the test item. Theming the data or moving from categories to a conceptual level based on the identification of recurring patterns, allows for definition of the construct of interest, namely the IELTS REM, as the construct of reading comprehension was operationalized by the test for the test takers considered in the study.

Figure 6.1 presents the framework used for data analysis moving from real data (retrospective verbal reports) to theory (the construct of the IELTS RCM). It lays out a schematic representation of the inter-relationship of “*categories*” (processes, strategies,), “*themes*” (reading, searching, answering) that emerged from coding of the test takers’ accounts. Guided by the research questions in the section which follows below, test taker accounts of each of the nine RCM tasks is provided along with a selective number of verbatim quotes. Overall 70 codes that emerged from the data provides a complete list of codes with test taker comments that serve as examples of the verbatim sources of the codes (See Appendix K, L, & M). The Appendix also includes an example of a coded

transcript to illustrate the coding in relation to test taker discursive accounts of their test performance.

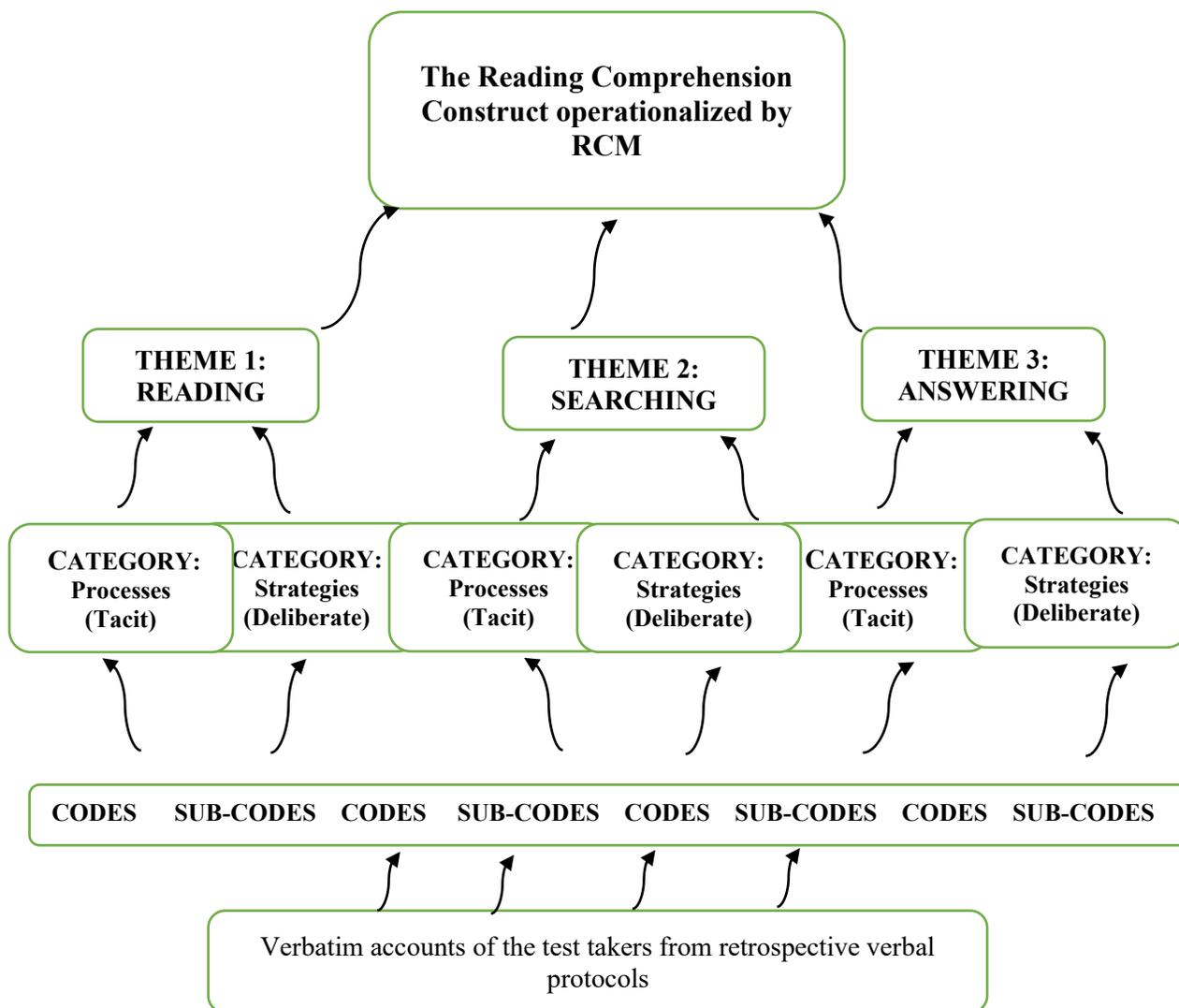


Figure 6.1. Schematic representation of data analysis framework moving from data to theory: The inter-relationship of *processes*, *strategies*, categories, themes and construct

6.2. Construct of the IELTS RCM Test Tasks

In this section, based on the results of test takers' accounts as reported in the first and second cycles of coding, the construct of each of the nine test tasks of the IELTS RCM in terms of the skills, knowledge sources, processes, and strategies (SKSPs) (cf. Gorin, 2006) is

presented. Also, similarities and differences observed among different groups of test takers are presented.

6.2.1. Construct of the True, False, Not given Task

The *True, False, Not given Task* asked test takers to compare the statements presented in the test task with the text and decide if they were true, false, or not given. As discussed in Chapter Five, See Section 5.2.1.1.) the items involved both literal and inferential comprehension.

As shown in Figure 6.2, two main patterns emerged from the test takers' performance. In the first pattern which was mostly used by L2 test takers, test processes involved highlighting key words in the item, scanning them in the text, skimming or careful reading of the information relevant to each item and get the answer to the item. One of the L2 test takers reported,

First, I looked at the task. I read the first and second items and underlined the key words to know what to look for in the text. For question 1, I underlined the words “*doubt, The Mary Rose, and sink*” and for question 2 the words were “*the only ship*” and “*battle*”. (Hash)

This L2 test takers clearly indicated that task performance involved highlighting key words in the item, scan them in the text to locate the relevant information, read the relevant information, to get the answer which was at sentence or inter-sentence level.

In the second pattern which was mainly used by the L1 test takers, they began with reading the whole text first and developed the gist of the text. Some test takers could recall the relevant information from their first reading and guessed an answer which they double checked with the text. They could quickly locate the relevant info in the text and answer the item. One of the L2 test takers who used recall information from first reading reported,

Then I went then to question number 1 and I remembered it at the very beginning of the first paragraph which talked about what had caused the ship to sink and that It had sunk before. So, I was pretty sure from reading the text the first time that they did not know for sure that the statement was true. So, I knew it is false. (Cour)

She clearly made use of her first reading in answering the test item without going back to the text. Other L2 test takers did not have an answer and went back to the text to find one. Two

of the more successful L2 test takers read the whole text carefully before doing the test task. So, they went through the same processes.

Test takers found the answers to the *True, False, Not given Task* in one or two sentences. For instance, one of the successful L2 test takers reported,

First, I carefully read the first two questions to help me locating the key words and the information in the text. Then I went back to the text and could locate the relevant info. I could find the answer in the sentence which reads, “accounts of what happened to the ship vary”. Based on this sentence, I could say the answer is true. (Hona)

He clearly indicated he used scanning the key words in the text and found the relevant information in one sentence which helped him answer the question.

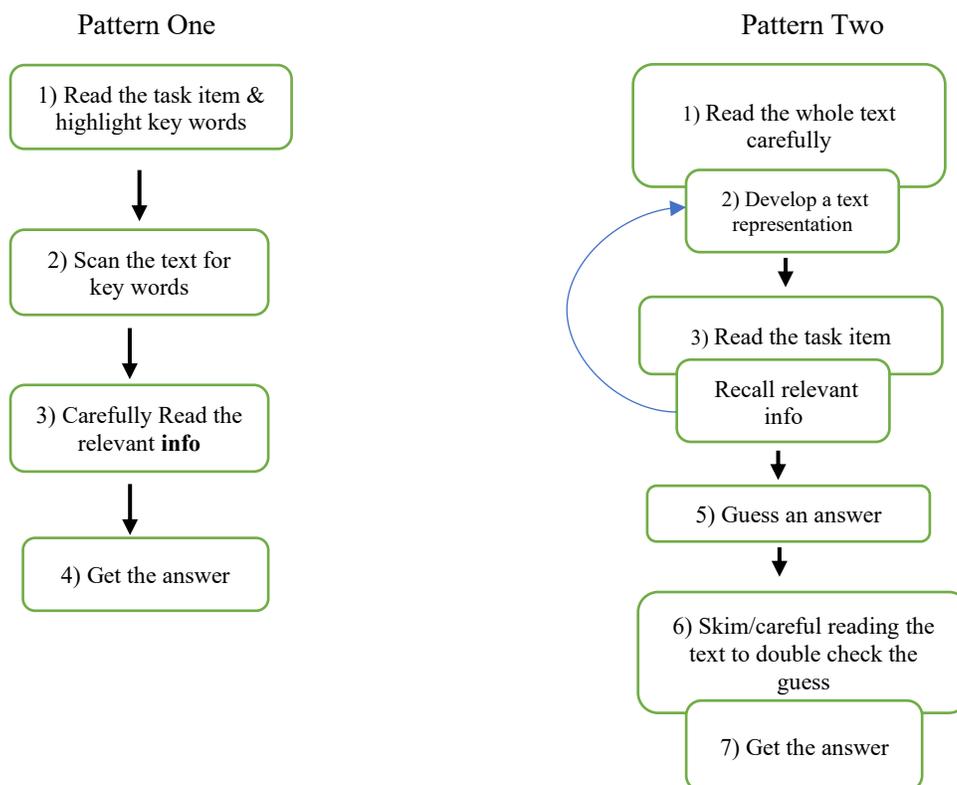


Figure 6.2. Test performance in the *True, False, Not given Task*

Groups of test takers differed in the way they did the task. one major difference between the L1 and L2 test takers was reading or not reading the whole text. The L1 test takers and two of the more successful L2 test takers read the whole text before they started doing the test task while none of the less successful L2 test takers did so. Reading the whole text before doing the

test task contributed to the consequent processes and test performance processes. Another major difference was related to the search for the relevant information in the text. For the L1 test takers, the search for locating the relevant information in the text, which was in the form of skimming, was faster and more efficient while for the less successful L2 test takers search processes involved careful reading of the whole paragraph and was very time consuming.

Analysis of the less successful L2 test takers' performance showed that the less successful L2 test takers struggled with understanding some of the questions and often times they could not comprehend some of the ideas presented in the text. They also struggled with locating the relevant info. One of the less successful L2 test takers reported,

I struggled to find the relevant info for question 3 and I could not finally find any relevant info. So, I chose Not Given for question 3. (Alip)

Also, in their search for locating the relevant information, they had to search more paragraphs before they could get to it. It took a significant amount of their time just to locate the relevant info. They could not efficiently skim the text to get to the relevant info. They were more confident to read carefully and miss no points that might be relevant to the item. In fact, they wondered they may miss some ideas had they skimmed the text. They thought by skimming they may miss some of the ideas that they need to answer the question. So, they preferred to read the text carefully and read more text to make sure they miss nothing. Careful reading of more text was cognitively more demanding and put more burden on their limited processing capacity. At the same time, it was also time-consuming which was the main source of anxiety they had in their test performance. As one of the less successful L2 test takers said,

I do not like "Not given" items because, finding the answer is hard. I need to read much more to answer it. I do not like it because it demands much more from me and you do not have much time. (Mosa)

While processing the text, they also faced several new words in the text which was an obstacle in relating the phrase and sentences together to get the main point of the paragraph. It is interesting to note that the "not given" items required inferential comprehension which is

obviously more demanding. For items addressing literal meaning, they felt more comfortable while for items involving inferencing their struggle exacerbated.

One more observation made in the test performance of the *True, False, Not given Task* was their use of lexical clues. All test takers tried to use some lexical clues that could help them relate the test items to the text. The clues were very conducive in getting the answers. The same clues, however, caused more confusion for the less successful L2 test takers because they did not know some of the words used. One of the more successful L2 test takers commented,

Or in question 3, [reading the item]* “*most of one side of the Mary Rose lay undamaged under the sea*”. The text uses synonym in paragraph 2 line 4 it reads, “*nearly all of the starboard half survived intact.*” If one knows the meaning of the word “*intact*”, he can get the answer. I could simply relate the two. For this question, I did not need to read the whole paragraph and I could read half-way through it and get the answer. (Mart)

*NOTE: [Square brackets] indicate my observation of what the participant was doing while they accounted for a test response.

She clearly indicated how the lexical clues saved her reading more text to get the answer. She used the lexical clues and could answer the item without reading the whole paragraph. The less successful L2 test takers could not attend these clues as much as the L1 test takers and the more successful L2 test takers.

As to the test taking strategies, some of the strategies used in processing *the True, False, Not given* items included; reading more to make sure nothing is missed, test-wiseness strategy, recall information from previous reading, recall question while reading, noticing the order of questions in the text, and tagging paragraphs with a word or phrase. For instance, one of the more successful L2 test takers commented, “*As I was reading, I knew in such tests it is important to have dates in mind. I also know that names are important.*” (Angl) She was test-wise and aware that dates, and names might be important in answering some test items.

What does The *True, False, Not given Task* Measure? Based on the skills, knowledge sources, processes, and strategies (SKSPs) used by the test takers, the *True, False, Not given Task* seems to measure literal and inferential comprehension of specific details at inter-sentence and paragraph level but not the general idea of the paragraph. All the questions focused on one

specific aspect of an idea explicitly or implicitly presented in the text. Some items required literal comprehension where the information is more explicitly expressed in the paragraph while other items are implicitly discussed and call for inferential comprehension. Task performance also involved skimming, scanning, and careful reading and use of lexical clues that depend on vocabulary knowledge of the test taker. The task also involved reading speed skill to cope with the time limit set for test performance.

6.2.2. Construct of the *Matching Features Task*

In the *Matching Features Task*, four statements and a table of dates were presented to the test takers and they were asked to decide what date matches each statement. Test takers had to use three sources to get the answer; the text, the statements, and table of date options. Some statements were literal while other statements involved inferential comprehension. This was noticed by some test takers. One of the L2 test takers reported,

The ideas in the *Matching Features Task* were implicit and you needed to read more carefully to relate it to the text. I read a whole paragraph for the dates. (Khei)

Task performance involved scanning the date, locating the relevant information in the text, careful reading of a couple of sentences or a paragraph related to a date and match the correct date with the statements. To get the answer, test takers had to go back to the text to study the dates and see when those events happened. Scanning the dates in the text was the starting point. Then, test takers skimmed the section around the date to understand what is presented then carefully read the sentences that seemed relevant to the item. As there were several dates in the text, test takers had to read the relevant section carefully. Answering the test items involved comprehension of several sentences in the relevant paragraphs. The relevant paragraphs included a number of ideas and events happening around a date, so it was vital to read all the paragraph carefully to make sure they get the answer right. For some items, it was not just one single sentence that provided the answer and more reading was needed. For instance, Mart, one of the L1 test takers reported,

I just looked at where dates were and read the sentence around the date. Then I said this has nothing to do with this statement and then I moved to the next date. This was basically

what I did. I focused on the dates in the text. I started reading the sentence with the date and if it had nothing to do with the statement then I moved to the next one.

Another L1 test taker reported, “... *If it (the sentence) was like it might have something to do with the item, I would read a bit more text after it to see if it is still related. I would read carefully then.*” (Mart)

Based on the skills, knowledge sources, processes, and strategies (SKSPs) observed in the test performance of the test takers, three different patterns were used for doing the *Matching Features Task*. In the first pattern (Figure 6.3), the starting point for some test takers was the text and the date mentioned in the text. Then the date was checked against the table of dates provided in the test task. If the date was included in the table, the test taker would continue reading around the date in the text and check if the idea discussed matched the statement. In case the date mismatched the date in the table, the test taker would go to other section of the text and read another part that contained a date. The process could continue until they get the answer. Task performance involved several back and forth movements between the text and the test items.

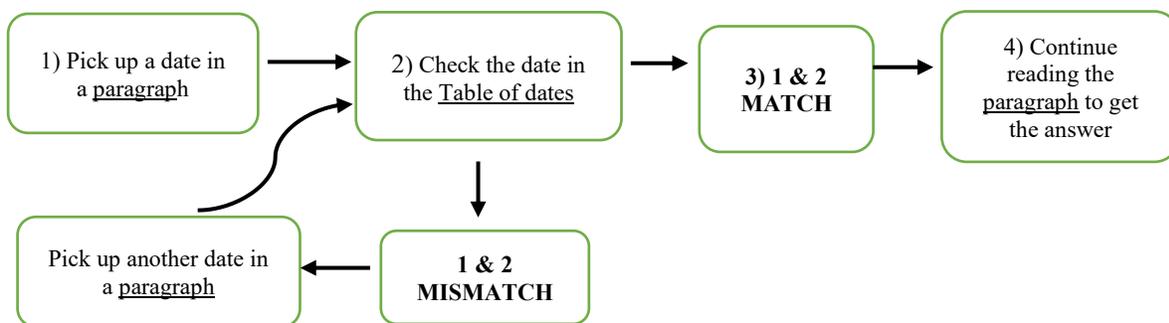


Figure 6.3. Test performance in the *Matching Features Task*: Pattern one

In the second pattern (Figure 6.4), test takers first skimmed the statements and picked up a date from the table of dates. Second, they went back to the text and scanned the date in the text and after locating the date, read the sentences before and after the date carefully. Third, they checked if the relevant information matched the statements. When they matched, they chose the correct answer otherwise they would repeat the same process by picking up another date from the table of dates and go through the same processes to get to the answer.

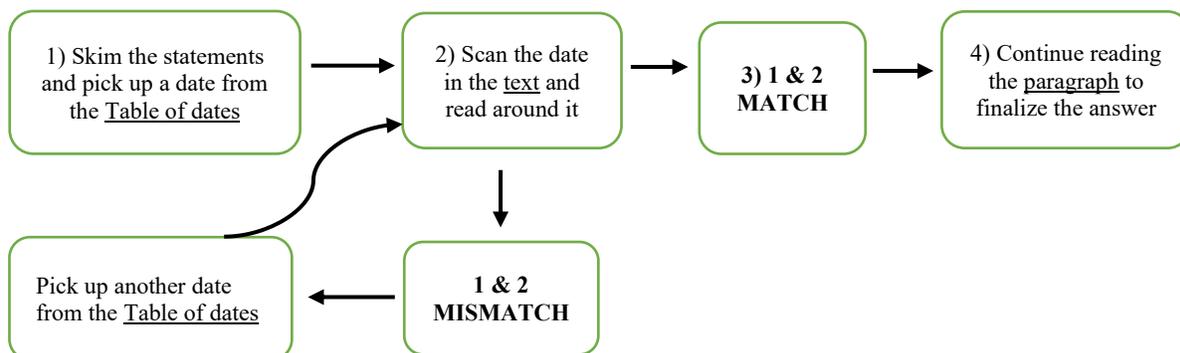


Figure 6.4. Test performance in the Matching Features Task: Pattern two

In the third pattern (Figure 6.5), test takers started with the text and read the text that contained one of the dates. Then, they checked the relevant information against the statements in the task without using the table of dates. In this pattern, the task was simplified by ignoring the table of dates. Paragraphs that contained a date were read and checked against the statements and they treated the task as the *True, False, Not given Task*. The one that matched the idea was chosen as the answer. The main feature of this pattern of performance was ignoring the table of dates altogether. Test takers simply read and compared the text and the statements to arrive at an answer. Some test takers used the table of dates just to make sure the date they had chosen was included in the table.

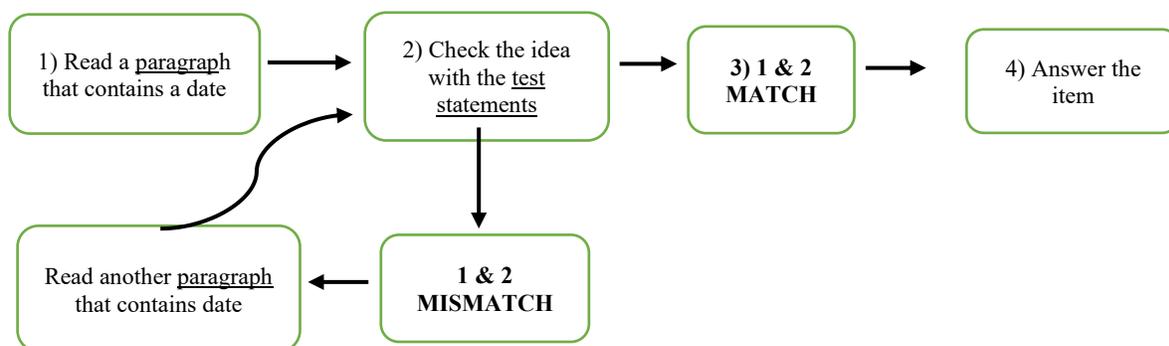


Figure 6.5. Test performance in the Matching Features Task: Pattern three

These patterns were randomly used by the test takers and no associations were found between the pattern used and the group of test takers, but the third pattern was more frequently used by the L2 test takers. They read a paragraph that contained a date and read the idea related to the date then they went back to the statements to check if what they had read matched or

mismatched the statements. However, the choice of using any of these patterns was basically based on reading less and saving some time. L2 Test takers thought the pattern they chose could save them some time.

Two main differences were observed between the L1 test takers and the less successful L2 test takers. First, L1 test takers and some of the more successful L2 test takers were more efficient and quicker in locating the relevant information. Second, L1 test takers did more skimming to locate the relevant information, while the less successful L2 test takers had to read the text around the dates more carefully which cost them a lot of time. One of them reported,

I read two or three paragraphs to find the information. As I was not sure what the text is saying, I had to re-read the sentences very carefully. I think I spent more time on this task.
(Zade)

With regard to locating the relevant information, test takers had to search three different paragraphs which discussed the chronology of finding “*The Mary Rose*”. Since the test task statements were not in order, the relevant information could be in any of the paragraphs. So, test takers had to search the three paragraphs to locate them. They achieved the search for the relevant information by skimming, scanning, and/or careful reading of at least 3 paragraphs. In locating the relevant info, in most cases the test takers read a few sentences before and a few after the dates, not the whole section or paragraph that contained the date.

One feature that distinguished the *Matching Feature Task* from the *True, False, Not given Task* was that the text presented several dates related to the events described in the text. Therefore, understanding chronology of events could help test takers better see what had happened and when. Such awareness of paragraph structure and text structure could facilitate locating the relevant information more efficiently. As one of the L1 test takers said,

The dates task may need understanding of the whole text. If you understand the whole text and the chronology, you can better do the dates. I underlined the dates because I knew there will be a dates task next. (Broo)

One strategy that mostly used in this test task was “Attention to certain features of the text”. The strategy was weakly associated with the *Matching Feature Task*. As the task dealt

with dates, test takers were aware of the importance of dates in the text and tried to pay due attention to them and focused on reading the dates and the information surrounding them. One of the L1 test takers reported,

[Reading the text] I knew the questions would do with dates, so I decided that I was going to basically just skim through the text for dates and see what the surrounding text is. (Isab)

What does the *Matching Features Task* measure? Based on the findings of the study, *the Matching Feature Task* targets literal comprehension of specific details at inter-sentential and paragraph levels. Awareness of paragraph and text structure, use of vocabulary and grammar knowledge can help task performance. Task performance also involved skimming, scanning and careful reading. Like other test tasks, reading speed skills are key in test performance within the time limit allowed.

6.2.3. Construct of the Diagram Completion Task

The Diagram Completion Task was basically asking for filling in the blanks of the diagram textboxes with two-word phrases from the text. Unlike the other tasks, the stem included two components; 1) the diagram and 2) the textbox with a blank. The textboxes were part of a diagram that illustrated stages of raising a shipwreck. The *Diagram Completion Task* included a diagram as part of the stem. It could help test takers do the task by building on the information that is included in the diagram. Test takers could use three sources of input for processing; the text, the task which involved a diagram and textboxes.

The main pattern of test tasks performance (Figure 6.6) was as follows. Test takers, first, studied the task and carefully examined the diagram to see what the task was asking. Then they read the textboxes that contained the blanks. Next, they highlighted some key words from the textbox and scanned them in the text. Most of the test takers knew where to look for the relevant information in the text simply because they had either read the whole text before (L1 test takers) or they knew no questions were asked about the last two paragraphs.

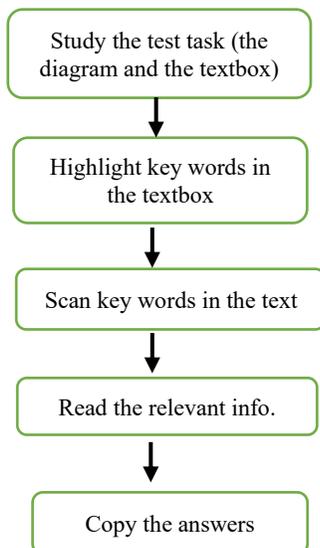


Figure 6.6. Pattern of test performance for the *Diagram Completion Task*

Most test takers were confused by the task and struggled with understanding details of the processes involved in “Raising the Mary Rose”. They looked for exact match of phrases and sentences in the text without having any clear view of the raising processes. They found the relevant text vague and not meaningful to them. The confusion was related to both the diagram and the textboxes. They did not have a clear picture of the sequence of the processes and the specific steps involved when they read the text. The whole processes described in the text were not clear to them. One of the L1 test takers who managed to do the task successfully reported,

But I had problems with the *Diagram Completion Task*. I tried to answer the first question, No. 9, and I was already confused so I read through the whole two paragraphs and I went back to the question and then I went back to the paragraphs because I still I do not understand it. I had problems with what I should be looking for because if I do not know what to look for, I am in trouble. (Nabi)

In spite of doing the test items successfully, most test takers, except for very few, did not have any background in such engineering projects and the technical tools used in the raising processes. The instruments mentioned in the text were mostly unfamiliar to them. Most of them did not find the diagram informative or helpful in clarifying the processes. They found it difficult to make sense of. In one case, the diagram caused confusion and the test taker thought the arrows are pointing to the same thing, so she chose the same answer for two different blanks. Some even

did not use the diagram at all and ignored how it illustrates the processes described in the text. Most of the more successful L2 test takers (3/5) and less successful L2 test takers (4/5) used just the textbox to do the task and did not make use of the diagram while most L1 test takers (6/10) used both the diagram and the text box to understand the processes.

Interesting to note that most of the less successful L2 test takers identified themselves as pictorial learners. One of the test takers commented, “*I am a pictorial person, but I should also say that I could not make use of the diagram and it did not help.*” (Mosa). They could not use the diagram especially the second part of the diagram and relied just on the text boxes provided in the task. Another successful L2 test taker (Angl) said, “I am not used to graphic representations and diagrams and it was technical to me.” They mostly relied on the text boxes which included some key words and lexical clues that guided them in searching and locating the relevant information in the text. As one of the test takers said, “*I cannot explain how they did the thing. I just know that they took several steps to raise the ship.* (Mitc)

Reading the textboxes, some test takers could see they are paraphrase of the same idea from the text with a number of same words repeated and a number of synonyms and phrases. One of the more successful L2 test takers said,

Then I underlined some key words used in the text boxes in the diagram so that I can track them in the passage. To me this is like a montage and I had to piece things together from the text in the blank. (Jaha)

Most L1 test takers used their knowledge and awareness of paragraph/text structure in their test performance. They were aware of the text structure and roughly knew where these processes were described. They mentioned they knew where to look for the relevant information. They knew the idea of *Raising the Mary Rose* was discussed in the last two paragraphs. They also knew it was a process paragraph and applied this knowledge of text type and paragraph structure to their search for answer. Knowing that the text discusses three stages of *Raising the Mary Rose*, they had an idea which items are related to stage 1 or stage 2. However, they could not fully comprehend the exact steps, instruments, and technical terms used in the text.

The relevant information for answering the questions were presented in the last two paragraphs. The second last paragraph accommodated the answer to questions 9-10 and the last paragraph provides the answer to questions 11-13. Test takers could locate the relevant section and narrowed down their search to few sentences from each paragraph. Their search involved skimming, scanning, and careful reading. Unlike the *Matching Feature Task* which involved searching one or even more paragraphs, for most test takers search for the relevant information was basically searching a few sentences in the paragraph. In their search for the relevant information in the text, the test takers focused on locating the same words or some synonyms in several sentences in the text which required some paraphrasing skills.

As the textboxes contained a number of lexical clues and synonyms, vocabulary knowledge played an important role in the *Diagram Completion Task*. Searching for words and phrases that matched the gap was at the center of the processes. The textbox contained some lexical clues to help the test takers locate the relevant information in the text. However, not all test takers could use these clues. The less successful test L2 takers faced many new words that added to their confusion of finding the answer in the text. one of them said,

My vocabulary is not rich, and I think this is why it took so much time to do the task. I found many new words in the text and I could not read as quickly as I thought. I needed to read more and reread to get the main ideas. I had to guess some of my answers in these tasks. I also tried to relate the sentences to other sentence and get some understanding of each paragraph. I could not fully and clearly understand the ideas in the text. There were several new words. (Mosa)

Test performance was also influenced by background knowledge which played a key role in the test takers' performance. Technicality of the target words and phrases was a challenge even the L1 test takers. Most test takers claimed the task calls for some technical vocabulary, background knowledge, and background in understanding engineering projects. One of the L2 participants who had some background in mechanical engineering was quite familiar with engineering projects and the technical terms and the technicalities involved in the task. He was very successful in answering all the questions in the task correctly, in spite of being less successful in his overall performance on the whole test. His background knowledge put him at an advantage,

My background in mechanical engineering helped me better understand the processes. I knew what “stabbing guides” are and the challenges involved in using them, I had some experience about it in my job. I was sure I could use it for (filling in) the blank. It made sense. It was much easier to choose. (Hash)

Background knowledge influenced even the reading and comprehension process. As one test taker said she avoided reading the last two paragraphs of the text in her first reading simply because she found it too technical to understand. This was in the context where she had already read the rest of the text and had developed quite a clear picture of the chronology of events and the ideas discussed.

Test takers also used different strategies for the *Diagram Completion Task*. As the diagram and the textbox were not clear enough for most test takers, they simplified the task by redefining and re-phrasing the text box into more sensible statements and ideas. One of the successful L2 test takers said,

Question 10 is asking for something that prevented the hull from being sucked into the mud. So, I knew what I am looking for. I was looking for something in the text the legs have gone into. (Jaha)

Another L1 test taker, said,

I was looking for question 10 which was something to prevent the hull from being sucked into the mud so with that in mind I just continued from where I had read. (Kyle)

Their account clearly indicated that they re-phrased the textbox to a more meaningful question so that they can have a clear idea what they are looking for. Other strategies they used included “use of paragraph and text structure knowledge”, “delay answering”, “use of background knowledge”, and “noticing grammar”. One of the L1 test takers said,

I read the textboxes first and based on the syntax of each sentence and the blank I guessed what kind of word I would be looking for, a noun. (Broo).

This clearly shows her attention to grammar of the sentence helped her determine the type of word she needs to fill in the gap.

In summary, the *Diagram Completion Task* aimed at measuring vocabulary knowledge in the context of process paragraphs, but as shown in the test performance of most test takers, it involved matching exact words rather than comprehension of the relevant sentences and paragraphs. The task failed to tap sentence or inter-sentence comprehension. Test takers struggled doing the task because they did not understand the processes described the text. The diagram was not clear to them, they had no background knowledge nor did they know the technical words used in the text. Most test takers could successfully do the task without understanding what these processes really were. As one of the test takers said, “*If I answer them right, which I think I did, it does not really mean that I really understood the processes the text explains.*” (Angl) They mostly relied on lexical clues to find the answer not comprehending the processes described in the text. Use of background knowledge and technical vocabulary were serious obstacles to doing the task. This was best described by one of the L1 test takers,

And I had trouble with like it is talking about these technical processes. I do not know what any of these things really are to even understand what they are talking about. I was not familiar with these technical issues at all. (Nabi).

Overall, the *Diagram Completion Task* was not an easy straightforward task to do. One of the less successful L2 test takers mentioned how she struggled to understand the process, “*The diagram was the most challenging because I feel it is vague to me and I could not clearly understand what the processes are.*” (Mosa)

What does the *Diagram Completion Task* measure? The *Diagram Completion Task* measured technical vocabulary knowledge in the context of word matching and phrase matching rather than literal comprehension of some processes and specific details at sentential and inter-sentential levels. Paraphrasing skill, grammatical knowledge, awareness of text type and text structure, and more importantly background knowledge were also involved in test performance.

6.2.4. Construct of the *Matching Headings Task*

The *Matching Headings Task* asked test takers to match each paragraph with one of the heading options. It involved both literal and inferential comprehension. Few heading options were straight forward and could be directly related to the text while most other headings involved inferential meaning and could only be inferred from the paragraphs by integrating details of the paragraph into a meaningful whole. For example, in paragraph G, different opinions about the Easter Island were discussed while the language used in the heading option is different. The option reads, “*two opposing views about the Rapanui people*”. For this item, test takers had to infer two opinions being inferred and that the opinions are opposing. None of these meanings were explicitly stated in the text. These inferences had to be drawn by the test takers before they could choose the right answer. Or in the heading option “i” which reads, “*evidence of innovative environment management practices*”, the test takers had to read through paragraph D and put several examples of such innovative practices together to choose the right option. The paragraph talks about Rapanui people protecting the resources of their wind-lashed, infertile fields, building thousands of circular stone windbreaks and gardening inside them, and using broken volcanic rocks to keep the soil moist which are all examples of innovative environment management. Piecing these practices together as innovative practices for management of the environment was not a straightforward task. It required applying some world knowledge and inferencing. Without inferencing test takers could not arrive at this heading option. As Mitch one of the L1 test takers put it,

This heading task was like involving a lot of interpretation and you needed to go back to the text read the paragraphs again and, you know, getting the entire reading skimming it and try to eliminate what I knew and then I think it needs more interpretation.

Another test taker noticed the inferential nature of the task and reported,

This (item 15) is talking about diminishing food resources which is what this question is about because it talks about how they cut down the forest for firewood. This guy Diamond thought they cut the forests and then they could not make canoes and ended up eating birds and soil erosion. All the horrible stuff they thought had happened to the island this is why I

put the diminishing food resources as the answer because it is what it was talked about in the paragraph.

Performance on the *Matching Headings Task* involved a brief look at the task to identify the type of task followed by careful reading of each paragraph one by one and going back to the heading options to find the best match. Test performance also involved several back and forth movements between the paragraphs and the heading options. As mentioned above, to choose the right heading test takers had to get the gist of each paragraph and comprehend some of the details discussed. Details could help test taker connect the dots and get the gist of the paragraph. The right heading could not be directly selected because most headings were abstract, and the test takers had to infer them from the text. Without integrating the details together, test takers could not choose the right heading because the headings are more abstract and need some kind of reinterpretation and inferencing. For example, one L1 participant said, “*Paragraph C talks about cannibalism, crop and soil, it should be about food and diminishing resources.*” These inferences could not have been drawn if it was not for the details mentioned in the text.

To draw the inferences, test takers had to do a lot reading and re-readings of each paragraph and develop a textual representation of the paragraph.

Most test takers read the first sentence or the first two sentences of each paragraph more carefully assuming that it contains the main idea of the paragraph. This proved wrong for most paragraphs because they did not have a topic sentence to state the main ideas. For instance, neither paragraphs A nor paragraph B provided the main idea in a single sentence and test takers had to read and re-read the whole paragraph to get the main idea.

Even comprehending the paragraph was not enough to choose the right answer because the heading options used different language and vocabulary. In fact, the key words used in the heading options were not used in the text, nor their synonyms be different from what they had comprehended, and they had to adjust their comprehension to accommodate the headings inferred from each paragraph. For some paragraphs, lexical clues and language was the main source of clues (paragraph E) while for most item (paragraph D) it was through creating or discovering the interconnection of ideas and inferencing that guided the test takers to the answer. Test takers had to move back and forth between the text and the headings, compare different

options and fill the gap between the more explicit ideas of the text and the more abstract idea of the heading.

One noticeable feature observed in the test performance of all test takers was their use of some key words, phrases or sentence from the text in finalizing their answer. In almost all the paragraphs there was a sentence or key phrase that provided the best clue that matched certain heading. These clues triggered the choice of the heading and test takers used some lexical clues that could support their answers. For example, in paragraph E, heading iv “*A theory which supports a local belief*” which is the correct heading, included the word “backs up Rapanui folklore” in the text. These lexical clues (back up-support/local belief-folklore) were very instrumental in finalizing the choice of the heading. One of the L2 test taker reported,

[Pointing to paragraph E] I could relate words such as, [Reading from the text], *back up-support, local belief-folklore*, from the text to the heading. They are synonyms. The connection is very telling, and I was sure this is the answer.

For some paragraphs the heading options were more explicitly related to what the texts was discussing. For example, paragraph F talks about things that were out of the Islander’s control and the lexical clues in the heading can be directly related to the text. The heading mentions “*outside inhabitants’ control*” and the text talks about “*people were not wholly responsible*” which are basically the same meanings and by use of lexical inferencing it can be chosen as the correct answer. Another example was Paragraph C, which included a key phrase, i.e., “self-destruction” which could help choose the right heading.

To get an overall understanding of each paragraph, the L1 test takers heavily relied on their first reading of the text. They started off by first careful reading of the whole text. Then they skimmed the headings and went back to each paragraph and read or skimmed them one by one to choose the right heading. For instance, one of the L1 test takers commented,

The second time I read the paragraph I did not read for any details I just skimmed through it to remind me what that specific paragraph had been about. I could recall a lot more from my first reading by seeing a couple of words. (Cour)

She clearly mentions that her second reading was not careful reading. It was just skimming and reassuring herself what she recalled is right. The L2 test takers, on the other hand, started by reading the first paragraph and tried to choose a heading immediately after they read the first paragraph. They had not read the whole text before they started answering the task. The L1 test takers either skimmed each paragraph or read it carefully to get to the answer while the L2 test takers did not use skimming and read the whole paragraph carefully to get a gist of the paragraph and develop an overall understanding of the paragraph.

An interesting and unique observation in test performance of the L1 test takers was their critical attitude towards both the text and the heading options. For some L1 test takers, the *Matching Headings Task* was confusing because they found a serious mismatch between the paragraphs and the headings. They criticized the headings for covering just a few sentences of a paragraph rather than a full coverage of the whole paragraph. One of the L1 test takers reported,

I found that the list of headings did not capture the paragraphs, especially the last three headings that I did A, B and G. None of the headings felt that they were really matched and the right thing for those paragraphs. So, I was struggling with those. If it was for me, I would not have written these headings.

According to some L1 test takers, some headings were not the best headings for the paragraphs and at best they could be a brief summary of a few sentences in the paragraph. In fact, they found the right heading as incomplete and inappropriate for the paragraph. As one of the test takers said, "*Had they asked me to write a heading for these paragraphs, I would have come up with none of these options.*" (Broo) They found the heading option incomplete and unclear to match a paragraph. Some also criticized the heading options on the ground that they matched more than one paragraph, being too abstract and very different from the way headings are used in a typical text. One of the L1 test takers commented on the quality of the paragraphs;

Also, I should say that some of the paragraphs included more than one idea. B and C were both talking about like what was going on with the environment and what can be done. It also mentioned different ideas why it happened, so they have some common things at the same time some of the ideas in them are completely different. (Rich)

Yet another test taker said, “*The paragraphs were not distinct to me. I had a difficult time to see how they are different from one another. Some ideas were repeated in different paragraphs.*” (Kyle)

According to some L1 test takers a heading should be explicit and provide a summary and gist of the paragraph or section, while the heading options in this task did not serve this function of the heading were more difficult and confusing than the text itself.

The less successful L2 test takers had problems with comprehension of the paragraphs. They were unable to put the sentences together and make sense of the whole paragraph because they found too many new words in the text which prohibited them from getting the gist of the paragraph. For less successful L2 test takers, vocabulary was the main challenge in understanding the gist of the paragraph and the details presented. To make up for their vocabulary deficiency, they tried to skip the unfamiliar words and sentences that were difficult to comprehend. Some also used the headings as clues that could help them comprehend the paragraphs. For some paragraphs the headings were so confusing that one test taker focused on the heading and made it a starting point for understanding the paragraphs. In fact, she tried to choose a paragraph for the heading options.

Developing a textual or mental representation was part of L1 readers’ reading processes which helped them use it as a resource in choosing the right heading. Development of the text or mental representation was indicated by the participants’ recall of some ideas from different parts of the text while going over the headings. Some headings reminded them of different ideas discussed in different paragraph. Using the information recalled, helped them choose the answer after double checking with the text. For L1 test takers, test performance was a two-phase process. First reading to develop an overall understanding of the text and the gist of the text while the second reading was more focused on details and specific features of the paragraph that could pinpoint the answer and choose the right heading and not the other way around.

Another observation made in the test performance of all test takers for the *Matching Headings Task* indicated uncertainty in choosing from among two or three options. Most participants delayed the first and second paragraphs to get more information and more background. Since they found these paragraphs especially the first paragraph problematic for not having a focus and a topic sentence and presenting two or more ideas in one paragraph, they

frequently used “delay” strategy or changed the initial answer. They changed their answers as they proceeded reading more paragraphs and cross compared them with other paragraphs.

Test taker also used other strategies in answering the headings; 1) tagging each paragraph with a phrase, 2) attention to the repetition of inter-related words (lexical cohesion), 3) using lexical clues in the headings, 4) attention to the topic sentence and concluding sentences, 5) narrowing down the options, 6) using the headings to comprehend the paragraphs, 7) re-reading the last sentence in each paragraph, and 8) delay answering. For instance, one of the L1 test takers commented,

To get a general idea of the paragraph I often knew that the beginning and the last sentence are kind of really important in knowing what this paragraph is about and the concluding sentence at the end. (Cour)

She used “attention to the topic sentence and concluding sentences” strategy. Or, one of the less successful L2 test takers who delayed answering the first three paragraphs reported,

Then I looked back at the headings, but I could not get an answer for it. I continued reading the other paragraphs and the first paragraph that I could answer was D. (Hash)

In terms of difficulty of the task, results showed that most paragraphs were challenging since they did not have an explicit topic sentence that provides the focus and gist of the paragraph. Except for paragraph C which had some direct clues and the main theme of the paragraph was expressed in the first sentence, the other paragraphs were non-conventional and did not have a solid and explicit topic sentence. This contributed to the difficulty of the paragraph which required more time and more processing. Another factor that made some of the paragraphs difficult was related to the amount of processing involved. Some paragraphs were easy to process and answer. They were straightforward to answer like paragraph G, while other paragraphs such as paragraph A were more challenging. Yet, another reason for difficulty of some paragraphs depended on the degree to which the heading reflected a summary of the whole paragraph or just part of the paragraph. Heading options that were true gist of the whole paragraph were easier to process. Finally, the lexical clues in the heading and the degree they matched the text contributed to the difficulty of the *Matching Headings Task*.

What Does the Matching Headings Task measure? The main skills tested by the *Matching Headings Task* is mainly inferential comprehension of both the main idea and details at paragraph level. Only few tasks required literal comprehension. The task involved skimming and careful reading. Vocabulary knowledge was key to getting the gist of each paragraph. Vocabulary knowledge was also helpful in attending to the details of the paragraph and the lexical clues in the text and the headings. Additionally, text structure knowledge was conducive in relating the heading option to different paragraphs in the text.

6.2.5. Construct of the Summary Completion Task 1

The summary completion task asked for choosing some words from the text to fill in four blanks in the summary text which was paraphrase of four loosely related and randomly selected sentences from two paragraphs. The language used in the summary and the text was essentially the same with some minor replacement of a phrase from the beginning of a sentence to its end. All the test takers had to do was filling the blanks with words from the text. As shown in Figure 6.7, the main pattern used for the task was quite straight forward. Most test takers began by highlighting some key words/phrases before and after each blank, Next they scanned them in the main text and located the relevant information. Last, they carefully read the relevant information and found the exact word for each blank. in their scanning for the relevant information, test takers first scanned the proper name “Diamond” which was part of the summary heading. This helped them narrow down their search to specific paragraph in the text.

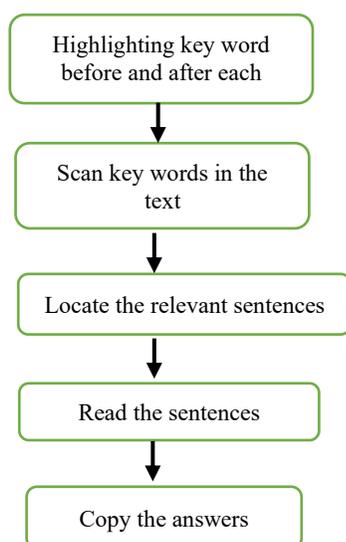


Figure 6.7. Pattern of test performance for the *Summary Completion Task 1*

Some L1 test takers and the more successful L2 test takers recalled the relevant information from the text and made reference to their previous reading. They could recall information from their first reading because they had read the whole text before they started doing the summary items.

Other test takers went back to the text and searched for the relevant information. They could locate the relevant info because they had read the whole text before or had skimmed it. The first reading or skimming of the text helped them in locating the relevant section and information.

An interesting distinction observed between the L1 test takers and the L2 test takers was that the L1 test takers searched the text for similar content and ideas while the L2 test takers searched for similar language and words. What the L1 test takers looked for was meaning while the L2 test takers focused on language and were not concerned with what ideas are being discussed. They did not explicitly notice or at least did not mention- ideas in their description of doing the task. This can be an indication of form-based approach to task performance versus meaning-based approach. What stood out for the L1 test takers was the ideas mentioned in the summary while for the L2 test takers it was the words and phrases that embodied the ideas that drew their attention. As one L1 test taker put it,

I just went through the paragraphs to find the key ideas so that I could match up the ideas with the words. I had already read the paragraphs for the previous task, so I was just kind of skimming through for the ideas that were mentioned in the summary. (Cour).

Her focus was on matching ideas rather than words. Another L1 test taker read the whole summary paragraph before she went go back to the text to search for answers. She said,

I kind of read for the idea and then because it was not word for word obviously, so I had to read for the idea in the summary and read for idea in the main text to find the word I was searching for. (Isab).

The less successful L2 test takers, on the other hand, had their attention on finding the same words and phrases in the text. One less successful L2 participant said, "*I had some rough*

idea where he (Diamond) was discussed in the passage. I read the relevant part to find a word that matched the sentence.” (Abed)

One strategy that was more frequently used in the summary completion task I was “noticing grammar”. Test takers were careful what kind of word in terms of parts of speech is needed to fill the gap. Other strategies were randomly used but their frequencies showed not association with the test task.

One issue that was frequently brought up by the test takers in different test tasks including the *Summary Completion Task 1*, was time pressure. Most L2 test takers felt time pressure. They claimed due to time pressure they could not do what they wanted to do. “Time pressure” was an issue that was frequently brought up by the L2 test takers especially the less successful L2 test takers in their verbal reports. One of the them commented,

I could not find an answer from the passage and I think maybe I should have also read other parts of the passage which I did not. I was stressed out not to spend too much time on reading the whole passage to find the relevant information. I did not know if this is a summary of the whole passage or just part or parts of the passage. (Alip)

This was a common challenge for most less successful test takers across tasks. They spent much time searching for the relevant info while for the L2 and the successful L2 test takers locating the relevant info did not take much time and processing. They used their first reading and the text presentation they had developed to locate the relevant info very quickly.

In brief, the *Summary Completion Task 1* tapped literal comprehension of few paraphrased sentences by asking test takers to fill few blanks with words from the text. The main challenge was locating the relevant info in the text. Since test takers were aware the same language was used in the task and the text, they searched for exact word in the main text. Even the less successful L2 test takers were comfortable with the task. They quickly highlighted and scanned key words of the summary sentences from the text.

As mentioned earlier, the pattern used in achieving the task included highlighting key words before and after each blank, scanning them in the text and reading the relevant information in the text and getting the answer. The process involved in reading the summary was in sharp contrast with the way a short summary paragraph is normally read. A short summary is normally read as a whole to develop and overall idea of the main points of the paragraph, but the summary

text was read sentence by sentence not as a whole. Unlike the L1 test takers, the L2 test takers finished answering one sentence before they moved to the next sentence. As the summary text was not a summary of the whole text or even part of the text, test takers preferred to do one sentence at a time. Task performance did not involve normal reading where a whole text or paragraph is read to get the main points made. The paragraph was broken down into smaller units, one sentence at a time then each sentence was read independent of other sentences.

What is measured by the *Summary Completion Task 1*? *The Summary Completion Task 1* tapped literal comprehension of specific details and vocabulary knowledge at inter-sentential and paragraph levels. The language used in the summary text was a copy of the same language used in the text with some minor paraphrasing. The task looked like a sentence matching task, where one sentence is in the summary and the other is in the text. Test takers scanned, skimmed, and carefully read the text to copy some exact words from the text into the summary task.

6.2.6. Construct of the Multiple-Choice (two- answers) Task

The *Multiple-Choice (two answers) Task* asked for two points of disagreement between the scholars mentioned in the text. As these points of disagreements were discussed in almost all the paragraphs, test takers had to read all these paragraphs and pay due attention to the details of the disagreements mentioned. Test performance involved comparison and contrast of the ideas suggested by the scholars in different paragraphs of the text.

Most test takers especially the L1 test takers chose the answers by recalling information from previous reading. They read the question and guessed the answer based on their overall understanding of the text which they had developed during the first reading. For L1 test takers, their first reading of the whole text was the main source for answering which shows they had developed a text representation. Interestingly, for these few items most test takers did not need to read the whole text and just relied on their understanding of the text which they had read for other test items.

Then, they double checked the answers with the text by skimming the paragraphs just to make sure the guesses are correct. Re-reading was just for confirmation of their guess. In many cases the test takers guessed one answer but had to go back to the text to search for the second answer. One of the L1 test takers best described his performance,

I read the options, B how the moai was transported they did not think they caused the collapse. They thought other environmental factors caused the collapse I had the information from the text, and I did not check the answers. I remembered from the text. (Mite)

This key pattern was observed in the performance of the L1 test takers while the L2 test takers mostly searched for the relevant paragraphs and information and carefully read them.

An interesting observation in the task performance of most test takers showed that some of them had a rough view of one point of disagreement but not the other. So, they went back to the text to find out the other one. For L1 test takers checking with the text was basically to double check the guess not to get the answer. One of the L1 test takers commented

I went back to make sure that I did not accidentally miss this point or another. they did not talk about the period when moai created. They did not talk about the origins of the people or how the moai were actually carved. (Broo)

Some less successful L2 test takers recalled part of the info (for one scholar) that was needed to answer the question, but they had to get the other answer from the text.

I was sure about answer B but I was uncertain if C or E is the next answer. I thought C is the answer. So, I went back to the text to find what the answer is. (Khei)

The less successful L2 test takers faced serious problems in locating the relevant info. They searched three paragraphs for locating the relevant information, yet they were confused what the answer might be. They also encountered many new words in the text which prohibited them from having a clear understanding of the text and struggled with understanding the ideas in the text. As they had not read the whole text in its entirety, they had not developed a text representation to use for choosing the answers. Failing to get the info required to answer the items, they randomly chose to guess one of the options as their last resort to do the task, but they could neither support nor reject their choice because their choice was based on their incomplete and vague comprehension of the text which they had got while doing the other test tasks and furthering reading of the text. So, they appealed to guessing. It was not an option for them. Most of the guesses they made were uncertain rough guessing rather than informed guessing. They

used it as the last resort to answer the items while for the L1 test takers and some of the more successful L2 test takers guessing was informed and the best choice to answer the items. They double checked their guesses to make sure they are right and in most cases they proved right.

Test takers used different strategies in their test performance. The main strategy was answering the task by “narrowing down and elimination of the options” which required development of a text representation wherein ideas of each scholar and the evidence they provided would represent one of the main points of the text. Another strategy used was “detailed analysis of the options”. The argument they put forward to eliminate the distractors showed that they had developed a text representation. They used the representation as a point of reference to compare with what the distractor said. One of the L1 test takers commented,

So, with that in mind I looked at the options I had and there was nothing in the text really that discussed the belief of the islanders and blame so I moved on from that and crossed it out. I just knew there was nothing to say about any of that in the passage. Then continuing to read. I found nothing that related to the that discusses the contradicting the points that Hunt and Lepo had with Diamond. So, I crossed it out. (Kyle)

Using the same strategy, another test taker said,

So, I read the options before I go to the text. I noticed moai is repeated in all the option which indicates the focus is on moai or the islanders. Role of the moai in the destruction of the island was something I had in mind. (Jaha)

Another strategy used by the test takers was “guessing”. In rejecting or supporting some of the options, test takers frequently referred to different ideas and details that were discussed in the text. Marta, one of the L1 test takers said,

I had a feeling the answers are B and C so I did not check the other options. I said let's check my guess, B and C and it worked. B C was the first thing that jumped out of me, so I got the answer from my previous reading. I just skimmed some sentences in these paragraphs. This is what it is. This is what they say.

She used guessing which she then double checked with the text.

In summary, *the Multiple-Choice (two answers) Task* performance did not involve much search for answer. For the L1 test takers, the task was mainly answered by using the text representation which they had developed from the first reading was the main source for answering the items. The second was mainly for double checking the guess they made for answering the task. The less successful L2 test takers, on the other hand, had rough guesses for the answer because they had a vague and incomplete understanding of the text. The L2 test takers used the process of elimination in choosing the right options more frequently than the L1 test takers unless necessary. The more successful L2 test takers could easily eliminate the distractors, but the less successful L2 test takers did not have the resources needed to examine and eliminate the options. L1s did not feel they need to eliminate the distractors. Finally, the more successful L2 test takers used either of these patterns. Those who scored higher on the whole test tended to do as L1 test takers did while those who scored lower tended to do use the same pattern of the less successful L2 test taker. The only difference was that they used the processes more efficiently and did not struggle doing the task.

What is measured by the *Multiple-Choice (two answers) Task*?

The *Multiple-Choice (two answers) Task* asked for two points (out of five) of disagreement between the scholars mentioned in the text. As these points of disagreements were discussed in almost all the paragraphs, test takers had to read all these paragraphs and pay due attention to the details of the disagreements mentioned. Test takers had to read almost all the text and compare and contrast of the ideas suggested by the scholars in different paragraphs of the text.

6.2.7. Construct of the *Multiple-Choice Task*

The Multiple-Choice Task involved both literal and inferential comprehension. Some of the items were facilitated by direct reference to the exact paragraph where the relevant info could be located. In doing the *Multiple-Choice Task*, as shown in Figure 6.8, two patterns were used in task performance. In the first pattern, test takers (mostly the L2 test takers) read the test items and highlighted some key words/phrases in the test item, went back to the text and scanned them in the text to locate the relevant information. Search for the relevant information was guided by the key words in the question. Next, they carefully read the relevant information and chose the right option. This pattern was more or less the same across all groups of test takers. In the second pattern which was mostly used by the L1 test takers and few successful L2 test takers, task

processes revolved around recalling information from the first reading and guessing an answer. The guess was then double checked with the text to make sure it is the right answer. Their second reading, in fact, served them to double check their guesses. As one L1 test taker said,

The second reading was just about finding the key words for making sure I give the right answer. The second reading was about the answer, based on just what I had first read. (Court).

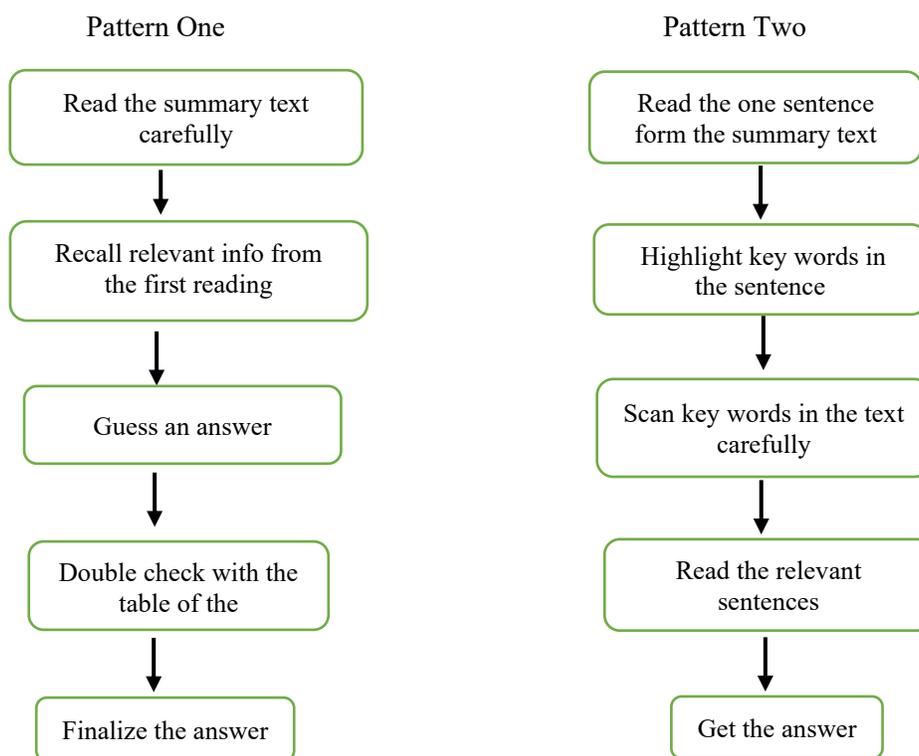


Figure 6.8. Patterns of test performance in The Summary Completion Task 2

However, variant forms of these patterns were observed. One L1 test taker who had started by reading the test items first, recalled the questions as she was reading the text. She could answer two of the task items by keeping the questions in mind while she was reading the text for the first time. This can indicate higher working memory capacity of the test taker to process both the text and the questions in parallel. Evidence of such parallel processing was not observed in L2 test takers. Yet another L1 test taker who was also very critical of the text for not

being well organized just relied on his mental representation of the text and memory without re-reading the text. He used the recalled information and eliminated some options.

Most test takers especially the L2 test takers also used eliminating strategy to narrow down the options.

Due to some limitations you do not go to read the rest of the options, but sometimes when you cannot find the answer you go for elimination of the wrong answers. I tried to see if there is a vivid answer first otherwise, I would go through elimination process. (Pari)

Some of the less successful L2 test takers failed to comprehend the text. So, they tried to find the answer by eliminating of the options. They could eliminate only one option based on their comprehension of the text, and they had to go back to the text to get the other option. In fact, they used both the relevant info in the text and elimination of options to arrive at the right answer. However, they were not very successful in eliminating the options because they did not have enough clue from the text to do so. This was a clear indication of their inability to fully comprehend the text and the relevant paragraph and info. L1 and more successful L2 test takers did not frequently used elimination of the options but when they used it, they were more successful in eliminating all the options. Eliminating the options was less frequently use by the L1 test takers which can be due to the more certainty of the answers they had chosen while for the L2 test takers, they still needed to make sure their choice is verified by the negative evidence they can collect from the distractors. L1 test takers used elimination strategy only when needed. They used it strategically. Some L1 test taker did not use elimination of the options because she found the answer was just there in the text and felt no need to eliminate other options. She was certain that her answer is the right answer. When one of the test takers asked why he did not use elimination process, he said,

The items were relatively easy. There were specific names that I could skim through and easily find them. Also, the task mentions where to look for it, in the 2nd paragraph, in the 5th paragraph and so on. I did not go through the process of elimination because the questions are asking me something that is easy to find. I can find the paragraph and the specific information and make sure what the answer is. (Mart)

In fact, for some of the L1 test takers, the *Multiple-Choice Task* was mostly a matter of matching the information in the text with the same information in the text. Therefore, elimination sounded irrelevant. One of the L1 test takers said,

This sentence matched the information with the right answer. I found the information in the text matched with the same information in the options, So I did not go for processes of elimination.” (Hele)

Two of the test items in the *Multiple-Choice Task* mentioned the paragraph in which the relevant information presented. For these items test takers especially the less successful L2 test takers could do the items more comfortably. Because they did not need to search several paragraphs for the relevant information. Another interesting feature observed in the test performance of the L1 test takers was their awareness of the paragraph structure. For example, three test items asked for results of the research studies reported in the text. The L1 test takers were aware of the paragraph structure and they mentioned that they were aware that results of the studies were the target of the questions and they were aware that results come at the end of the paragraph. One of the L1 test takers just skipped reading the whole paragraph and directly went to the last sentence which reported the results of the research. This awareness of text type and paragraph structure was not observed in the L2 test takers.

Another process that distinguished the L1 test takers from the L2 test takers was the frequent references the L1 test takers made to the text representation they had developed in their first reading of the whole text and mentioned. For example, in explaining how they arrived at the answer for the *Multiple-Choice Task* they could clearly indicate that option B or D was not discussed in the text or C was completely wrong because the text did not talk about this aspect of the topic. For example, one L1 test taker said,

For number 30, I crossed out B and D mentally first because the text does not look at the abstract art rather it talks about the mentality of appreciating art. So, B was completely wrong. D is about public opinion which is not mentioned in the text at all so D is also out. A was true but I went for C because it was a better choice. It talks about representational art down here in the text. It talks about both of them and it does kind of comparison near the end. That is why I ended up going for C. (Rich)

This account clearly shows that the test takers had developed a text/mental representation which helped them eliminate some distractors.

One more feature observed in the test takers' account of their test performance the reference they made to the use of some key sentences and phrases in the text that triggered the answer. Many test takers mentioned certain sentences as key in deciding the choice of answer. However, the way L1 test takers and L2 test takers mentioned the use of the sentence differed in a meaningful way. The L1 test takers focused on the idea presented in the sentence with no mention of formal aspects and the lexical clues in the sentence. They focused on the meaning and idea of what they read and justified the answer for having the same meaning or saying the same thing while the L2 test takers relied on form-related clues and aspects of the sentence. For instance, one L1 test taker said,

The paragraph talks about people and if they spend a longer time trying to figure out what the work of art is, they found it more satisfying so I went to the first option and it was not that, so I found it in the next option B which matched what the paragraph said. (Nabi).

So as indicated by this test taker, what she focused on was mostly the meaning expressed rather than the language while an L2 test taker had a different description for the same item,

I read the options, option B, (which reads) "to find it satisfying to work out what a painting represents" was the answer. The words rewarding and satisfying were so close in meaning. I was sure it is the correct answer. (Horr)

Based on the processes and strategies test takers used, it can be argued that for L1 test takers task performance processes seemed to be more meaning-based while for the L2 test takers the process was mostly form-based and linguistic elements played a more vital role in their processing. They were more aware of lexico-grammatical features that could guide them in choosing the right answer.

It is worth noting that the last item of the *Multiple-Choice Task* differed from other items. It asked for a sub-title for the text which required comprehension of the gist of the whole text. This item was answered based on the test takers' overall understanding of the gist of the text. Interestingly, all the more successful L2 test takers got the answer right while 30% of the L1 test

takers and 50% of the less successful L2 test takers missed the correct answer to this item, suggesting the task had high discrimination power.

In terms of strategy use, two strategies were more frequently used in *the Multiple-Choice Task*, “eliminating/ narrowing down the options” and “developing the gist of the text” which were moderately and weakly associated with the task, respectively. The first strategy had to do with the nature of the *Multiple-Choice Task* which included four options. Other tasks did not have any room for use of such strategy. However, developing the gist of the text is a more general strategy that can be adopted for other tasks too.

In summary, two main patterns were observed in the test performance of the test takers; the first one mostly used by the L1 test takers and few of the successful L2 test takers and the second pattern mostly used by some successful L2 test takers and all the less successful L2 test takers. In the first pattern, the L1 test takers used the text representation they had developed through the first reading the text and recalled the relevant information to answer the item. They guessed an answer and then they went back to the text and double checked their guesses. However, for some items they could not recall any information, or they were totally uncertain, so they went back to the text to get the answer. For them the second reading involved skimming, scanning, and careful reading of the relevant information. In the second pattern the less successful L2 test takers could not recall the relevant information and had to get all the answers from the text. So, they first highlighted the proper names or key words/phrases in the questions which they scanned in the text and located the relevant information. Finally, they read the relevant information which was mostly few sentences and got the answer.

What does the *Multiple-Choice Task* measure? Test performance of the test takers indicated that *the Multiple-Choice Task* measured specific details and the main idea at inter-sentential and paragraph levels. Both types of comprehension literal and inferential are tapped by the *Multiple-Choice Task*. Vocabulary knowledge and grammar knowledge can help test performance. The test task involves different reading skills such as skimming, scanning and careful reading.

6.2.8. Construct of the Summary Completion Task 2

The *Summary Completion Task 2* was very straight forward for all test takers. It asked test takers to fill in four blanks in a summary text with words from a table of words. The

summary was a short paragraph of few sentence which involved comprehension and matching sentence in the summary text with the same or similar sentence in the text.

Two main patterns emerged from the test performance of the test takers (Figure 6.9). In the first pattern, most L1 test takers and some of the successful L2 test takers who had already read the whole text used the text representation they had developed and recalled relevant information. They guessed what the answers might be and used the table of words to double check if their answers are right. Most of them did not go back to the text to double check the answers. In fact, they relied on the information they recalled from their previous reading and most often they filled most blanks just by looking at the table of options and did the task without going back to the text or re-reading the text. One of the L1 test takers reported,

So, for this one I just read the summary and filled it in. I actually had some answers before I look into the table of words. I did not need to go back to the text. I had the ideas in mind as I read through the summary. Maybe, if the options were not given, I would have checked one of the answers to make sure it is the right answer. I could also see that the summary is on different parts of the main text. (Rich)

They had no problem with the task. The successful L2 test takers were quick and efficient in doing the task correctly. In case they needed to get the information from the text, they did it quite easily and quickly. They just scanned the text for the names or key words mentioned in the summary and located the relevant information. They looked for matching the same information and finding words that made sense in the blanks. The short amount of time spent on the task is a clear indication of how easy they could do the task. In the second pattern which was mostly used by the less successful L2 test takers and some of the successful L2 test takers, they first highlighted some key words from the summary and went back to the text to scan them and read the relevant information and get the answer. One of the L1 test takers best summarized this pattern,

... and then I began reading the second sentence, reading up to the second blank and it talked about Alex and I scanned the text for his name and found it in the first paragraph in the first page and so I read the paragraph having in mind what the question was asking

about which was the precise degree of, and I found about functions and motives so I chose B which was complexity. (Isab)

The less successful L2 test takers struggled with relating the summary to the text and locating the relevant information.

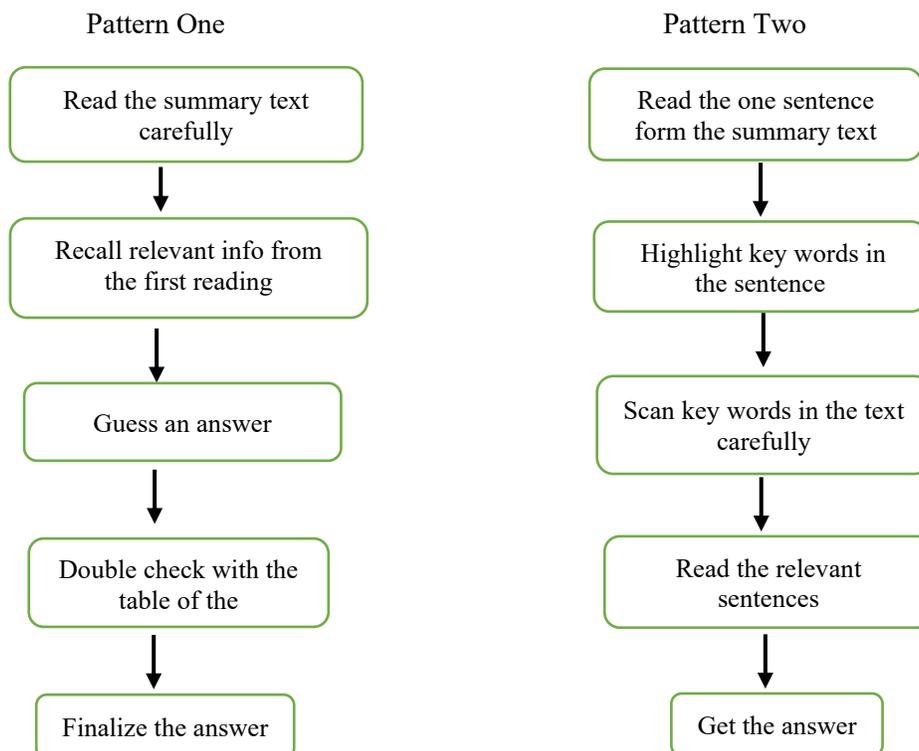


Figure 6.9. Patterns of test performance in the *Summary Completion Task 2*

Variant forms of these patterns were observed in the test performance of the test takers. For the first pattern, one L1 test taker could recall some information about one blank, but he had to go to the text for the other two blanks. So, L1 test takers used a mix of two resources to answer the items; recalling information and using the source text. Another L1 test taker read the summary and filled in all the blanks based on his memory. He double checked his answers with words in the table but not the text. Yet another L1 test taker read the summary carefully, answered one item by recalling information from previous reading and guessed the two other answers which she then scanned for in the text to get the answer. The L1 test takers used two patterns; in the first pattern they recalled information from the previous reading and just double checked it with the text and/or the table of words. Double checking the guess was

straightforward and easy. They could quickly locate the relevant information by scanning key words before each blank. One of the L1 speakers said,

First, I read the little (summary) text in the question to know what part of the text to compare with or to find the answers and it talks about impressionist paintings and I know it is in the first paragraph so I knew where I should go. (Hele)

Another test taker commented,

I was looking for the name Alex and could quickly locate it. I had read the text and I remember where it was talked about which is down here, so I had to just look for her name and read the text so that I find the answer. When you know it is in this paragraph you can find the answer. (Nabi)

In the second pattern, test takers had a rough guess and searched for an answer from the text. So, they scanned the text for some key words from the task and carefully read the relevant information to find the answer.

Most less successful L2 test takers searched the text for answer because they could not recall any information from previous reading, nor could they guess and answer. Both groups, read the summary text carefully especially the words and phrases preceding and following the blanks. One test taker commented, *“Based on the words before and after the blank I could find the word “complexity both here and there.”* (Abed)

They highlighted some proper names and key words and phrases in the summary text which they later scanned in the text. They looked for exact same words in the text. Also, time pressure forced them not to waste their time guessing an answer and preferred to rely on the text for their answer. For the L1 test takers, locating the relevant section and information was easy and quick because they had read the same section for the previous tasks. Previous reading allowed them to locate the relevant inf very quickly and easily.

Variant forms of Pattern two were observed in the test performance of the less successful L2 test takers who searched for an answer in the text. Only one of them answered two of the items based on recall. She used the table of words to save some time. Two of them used table of options to double check their guesses. After reading the relevant information, they used the table

of options to make sure the word they had selected as an answer is there. One of the less successful L2 test takers used test-wiseness strategy and lexical clues in the summary text to search for an answer in the text. He avoided guessing simply because he had problems with comprehending the text. He said,

I had no clear idea of the text so I could not rely on my guesses. Then I went back to the previous blank, I thought it should be related to paragraph 1 because there was no question related to this paragraph in other tasks. Knowing that the task is basically a copy of the original text, I focused on “for example” I searched the text for the phrase “for example” in the paragraph and then the other words and I could find the word that fits the blank. (Hash)

Finally, another less successful L2 test taker guessed three items, scanned key words and double checked the guesses in the text. Only one of them proved right. Then she went back to the text to get the answer. In brief, they mostly got their answers from the text after they read the relevant sentences in the text.

In sum, some differences were observed between the performance of the L1 and L2 test takers. First, L2 test takers looked at the text to double check their answer while for the L1 test takers the text was used for getting the answer. Second, recalling information from previous reading was mostly used by the L1 test takers and some of the more successful L2 test takers. Third, the table was not used by the less successful L2 test takers because they had no clue what the answer might be and needed to get the answer after reading the text. Finally, most of the L1 test takers carefully read the whole summary paragraph before answering the questions but this was not observed in the L2 test takers who read each sentence separately and did not go to the next sentence before they finished answering the previous sentence.

The only strategy that was more frequently used in the *Summary Completion Task 1* was noticing grammar where test takers considered parts of speech needed to fill the blank. However, its frequency was very low and had very weak association with the *Summary Completion Task*

What does the *Summary Completion Task 2* measure? Based on the content analysis of the test and test performance of the test takers, the *Summary Completion Task 2* measured vocabulary knowledge and awareness of lexical clues in the context of sentence and inter-

sentence level of comprehension. The task also involved reading at paragraph level, scanning, and careful reading.

6.2.9. Construct of the *Yes, No, Not given Task*

The Yes, No, Not given Task included six statements that had to be judged against the text. The statements were partially literal and partly inferential and evaluative. The statements were all related to three last paragraphs of the text. The main processes used in the *Yes, No, Not give Task* included reading the statement in the task, locating and reading the specific sentences or paragraph that contained the relevant info, matching some features such as phrases and words from the text to the statement, analyzing the specific meaning expressed, using some background knowledge when necessary and inferencing the answer. These processes were mostly used by the L1 test takers and some successful test L2 takers. The items could not be answered without specific attention to some very specific details that were expressed in words and phrases embedded in few sentences. Those test takers who could attend these lexical features and clues had no chances to answer the items correctly. One of the L2 successful test takers reported,

(For question) 35, I looked for “*mirror neurons*” I search the text and found it in paragraph 8. The question is about *further verification*. This was key in the question. I am confident the answer is yes because the passage says, “*it needs to be thoroughly tested*” again this phrase and the word *verification* are kind of synonym. I could easily answer this one.

(Jaha)

She indicated how her good vocabulary knowledge and attending to lexical clues in the text could help her get the answers. Another successful L2 test takers commented, “*For item 37 the question says it is easier while the text says they are different not easier. So, I chose not given.*” (Angl) In this particular item, the test takers could not get to the right answer, had she not noticed the words (easier-different) and the contrast between them.

The L1 test takers and some of the more successful L2 test takers used the text structure knowledge to locate the relevant section or paragraph. With a text representation in mind, they could recall information or go back to the text and locate the specific relevant information they thought they need for answering the question. One of the L1 test takers commented,

I read the statements one by one. So, for question 34, I answered it very quickly because it is talking about it in the last paragraph how abstract and representational art show signs of fractal repeated motifs recurring in different scales blah blah. This is what she is talking about. This is her view. This is why it is easy to answer no. She is not contradicting fractals. I looked back and checked in the text. I had some idea and I used the information in the text to answer the question. (Rich)

He clearly indicated that he had developed a text/mental representation and could recall the relevant information. Further he used the text to double check his answer not to get the answer. The evidence for the development of a text representation was presented by the test takers in their verbal report where they made frequent reference to different paragraphs and what each was discussing. In another example, Jack said,

Item 34 was the last item I did because I remembered Foresight. I knew that is somewhere at the bottom of the first page. I knew where that info was, so I knew I had to go back to the passage.” (Jack)

This clearly indicates that he was quite aware of what was being discussed in each part of the text.

Closer analysis of the L1 test takers’ performance showed that they noticed meaning of the ideas discussed in the statements and the text and not the language and wording used in the test items. For instance, one L1 test taker said,

I found the statement that clearly contradicts what is said in the question “it would be foolish to ... a set of scientific laws” which is a direct contradiction to what is said in the question, so I decided to choose No. (Kyle)

His focus was on the meaning of the statement and not the individual words and lexical clues. It is interesting that he talks about “contradiction” which is what we usually use for ideas not words. In another example the same test taker talks about his attempt to find an idea in the text that proves otherwise. “

For item 38 (which reads) “*Art appreciation should always involve taking into consideration the cultural context in which an artist worked*”. So initially I had an idea that

this information was not given but I decided to skim through to see if I could find anything that proved otherwise and I could not really find anything discussing the cultural context of the artist so I decided to choose Not given.”

Since the answers to *the Yes, No. Not given* items were not explicitly mentioned in the text, test takers had to infer them or use lexical inferencing to get them. The less successful L2 test takers could not infer the implied meaning and answered the items only based on their rough understanding of the text. This is why they were uncertain if their answers are right.

The more successful L2 test takers, on the other hand, used their vocabulary knowledge and noticed the lexical clues used. For instance, one of them said, “For 36, “*trend*” and “*fashion of the time*” have the same meaning and helped answer the question.” Another L2 test taker said,

For 38, I had to do with the specific language. It was using the term cultural context in the question even though it was not in the section that I was reading I could see the connection like they are talking about the artistic environment of the time, that kind of thing. So, I could see it as another way of phrasing the same idea. So, it is talking about the same thing not something different. (Mart)

She noticed paraphrasing and some language features used in the item and how the answer was implied in the text. This item involved inferential comprehension. Another successful L2 test taker talked about the lexical clues he used in the text to arrive at the answer.

For 37, I noticed key words in the statement, “*precise rules*” and “*art appreciation*”. I continued reading and “*It would, however, be foolish to reduce art appreciation to a set of scientific laws.*” gives the answer. We have synonyms *scientific laws* and *precise rules* and it says *it is foolish*, so the answer is No. (Hona)

The more successful L2 test takers were more aware that the items are asking for inferential meaning not literal comprehension. one of the L2 test takers said,

I knew the writer’s ideas are being questioned not the sentences. It is a kind of evaluating the ideas. It requires much more inferencing, more wholistic meaning and getting the intention of the writer not just what he has said. It is also more abstract. (Jahan)

She had read the whole text and had developed a textual representation which she could use in answering the items. The successful L2 test takers were also more test-wise. For instance, one of them could choose the right answer for two of the items because they included some absolute terms such as. “*precise*” and “*entirely*”. She said,

I did not go back to the text (for this item) because when they use a word like “entirely”, it is usually wrong. Also, Item 37, says “define precise rules”. I did not think it is precise. It is one of those words that make a sentence false. (Angl)

Some of the L2 test takers especially the less successful L2 test takers had a hard time locating the relevant section in the text. They had serious challenges in understanding both the statements and the paragraphs. Nor could they relate the statement to the text. Some new words in the text added to their challenge of comprehending the text.

I had two key words, Foresight and fractal from the question. I went back to the text for them. I found them in paragraph 7, but I had problem with this task because (the *Yes, No, Not given Task*) I have problem both with understanding the question and also the text. They add up and cause more challenges. I did not get the question. The topic is vague to me and the question itself is part of it. I understand the words there, but I cannot add them up and develop an understanding. I did not understand the question, so I looked for something in the text that kind of matches the first question. I was not sure what the answer is. (Abed)

He best expresses how the task is so difficult for him. He fails to understand the question and topic unfamiliarity adds to the challenge of getting the gist of the relevant paragraph. For the less successful L2 test takers, the task was tough. Another less successful L2 test taker had the same problems with the task. She said,

I did not understand the question, so I looked for something in the text that kind of matches the first question. I was not sure what the answer is. (Zade)

This illustrates the confusion they faced in comprehending the questions and relating it to part of the text that is relevant. Since the ideas expressed in the statement were not explicitly stated in the text, they failed to locate the relevant information, so they went back to the text to get an answer.

They answered the questions just randomly and with full uncertainty. They ended up not answering some questions or gave the wrong answer.

I could not get the last three questions to the text. They kind of looked hanging there. I could not locate any relevant information in the text. I think I should have inferred them from reading the text, but I could not. They were not much factual. I sensed a gap between what the questions say and what the text says but I could not address this gap. I should have read the last three or four paragraphs more carefully. (Beig)

She failed to comprehend the gist of the paragraph and since the items require inferencing, she could not easily and confidently infer any of the statements from the text. Nor could she spend more time on the task.

It is very interesting to note that two strategies most frequently used in processing the *Yes, No, Not given Task* included “delay answering” and “use of background knowledge” which speak volumes how difficult the task was. For some items, they could not get the answer directly from the text and had to include their background knowledge to infer the answers. One of the L1 test takers commented,

Based on that I said No because even though style of the time make style popular. There is style that is against the currents of time. There is style that stays for ages. That was the kind of thing that I was going for. So, looked for this in the text. (Rich)

This clearly shows he used his background knowledge for achieving the task.

In brief, test takers who did better on the task highlighted key words in the statements, scanned and searched for relevant info in the text, read the information carefully, and used lexical clues in the items and the text to correctly infer meaning and answer the questions. L1 test takers did not make as many references to the use of lexical clues as the successful L2 test takers did. They seemed to have relied on their comprehension ability and the textual representation to

get the answers. One L1 test taker reported, “*For 36, I looked at the second last paragraph for this item. I knew it is near the end of the text and easy to find.*” This clearly shows the test taker had developed a map of the text. The less successful L2 test takers, on the other hand, were confused and struggled to understand the statements and the text. Some of them could not even relate the questions to the relevant information in the text. This was mainly due to the inferential nature of the task.

What does the *Yes, No, Not given Task* measure? The test task measures test takers’ ability to comprehend implied inferential meaning at paragraph level. The task items are mostly inferential in nature. They involve either lexical inferencing or inferencing. The task measured high level processing of reading comprehension. Key to doing the task was noticing the lexical clues used in the text. Therefore, vocabulary knowledge was vital in task performance. The task involved reading, skimming, and scanning two paragraphs.

In conclusion, results of the SKSPs used by different groups of the test takers during test performance as reported in their retrospective verbal reports showed that each test task tapped into different components of the construct. Most test tasks shared some similarities in terms of the type reading involved (careful-expeditious), the level of processing (low-high), the specific skills measured, the type of knowledge used, etc. However, each test task was operationalized differently and involved use of different SKSPs which defined the its sub-construct. The details described in this section provided a clear picture how each test task contributed to the construct of the test which is dominated by literal comprehension of specific details at low levels of sentence and inter-sentence. Findings of the test takers’ verbal accounts of each test task also showed that IELTS RCM test construct alters across test takers. Depending on their language background (L1-L2) and their level of language proficiency (low-high) the SKSPs used for test performance differed.

6.3. Models of IELTS RCM Test Performance of L1 and L2 Test Takers

Based on the results of the SKSPs used by the test takers in their test performance two models of test performance for the IELTS RCM emerged. This section presents these two models.

As shown in Figure 6.10, reading comprehension in the context of the IELTS RCM context manifested itself in quite different patterns. For the L1 test takers, the texts were typically carefully read to develop an overall understanding and identify some of the details that were discussed in the text. All L1 test takers, on the other hand, read to comprehend before answering the test items. First and foremost, they focused on comprehending the text and developing a representation of the text. All the L1 speakers read the whole text as they would normally read their academic texts. Their focus was on reading and comprehending the whole text. Searching for answer to the test items was achieved in light of the text representation that had been already formed.

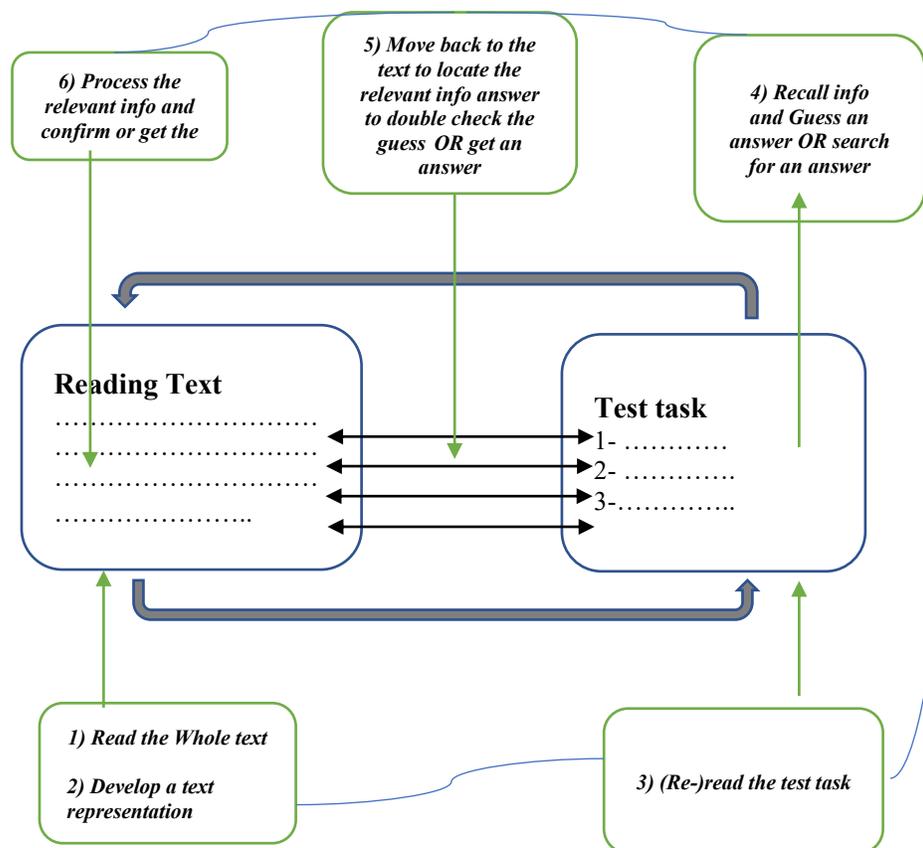


Figure 6.10. Model of IELTS RCM test performance of L1 test takers

As shown in Figure 6.11, the L2 test takers started with reading the few items in the first test task before reading the text. All they focused on was answering the test tasks rather than reading and comprehending the text. None of them read the whole text before going to the test tasks. They read just what they thought relevant to the item. They searched the text to locate the

relevant information required. Their reading was controlled by demands of the test tasks. Reading comprehension was not the purpose of reading. The purpose was to do the test tasks. Unlike the L1 test takers, for the successful L2 test takers reading was more strategic and selective. It was strategic reading because they did a lot more deliberate scanning and moving back and forth between the text and the test items instead of reading the whole text. These results suggested that the construct of reading comprehension operationalized by the IELTS RCM was not constant; rather, it changed from one group of test takers to another. In other words, the construct that was operationalized by the IELTS RCM was highly dependent on the characteristics of the reader/test taker. Readers with higher language proficiency processed the tasks differently.

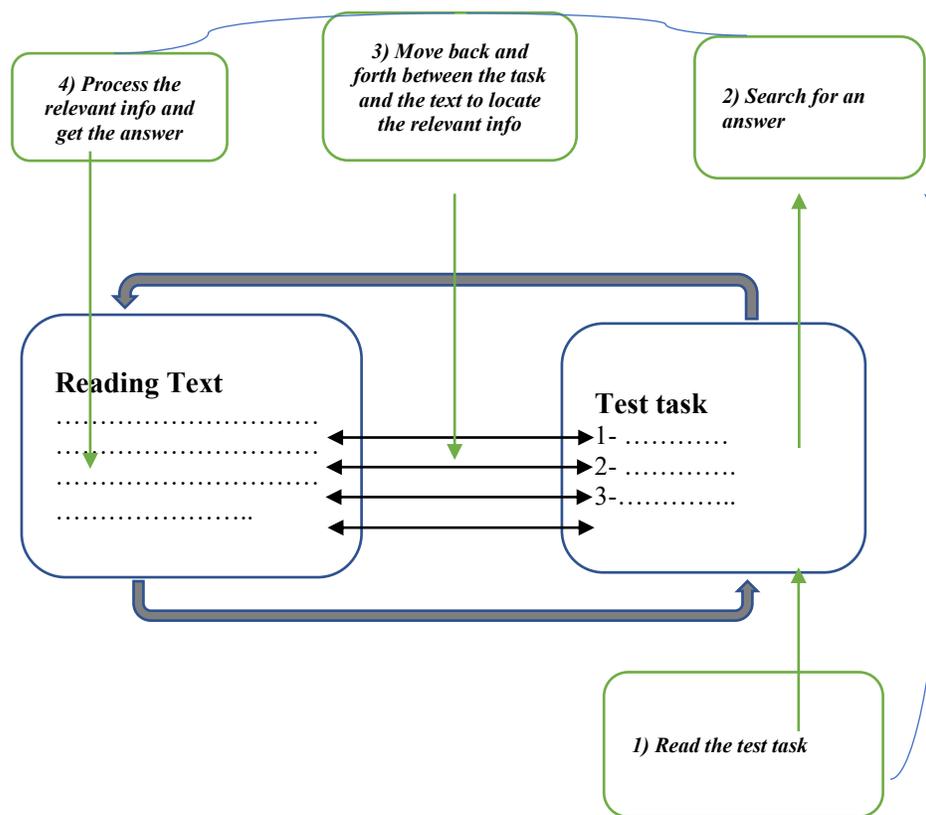


Figure 6.11. Model of IELTS RCM test performance of L2 test takers

6.4. Results of the First and the Second Cycle of Coding

This section reports on the results of coding of the test takers' accounts of the processes and strategies used in their performance (Research Question Two). Given that test task features

and item features influence reading processes and test performance (Bachman and Palmer, 2010; Grabe & Stoller, 2011), test takers may use a variety of skills, knowledge sources, processes and strategies (SKSPs) to answer each test task and test item. To explore and identify these processes and strategies used in the IELTS RCM test performance, two cycles of coding were conducted.

To begin with, as discussed in Chapter Four (See Section, 4.7.3.3) for coding the immediate retrospective verbal reports, Saldaña's (2009) coding manual was adopted which suggests different coding methods for each cycle of coding. The first cycle aims at capturing and summarizing the main features observed in the data while the second cycle focuses on grouping and categorizing the code and sub-coded to "*categories*" which help extract patterns in the data for developing themes and global understanding of the data. As each cycle of coding proceeds, codes are grouped into categories and categories into themes. Finally, themes generate a theory that accounts for the phenomenon under study.

As discussed in Chapter Four, results of the first cycle of coding produced a code book consisting of a long list of more than 72 codes (See Appendix K, L, and M) covering different cognitive dimensions (SKSPs) of test performance. Some of these codes were related to reading and processing the text and some to reading and processing the test tasks. Yet, there were some processes and strategies used to relate the text to the test item of the IELTS RCM. The focus of the second cycle of coding, on the other hand, was on integrating these initial codes into coherent categories. After much deliberation and consultation with the existing literature on models of reading comprehension especially, Khalifa and Weir's (2009) reading comprehension model, codes from the first cycle of coding were coded into some conceptual categories. At first, the processes and strategies used by the test takers seemed to be isolated activities, but they could be integrated into different categories. As coding proceeded, they were then fine-tuned for accuracy and consistency. For instance, codes that helped test takers locate the relevant information in the text were distinctly separate from codes that were related to test takers' effort to read and process the test tasks. Codes that shared some feature(s) in common were then integrated and categorized into separate categories. These categorizations and distinctions helped grouping the codes into more global categories and themes. In this section, results of first and second cycle of coding are presented in more details.

However, before reporting the results of the first and the second cycles of coding a few points need to be made about the terms used in naming and describing the codes and the

cognitive processes they represent. Terms such as *processes*, *strategies*, *categories*, *themes*, *metacognition*, *careful reading*, *expeditious reading*, etc. are subjective terms and open to different interpretations. Therefore, they need to be precisely defined. Readers are advised to look at the Glossary which provides an operational definition for these terms.

Based on the coding procedures described in Chapter Four, (Section 4.7), codes and sub-codes that emerged from first cycle of coding were closely studied and similarly coded data were grouped into categories. Three main themes emerged from the second cycle of coding each consisting of several codes and sub-codes; 1) Reading Theme, 2) Searching processes, and 3) Answering Theme (See Figure 6.12). These categories were all conceptual and were not directly mentioned in the retrospective verbal reports of the test takers. Each category accommodated and integrated a number of codes and sub-codes (processes and strategies) that had emerged from the first cycle of coding.

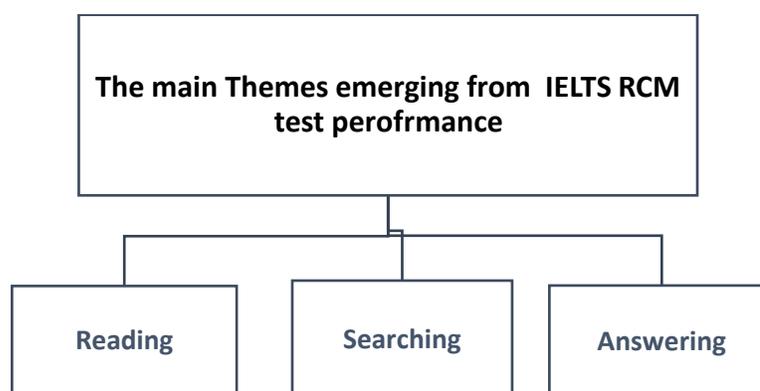


Figure 6.12. The main themes emerging from test takers' IELTS RCM test performance

6.4.1. Results of the “Reading Theme” in test performance

Codes from the first cycle of coding showed that test takers used different processes and strategies to manage each item and test task. Some test takers read the whole text or just parts of the text as required by the test item. Some read the test items first and then the text, while others read the text first. Some skimmed the whole text. Others identified type of test task first, before reading the part of the text to locate the relevant information. All test takers moved back and forth between the text and the test tasks to process the task, and so on. These processes and strategies were coded into “reading theme” helped test takers accomplish the test tasks. “Reading theme” consisted of several processes and strategies that were used to do the test task and moved the test performance forward. Without these reading processes and strategies test takers could not have answered the items on the test.

It is worth noting that some of these activities were specific to L1 test takers while some other activities were observed just in the performance of the more successful L2 test takers. The variations observed in the use of these activities across test takers can be very revealing about the nature of test performance and test construct. Figure 6. 13 presents the processes and strategies that comprised the “Reading Theme”.

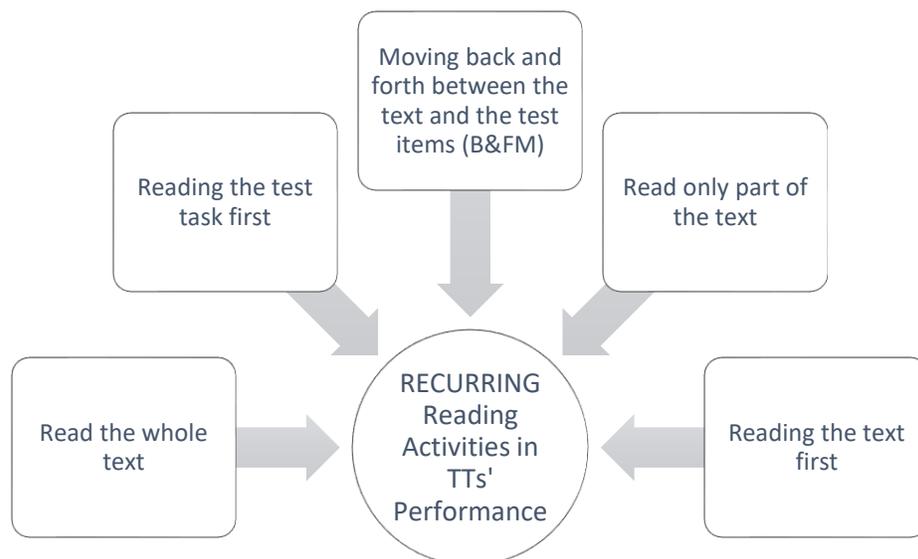


Figure 6.13. Some processes and strategies used in “Reading Theme”

To examine details of the “reading theme” in test performance, reading processes, strategies and activities that formed the theme were compared across test takers, Table 6.1 presents result of frequency of these processes and strategies for different groups of test takers.

Table 6.1

Results of recurring reading activities in test performance of different test takers

Reading Activities	L1	L2: More successful	L2: Less successful	Total
Move back and forth between the text and the task	10(100%)	5(100%)	6(100%)	21
Read the whole text carefully	10(100%)	1(20%)	0(0%)	12
Read the test task first	3(30%)	2(40%)	5(83%)	10
Identify task type first	2(20%)	3(60%)	1(17%)	6
Read the text first	4(40%)	0(0%)	1(17%)	5
Skim the whole text	0(0%)	2(40%)	0(0%)	2
Read few sentences from the text first	0(0%)	0(0%)	1(17%)	1

As shown in Table 6.1, different activities were used in test performance of the test takers. However, each group showed different patterns in their performance with some areas of overlap. In starting their test performance, L1 test takers paid equal attention to both the text and the test items. Four of them started by reading the text. Two other L1 test takers started by

identifying the type of test task by just looking at the test task but did not carefully read it. However, all L1 test takers (10/10) read the whole text carefully sentence by sentence before they started answering any item of the test. As most of them indicated they read the text normally as they read their academic texts. When they finished reading the whole text, as they later claimed, they had good grasp of the main points of the text and had developed a representation of the text organization. They knew the gist of the text and some relevant details. They were aware of what was discussed and where it was discussed. This was specific to L1 test takers and only two (2/10) L2 test takers who scored high in the test. They carefully read the whole text because they were confident enough, they can do it within the time limit of the test. They preferred to build a text representation before they start answering test items. In fact, every item they tried to answer later was done in light of text representation they had developed through careful reading of the whole text. As shown in Table 6.5, for L2 test takers, the reading theme was quite different. For example, the less successful L2 test takers all started with reading the test task and did not read the whole text at all. This single behavior can alter the nature of test construct for other test takers which will be discussed later.

The processes and strategies used by the test takers were put together to see what patterns might emerge from the sequence of their test performance. The patterns that emerged differed for each group of test takers. In this section, these three patterns are presented.

As Figure 6.14 shows, L1 test takers started with reading/skimming the first test items. Then they carefully read the whole text sentence by sentence. After they finished reading the whole text, they went back to the test items and read the test items one by one and tried to answer them by either 1) recalling information from the text, 2) guessing an answer which was double checked with the text, or 3) going back to the text and search for the relevant information and find the answer.

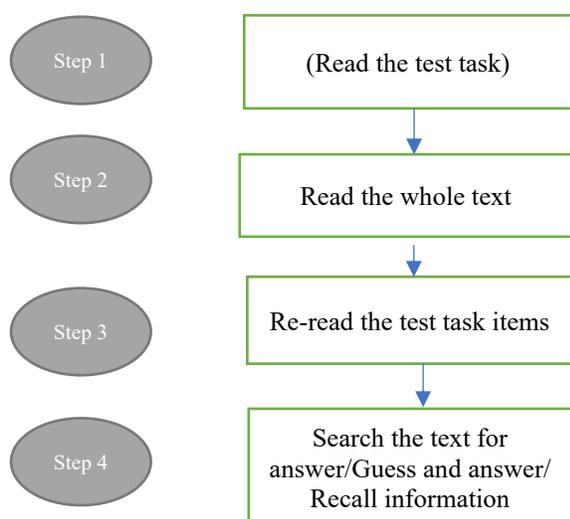


Figure 6.14. Processes and strategies used by the L1 test takers in reading theme

As shown in Figure 6.15 for the more successful L2 test takers, three different patterns of performance emerged. They used different processes and took different steps to do the test tasks. In their first step, two test takers (2/5) focused on the test items. First, they skimmed the first few test items to identify the type of task before reading the text. Then they read the whole text carefully and developed the gist of the text and the main points discussed. In the third step they went back to the test tasks and carefully read each item of the task. In the fourth step, they went back to the text to search for an answer or double check their guess. These test takers got the highest band score of 9 in their performance. This pattern was exactly the same as that of the L1 test takers.

In the second pattern, one test taker (1/5) started with careful reading of the test items and going back to the text to search for the relevant information and get the answer. She did not read the whole text, nor did she skim the whole text. She focused on doing the test task instead of reading and comprehending the text before answering the test items. She focused on searching and reading only that part of the text that seemed relevant to the test item. She did not read carefully because she worried that she might not be able to finish the tasks within the time limit of the test, as she claimed in her introspective verbal report.

The third pattern of performance which was used by two other successful L2 test takers (2/5) included the following steps. They first began with skimming the first test task to identify

the type of test task. Then in the second step, they skimmed the whole text by reading few sentences from each paragraph to get an idea what the text is all about. Next, they went back to the test task and read the items carefully one by one. In the fourth step, they went back to the text to locate the relevant information and get the answer. They justified skimming the text instead of careful reading for lack of enough time. They believed they could not read the whole text and answer the items within the time limit, so they chose to quickly skim the text instead of reading it carefully.

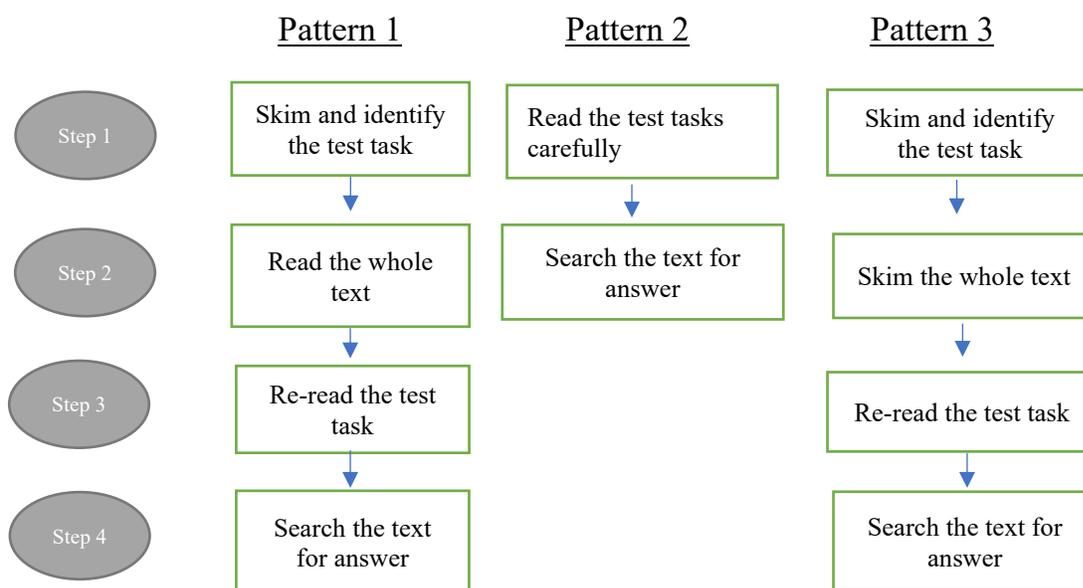


Figure 6.15. Processes and strategies used by the L2 test takers in reading theme

As shown in Figure 6.16, for the less successful L2 test takers only one pattern emerged. The pattern was very straightforward. They did not read the whole text, nor did they skim the text. They just read the test items one by one and searched for the relevant information in the text. They had no idea what the main points and arguments of the text were. They just focused on some key words and lexico-grammatical features of the test item and tried to locate them in the text and find the relevant information. As they had neither read the text nor skimmed it before, they had more challenges in locating and reading the relevant information. Using their test-wiseness skills, they know that the answer for the first few items is at the beginning of the text. Some of the less successful L2 test takers searched the first few paragraphs in the text to locate the relevant information for the first test task and looked for the relevant information of the last test tasks somewhere at the end of the text. This strategy helped them locate the relevant information in the text. When asked why they did not read or skim the whole text before doing

the task, they all said that they could not have finished the test tasks within the time limit assigned, had they chosen to carefully read the whole text or skim it.

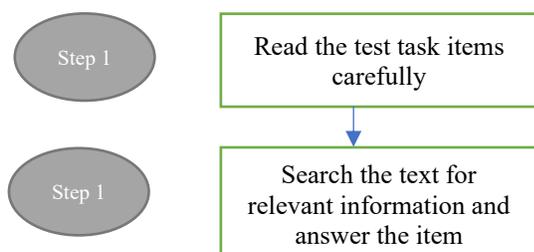


Figure 6.16. Processes and strategies used by the more successful L2 test takers in reading theme

In summary, the processes and strategies and reading activities of the test takers showed that depending on their language background and levels of language proficiency, test takers used different sets of processes, and strategies to respond to the test tasks. L1 test takers read the whole text first, then they tried to answer each test item by recalling information from the text, guessing and double checking with the text, or searching for an answer in the text. The more successful L2 test takers, on the other hand, showed three different and distinct patterns of behaviour. In the first pattern, two test takers did exactly the same thing the L1 speakers did. In the second pattern, two other test takers first read the test task and then quickly skimmed the whole text to get an idea about the text. Then they carefully read each test item and searched for the relevant information and answer in the text. In the third pattern, one other successful L2 test taker just read the test item and went to the text to search for the relevant information and answer. She neither read nor skimmed the text before starting to answer the test items. She read only parts of the text that seemed relevant to the test task. She chose to focus on doing the test task rather than building a text representation. The main reason most L2 test takers did not carefully read the text was time pressure. They worried they might not be able to finish the tasks within the time limit of the test, as they claimed in their introspective reports.

Based on these observations, it appeared that reading comprehension of the test takers was strategic. Based on the assessment of their own reading proficiency, they decided to adopt reading behavior that helps them do the task. L1 test takers chose to read the whole text carefully. They developed a text representation then answered the test items. The same observation was made for only two of the L2 test takers who scored very high on the test. Three successful L2 test takers chose different processes and strategies to achieve the test tasks which was very similar to the less successful L2 test takers' approach.

For most L2 test takers, achieving the test tasks was the main focus not reading and comprehending the whole test. In fact, they did not do reading in the normal sense of the word where a text is fully read to get the main points along with some details. For them reading and understanding the text was totally dependent on the task demand. They read and comprehended only what the tasks asked for, while for L1 test takers the test task made sense only after they finished reading the text. In brief, the activities used during test performance showed that for the L1 test takers test performance was comprehension-based while for the L2 test takers it was mostly task-based and related to task management.

6.4.2 Results of “Searching Theme” in the IELTS RCM Test Performance

Another theme that emerged from the processes and strategies used was “Searching” which is defined as using different processes, strategies, and activities to locate the relevant information for answering a test item. Identifying the test task and what it asked for, most test takers did not have an answer at hand for the questions and the first thing they had to do was to go back to the text to find the relevant information there and read it to get the answer. In fact, before answering the test item, test takers needed to locate the information that they thought might help them answer the question. However, this information was not readily available in the text and they had to locate it in the text. The information they needed for answering the question was somewhere in the text and they needed to locate it first. They had to locate it, read or re-read it and understand it to get to the answer. Even when they guessed an answer, they still needed some verification which still needed going back to the text. Search processes became more important when the test tasks asked for some specific information which was not normally remembered. So, searching for the relevant information was indispensable.

Based on the first cycle of coding, a set of codes emerged that were all part of the test taker’ effort to search the text to locate the information relevant to the test item. They shared one feature in common, i.e., they were all part of the test takers’ attempt to locate the relevant information in the text. The processes and strategies used were categories that were coded into a theme, i.e., “searching the text for locating the relevant information” for short, “searching theme”. So “Searching” referred to a set of process categories and strategy categories that were related to the test takers’ efforts to locate the relevant information in the text ((See Appendix L, for definition of the process and strategies used and actual examples from the test taker transcripts). Test takers searched different parts of the text to locate the relevant information.

Some read a whole paragraph, some skimmed few paragraphs, and some re-read few sentences. It included three distinct categories; 1) processes and strategies related to searching different sections of the text such as phrase level search, sentence level search, paragraph level search, and text level search and 2) the specific process and strategies used for searching these categories which included skimming the text, scanning specific words/phrases in the text, careful reading of a certain part of the text, and re-reading and 3) miscellaneous processes and strategies.

The search level included sentence level search, paragraph level search, several paragraphs search, and text level search. These were all related to the scope of search. For some items the whole text was searched while for some other items search was limited to a few sentences. In addition, test takers also used some specific processes to locate the relevant information. They skimmed the text, scanned specific words and phrases in the text, read a paragraph carefully, re-read certain section of the text, etc. For some test tasks the text was scanned while for some other task it was carefully read. Other processes and strategies such as “struggle with comprehension of the paragraphs”, “focus on doing the test task not comprehension of the text” were also observed in the search processes. As they were related to none of the two categories mentioned above, they were referred to as “miscellaneous processes” (See Figure 6.17).

It is worth mentioning that searching focused on locating the relevant information not reading and comprehending it. Processing the relevant information is discussed under a different theme. Therefore, a distinction needs to be made between “searching to locate the relevant information” and “processing the relevant information”. The former involved what test takers did to *locate* the relevant information while the latter involved what test takers did to *process* the relevant information. It seemed that unlike the L1 test takers, L2 test takers had serious problems locating the relevant information and had to spend a lot of time to locate the relevant information.

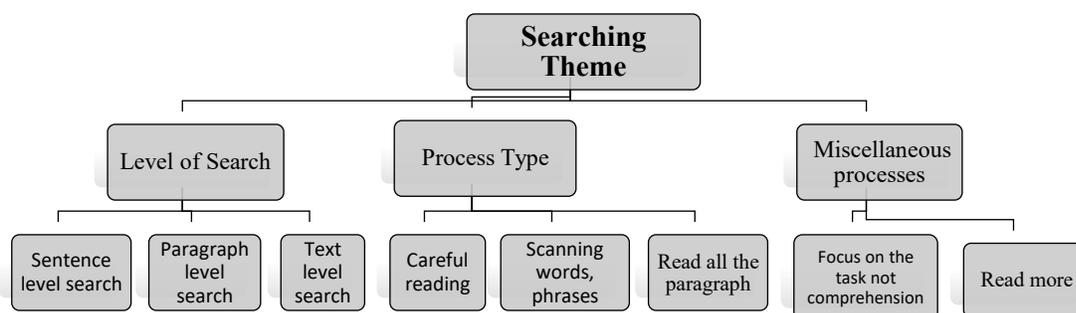


Figure 6.17. The main processes and strategies used in “Searching Theme”

To develop a detailed description of the processes and strategies used in “Searching Theme”, they were further analyzed by use of NVivo crosstab query which could show how the search processes and strategies were distributed in different groups of test takers. Table 6.2 shows results of cross tab for the use of these sub-processes across different groups of test takers.

Table 6.2.

Results of crosstab analysis of searching theme across the three groups of test takers

Search category	L1 Test takers	L2: More successful Test takers	L2: Less successful Test takers	Total
<i>What: Search level</i>				
1. Search at paragraph level	10(100%)	5(100%)	6(100%)	21
2. Search several paragraphs	9(90%)	5(100%)	6(100%)	20
3. Sentence level search	8(80%)	3(60%)	4(66%)	15
<i>How: Type of search process</i>				
4. Scanning specific words, phrases, or ideas	9(90%)	5(100%)	5(83%)	19
5. Read all the paragraph carefully	9(90%)	4(80%)	4(66%)	17
6. Skim the text to locate the relevant information	10(100%)	3(60%)	3(50%)	16
7. Recall info from previous reading	10(100%)	3(60%)	2(33%)	15
8. Careful reading	8(80%)	1(20%)	5(83%)	13
Miscellaneous processes				
9. Read more text to make sure nothing is missing	8(80%)	4(80%)	3(50%)	15
10. Focus on doing the task	2(20%)	3(60%)	4(66%)	9
11. Struggle with comprehension of the paragraph	2(20%)	1(20%)	6(100%)	9

It is worth noting that the frequencies reported refer to the number of participants who used these processes not the number of times these sub-processes were used. As shown in Table 6.2, all the 21 test takers, irrespective of their L1 or level of language proficiency, searched for relevant information at “paragraph level”. This means that to get the relevant information, test takers looked for the relevant information which they thought was located in a certain paragraph. Likewise, searching several paragraphs for locating the relevant information was used by 20 participants. For some test tasks, the relevant information could not be located in one paragraph, so test takers continued reading the next paragraph(s) to locate it. For some test tasks the search was limited to a single sentence. “sentence level search” was used by 15 test takers.

In terms of the how the processes of search category were achieved, 19 test takers used “scanning specific words/phrases”. “Skimming parts of the text” was used by 16 test takers. Another search process used was “careful reading” which was used by 13 test takers. “Reading all the paragraph carefully” and “reading carefully” were also very frequent processes and were used by 17 test takers.

“Recall information from previous reading” was used by 15 test takers. They recalled relevant information from their previous reading and tried to answer the question by using that information. instead of going back to the text. In fact, the relevant information was readily available in the test takers’ mind. Except for one test taker who finalized the answer without even double checking with the text, most test takers double checked their guess and searched the text for the relevant information.

Three other miscellaneous processes were also observed in test takers’ performance in their search category; 1) “reading more text to make sure nothing is missing” 2) “focusing on doing the task instead of focusing on comprehension of the text” and 3) “struggling with comprehension of the paragraph”. As these three processes could not be categorized under search level nor could they be part of type of search process. they were called miscellaneous processes. They were used by 15, 9, and 9 test takers, respectively.

Based on these results, it can be argued that searching could help test takers attend to details of test items and the specific sections of the text that explicitly or implicitly addressed the test item. In general, L1 test takers could locate the relevant information more quickly and efficiently simply because they had read the whole text carefully before they chose to answer the test items. They had the text representation in mind and could use their mental map for locating

the relevant information in the text. They frequently used skimming the text to locate the relevant information. Most of the L2 test taker (8/11), on the other hand, had not read the whole text and with no text representation in mind they could not easily locate the relevant information. They relied more on “careful reading” and “reading the whole paragraph” to locate the relevant information. In search for locating the relevant information, all the less successful L2 test takers “struggled with comprehension of the text” which can explain why they needed more time to the test. In fact, one main reason the amount of time spent on the test differed for the three groups of test taker was the search processes which took more time for the L2 test takers especially the less successful ones.

6.4.2.1 Distribution of the “Searching Theme” across different test tasks

To specify how searching processes and strategies were used across different test tasks, they were also closely examined across different test tasks. NVivo has a query option that allows looking closely at the specific context in which these search processes are used. This option allows researchers identify the test tasks for which participants used a certain search process. For example, one can look into the coded transcripts and find out who used “search at paragraph level” and for what test task(s) it was used. For a closer examination of the distribution of each process across t test tasks, all the coded data were examined. (See Appendix K, for definition of the process and strategies used and actual examples from the test taker transcripts.) Results of searching theme across different test tasks are also presented in Appendix K.

Based on the frequency of their use, the processes and strategies used in searching for each group of test takers were calculated. Table 6.3 presents result of the total frequency and mean frequency of the search category for different test takers.

Table 6.3.

Summary of the frequency of searching theme in different groups of test takers

	L1 test takers	L2: Successful test takers	L2: Less successful test takers	Total
<i>What: Search level</i>				
1. Search at paragraph level	130(49%)	84(31%)	56(20%)	270(100%)
2. Search several paragraphs	26(47%)	14(26%)	15(27%)	54(100%)
3. Sentence level search	12(50%)	7(29%)	5(21%)	24(100%)
<i>How: Type of search process</i>				
4. Recall info from previous reading				
5. Scanning specific words, phrases, or ideas	84(55%)	29(19%)	40(26%)	153(100%)
6. Skim the text to locate the relevant information	30(55%)	15(27%)	9(16%)	54(100%)
7. Read the whole paragraph carefully	17(24%)	18(25%)	36(51%)	71(100%)
8. Use previous reading	28(65%)	8(19%)	7(16%)	43(100%)
9. Careful reading	8(20%)	10(25)	22(55%)	40(100%)
<i>Miscellaneous processes</i>				
10. Read more text to make sure nothing is missing	10(50%)	4(20%)	6(30%)	20(100%)
11. Focus on doing the task	3(14%)	6(28%)	12(58%)	21(100%)
12. Struggle with comprehension of the paragraph	3(8%)	5(14%)	28(78%)	36(100%)
Total	351(45%)	200(25%)	236(30%)	786(100%)
Mean frequency per person	34	40	40	

As shown in Table 6.3., the total number of processes and strategies used in “searching” were 351, 200, and 238 for the L1, the more successful, and the less successful L2 test takers, respectively. The mean frequency, which is more meaningful representation of the frequencies, however, showed that L2 test takers used searching processes and strategies more frequently than the L1 test takers. This indicates that for the L2 test takers test performance involved more searching. This makes sense if one considers the L1 test takers’ reading behavior who read the whole test before answering the questions. They did not need to search much for the relevant information in the text because they had some of the ideas of the text in mind and could recall it for answering some of the questions.

Results also showed that the mean frequency of searching for the L1 test takers, the more successful and the less successful L2 test takers were 35, 40, and 40, respectively, lending more support to the raw frequency results.

To examine the observed differences in the frequencies of the “searching theme”, Chi-square goodness of fit was applied to the mean frequencies. Result of Chi-Square goodness of fit for the mean frequency of the searching theme is presented in Table 6.4.

Table 6.4.

Results of chi-square goodness of fit for the mean frequency of the search category

	Observed	Expected	Difference	Difference Sq.	Diff. Sq./Exp. Fr.
L1 test takers	34	38	-4.00	16.00	0.42
L2: more successful TTs	40	38	2.00	4.00	0.11
L2: less successful TTs	40	38	2.00	4.00	0.11
					.632

As shown in Table 6.4, the Chi Square value obtained was 0.63 which is not significant at $p < .05$. with the p -value is .73. This simply means that there is no significant difference between the frequency of processes and strategies used by different groups of test takers.

6.4.2.2 Variation of searching theme across test takers: Similarities and differences

Searching revealed some of the hidden processes of reading comprehension used by the test takers. It showed what the test takers do before they actually answer each test item. In many items that asked for inferential comprehension, test takers had to read more, process more, and strategize more to get to the answer. Test takers needed to unpack the implicit meanings in the text by using different processes and strategies. To do so they needed to read, re-read, and to collect more information from the text. They needed to locate the relevant information, read it carefully, relate it to the test question to see if it was compatible or incompatible with the gap in the test item. In many test items, the test task went beyond a simple matching of one idea in the test question with a statement in the text. They had to infer some ideas from the text and use it for answering a test item. They needed to read the test question, comprehend it before they go to the text and search for an answer. This contrast between the less successful L2 test takers and the other two groups in their search can explain high discrimination power of IELTS RCM. In fact, the scores discriminated the three groups of test takers which can be partly attributed to the different processes they used in searching for the relevant information in the text. Searching processes significantly enriched the SKSPs used in test performance and could reveal areas

where the less successful L2 test takers failed to achieve the test tasks. In general, searching theme revealed part of the complexity of the construct the IELTS RCM and highlighted the internal dynamic of test performance.

Based on the processes used in searching, a number of similarities and difference can be observed in the search processes of the three groups of test takers. First, there was sharp contrast between the less successful L2 test takers and the other two groups in their search for the relevant information. All the less successful L2 test takers had serious problems with comprehending the three main texts and locating the relevant information. They struggled with answering almost all types of test tasks, especially the *Diagram Completion Task* and the *Yes, No, Not given Task*. The only tasks they had no comprehension problem with were the *Summary Completion Task 1* and *2*. The more successful L2 test takers and the L1 speakers, on the other hand, showed no signs of problems in their reading, processing and comprehending the text and doing the test tasks. The only exception was the *Diagram Completion Task* which was somehow technical and required some background knowledge.

Second, L1 test takers searched a very specific section or paragraph by quick skimming through that section while L2 test takers seemed to search a larger part of the text to locate the relevant section and the relevant information it contained. L1 test takers most often had a clear idea where the relevant information lied and just scanned the information they were looking for while L2 test takers had problem locating the relevant information. For them, it was a two-phase process of locating the relevant section and then locating the relevant information within the section. Such processing differences can be explained by memory aspect of reading comprehension and development of a text representation. L2 test takers could not efficiently develop such a representation because they did not read the text in its entirety. Had they read the text in its entirety, probably they could have developed a text representation to refer to during their search processes.

Third, another main difference observed in the searching processes and strategies of the three groups was “recall information from previous reading”. For some items, L1 test takers used no search. While all L2 test takers at both levels of language proficiency had to go back to the text to get the answer. The L1 test takers relied on their memory of the text and answered the question. They just read the test item and recalled information from their reading and answered the item. They had carefully read the whole text and developed a text representation which they

could use for retrieving the relevant information. This indicates that they had developed a textual representation after reading of the whole text otherwise how could they recall information from the text. Another evidence that supports the development of text representation by the L1 test takers was use of guessing strategy and double-checking strategy. For some items, instead of searching for an answer in the text, L1 test takers guessed an answer and went back to the text just to double check it with the text and to examine if it is true.

Another feature of search category for L1 test takers was the speed and efficiency of the process. Unlike the less successful L2 test takers, L1 test takers did not show any sign of confusion for finding the relevant information. Based on their textual representation an awareness of the text structure they knew where to look for the answer. The speed with which they searched the text can be explained in terms of the role of memory and recall during reading processes.

Finally, locating the relevant information was not straight forward because there were many synonyms and paraphrasing involved in each test task. For many items, the language of the test question and the language used in the text differed which added to the complexity of and necessity of the search processes. Use of different lexical items and grammatical structures called for use of vocabulary knowledge, lexical inferencing, and paraphrasing skill which in turn require good mastery of grammar.

6.4.2.3. Association of test tasks and searching theme

As mentioned earlier the searching processes and strategies were not used randomly. Closer inspection of their use showed association of these processes and strategies with certain test tasks. It seemed that task features and task demands exert influence on the searching processes and strategies. For tasks that asked for specific details, the search was limited to smaller chunks of the text such as part of a paragraph or sentence. For example, the *Diagram Completion Task* asked for two-word phrases as an answer or *the Summary Completion Tasks* that asked for single word answers, the search was conducted at much smaller units such as sentence and phrase while for the *Matching Headings Task* which asked for more global meaning the search was conducted at paragraph level. Another example that showed the influence of test task characteristics on the searching processes and strategies could be observed in the *Multiple-Choice (with two answers) Task* which asked for the main idea of the whole text. Hence, test takers searched for the relevant information in several paragraphs. Such associations

between searching processes and the test tasks were closely examined and tabulated. Based on the frequency band of the process and strategies used, six levels of associations were formed. Based on these frequency bands strength of association between the sub-processes and different test tasks were determined. Table 6.5 presents these degrees of association.

Table 6.5.

Degree of association between test tasks and search category

Frequency band	>50-40	30-39	21-29	11-20 weak	0-10
Intensity of association	Very strong	Strong	Moderate	Weak	Very weak

The frequencies ranged between 0->50. Frequency interval of 10 was used to them divided into 6 frequency bands. Based on these frequency bands, the strength of association between the searching process and strategies and test tasks were tabulated. Table 6.6 shows the association of the search processes with the test tasks.

Table 6.6.

Association of the search processes used with the test tasks

Search processes	<i>Test tasks</i>	<i>Strength of association</i>
<i>What: Search level</i>		
1. Search at paragraph level	All task except for Summary Task 2	----- Strong
2. Search several paragraphs	MC (two answers)-True, False, Not given-Matching Features-	----- Moderate
3. Sentence level search	Summary tasks	----- Weak
<i>How: Type of search process</i>		
4. Scanning specific words, phrases, or ideas	All test tasks except for Matching Headings	----- Moderate
5. Use previous reading	Summary task 2- MC (two answers)	----- Weak
6. Skim the text to locate the relevant information	True, False, Not given	----- Weak
7. Read all the paragraph	Matching Headings	----- Strong
8. Read more to make sure nothing is missing	True, False, Not given.	----- Very weak
9. Focus on doing the task	Few test tasks	----- Weak
10. Struggle with comprehension of the paragraph	Diagram Completion Task	----- Strong

“Paragraph level search” was highly associated with the *Matching Headings Task* (f, 64) and the *Multiple-Choice Task* (f, 53). This process was also moderately associated with the *Multiple Choice (two answers) Task*, the *Matching Features Task*, and the *Diagram Completion Task*. The *Summary Completion Tasks 1 and 2*, on the other hand, had the least association with paragraph level search. Based on these results, it can be concluded that except for the *Summary Completion Tasks 1 and 2*, in all other test tasks paragraph was the level at which relevant information was searched. Sentence level search did not have much association with most test tasks. It was only weakly associated with the *Summary Completion Task 1 and 2*, and poorly

associated with the *Diagram Completion Task* and the *Yes, No, Not given Task*. In the *Summary Completion Task 1 and 2*, the test takers read each sentence in the summary and searched for similar language in the text.

A quick look at the results of the observed associations indicates that the processes and strategies used in searching theme were not randomly used across test tasks, rather they seemed to be systematically related to the type of test task at hand. Some test tasks specifically asked for more global ideas which required searching several paragraphs while other test tasks asked for the general idea of a paragraph which demanded paragraph level search. Still other tasks sought to measure comprehension of specific details which required sentence level search. For instance, searching several paragraphs was used only for one test task, i.e., the *Multiple Choice (two answers) Task* and the last *Multiple-Choice* item which asked for a subtitle and required comprehension of the whole text. These tasks were asking for global points in the text and required processing more the text. The answer to the *Multiple-Choice (two answers) Task* was searched for in several paragraphs simply because the relevant information was discussed in several paragraphs. The same was observed for the last multiple-choice item (item 40) which asked for an appropriate sub-title for the passage. Most test takers, however, answered this item based on their understanding of the text and did not go back to the text to search for an answer.

In summary, based on the results of second cycle of coding, searching theme proved to be one of the key components of test performance. In addition, the processes that were identified in the searching theme were not randomly used, rather they were associated with different test tasks. Second, some search processes were common to more test tasks while others were more restricted to few test tasks. Finally, choice of searching processes and strategies was influenced by the type of test tasks and test takers language proficiency.

6.4.2.4. Patterns of searching theme for different test takers

As shown in Figure 6.18, three different patterns were observed in L1 test takers' performance; 1) they read the text to search for the relevant information and get the answer, 2) they guessed an answer and they searched the text to check their guess/tentative answer, and 3) they did not read at all and relied on their recall of information from the text. As some test tasks asked for some specific information, test takers had to go back to the text and see what the answer is. This pattern was observed in all groups of test takers especially the less successful L2 test takers who relied on this pattern as the main way to approach the task. As to the second

pattern, for some test tasks, test takers had a rough idea what the answer is and they guessed an answer, but they were not sure if it is correct. So, they went back to the text to check their guess and decide on the final answer. This pattern was mostly observed in the L1 test takers and some of the more successful L2 test takers. This search pattern was guided by the test takers' text representation developed from the first reading of the text. In the third pattern, test takers relied on the text representation and answered the test item by recalling information from their memory and the text representation they had developed during reading the whole text. This pattern was observed in few L1 test takers.

Most L1 test takers used all the three patterns in their test performance. More successful L2 test takers, on the other hand, mostly used pattern one and to a less extent pattern two in their test performance. In few cases they could guess an answer and double check it in the text. The less successful L2 test takers used only one pattern in their performance. They just read the question and searched for an answer by reading the text. They read item by item and proceeded reading the text as required by the test item. They used no guessing, nor did they use recall information from previous reading because there was no previous reading. They started with reading the test task and focused on answering the task by reading only what is relevant to the item. This was the only pattern they used in their performance. These patterns do not seem to be optional and test takers used them based on the requirement of the test item and their own reading skills and ability.

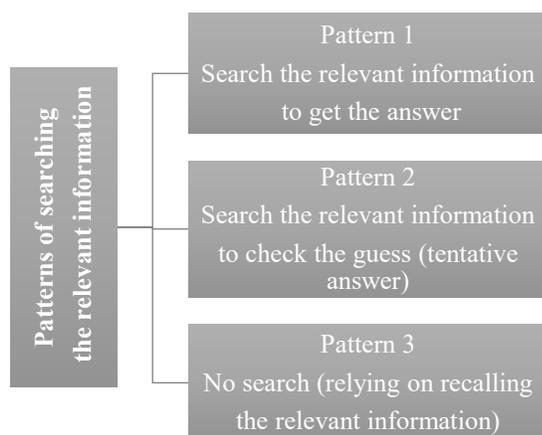


Figure 6.18. Patterns of searching the relevant information

6.4.3 Results of Answering theme in the IELTS RCM Test Performance.

As mentioned earlier, test takers used a number of processes and strategies in their test performance. Some of these processes were grouped under “Reading Theme” and “Searching Theme” as described in the previous sections. The third main theme that emerged from the data was “Answering Theme”. This section reports details of the processes and strategies used in answering theme and results of crosstab across the three groups of test takers, distribution of the sub-processes in test tasks, and association of the processes and strategies used in answering theme with each specific test task.

Based on the results of the first cycle of coding a number of codes that were thematically related to processing the relevant information in the text emerged from the data. The common thread that connected some of these process categories and strategy categories was that they were all related to analyzing and processing the relevant information. These codes were integrated and coded into “Answering Theme”. Identifying the information relevant to the test item, test takers had to process it to arrive at an answer. Those who guessed an answer needed to double check their guess with the relevant information, but for those who had no idea what the answer might be, they had to read the relevant information more carefully to get the answer. In either case, test takers had to go through a number of processes, strategies, and use different knowledge sources and skills to get the answer. For instance, after locating the relevant information, test takers attended to lexical clues given in the text or test item and used “lexical inferencing” to answer the test item. Or they noticed the similarity of the language used in the text and test items. So, they used “paraphrasing” to get to the answer. Another important answering process was “inferencing” the answer based on the information provided in the relevant information. “Answering theme”, then, referred to the specific processes, strategies, and knowledge sources and skills that were used in reading, processing, and comprehending the relevant information to arrive at an answer. they included inferencing, lexical inferencing, paraphrasing, careful reading of the relevant information, attending to lexical clues, and use of vocabulary knowledge. (See Appendix L, for definition of the process and strategies used and actual examples from the test taker transcripts.) The processes and strategies used for processing the relevant information were carried out at different levels of sentence, paragraph, and text

As shown in Figure 6.19, different processes and strategies were used to arrive at an answer. They were grouped into two main categories; 1) level of processing, and 2) the specific

processes, strategies, and activities used. Level of processing included sentence level process, paragraph level process, several paragraphs process. The strategies used to process these levels of the text included “attending to literal meaning”, “answering the question without re-reading the text”, “careful reading of the relevant information”, “inferencing”, “using vocabulary knowledge”, “attention to paraphrasing”, “recall information from the text”, “re-reading the relevant information or the text”, “lexical inferencing”, “highlighting the answer in the text”, and “answer the question while reading”.

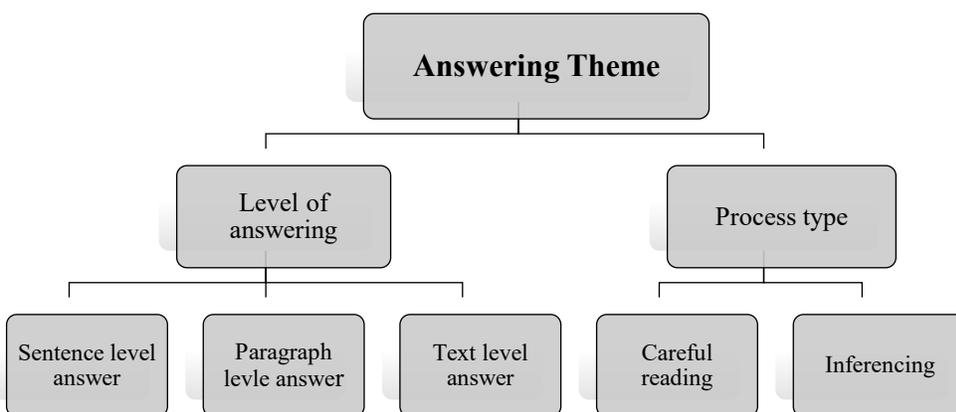


Figure 6.19. Some processes used in “Answering Theme”

6.4.3.1. Answering theme across test takers

To get a better picture of how different test takers used these processes and strategies they were further analyzed across the three groups of test takers. To this end, NVivo crosstab query was used. Table 6.7 reports the frequency and percentage of each process across the three groups of test takers. The percentages for each group are reported in parentheses.

Table 6.7.
Frequency and percentage of “answering processes” across groups

Answering processes/strategies	L1 test takers	L2: More successful	L2: Less successful	Total
Level of processing				
Sentence level answer	10(100%)	5(100%)	6(100%)	21
Paragraph level answer	10(100%)	5(100%)	5(83%)	20
Several paragraphs level	9(90%)	5(100%)	4(66%)	18
Phrase level answer	5(50%)	3(100%)	2(33%)	10
Processes, strategies and activities involved in answering process				
Use of vocabulary knowledge	10(100%)	5(100%)	6(100%)	21
Attending to the literal meaning	9(90%)	5(100%)	5(83%)	19
Careful reading of the relevant info	8(80%)	5(100%)	6(100%)	19
Inferencing	10(100%)	5(100%)	4(66%)	19
Paraphrasing	8(80%)	4(80%)	4 (66%)	16
Recall info from the text (no-re-reading)	9(90%)	3(60%)	4(66%)	16
Re-reading paragraph or the relevant info.	7(70%)	3(60%)	5(83%)	15
Lexical inferencing	7(70%)	5(100%)	3(50%)	15
Highlighting the answer in the text	3(30%)	2(40%)	1(16%)	6
Answer the question while reading	1(10%)	0(0%)	0(0%)	1
Total	10	5	6	21

As mentioned earlier the frequencies reported in Table 6.7 refer to the number of participants who used these processes not the frequency of their use. As shown in Table 6. 43, most of these processes were used by the three groups of test takers. The frequencies and percentages show the number of participants in each group who used these processes. For some processes, they were used by all members of the groups while for some other processes they were used by only a fraction of each group. For example, “highlighting the answer in the text” was used by only 6 test takers. Almost all the test takers, irrespective of their L1 or level of language proficiency, sought the answers at sentence level (100%) and paragraph level (95%). Processing the answer in several paragraphs and at phrase level was used by (85%) and (38%) of the test takers, respectively. Results also showed that most of the processes were frequently used by all the test takers. Processes such as “use of vocabulary knowledge” was used by all test takers and other processes were used by 90%-71% of the participants. Other processes and strategies such as “paraphrasing” was used by only 80% of the L1 test takers and more successful L2 test takers and 66% of the less successful L2 test takers. Also, it is interesting to note that some strategies such as “answer the question while reading the text” was used by only one participant. Overall, high frequency of most of the processes and strategies used in “answering theme” showed that they were indispensable part of the test performance.

To have an overall picture of the frequency of the processes and strategies used for answering, they were tabulated and summarized. Table 6.8 presents summary of total frequency of the answering processes for the three groups of test takers.

Table 6.8.

Summary of frequency of in answering theme by different groups of test takers

	L1 test takers	L2: More successful	L2: Less successful	Total
Level of processing				
Sentence level answer	103(58%)	47(26%)	27(16%)	177(100%)
Paragraph level answer	59(57%)	31(30%)	13(13%)	103(100%)
Several paragraphs level	11(44%)	5(20%)	9(36%)	25(100%)
Phrase level answer	9(60%)	3(20%)	3(20%)	15(100%)
Processes and strategies in answering theme				
Use of vocabulary knowledge	52(36%)	44(30%)	50(34%)	146(100%)
Literal meaning	30(45%)	26(30%)	11(25%)	67(100%)
Careful reading of the relevant info	36(50%)	14(20%)	22(30%)	72(100%)
Inferencing	54(68%)	20(25%)	6(7%)	80(100%)
Paraphrasing	29(58%)	15(29%)	7(13%)	49(100%)
Recall info from the text (no-re-reading)	37(70%)	12(22%)	4(8%)	53(100%)
Re-reading paragraph or the relevant info.	18(26%)	11(16%)	40(58%)	69(100%)
Lexical inferencing	14(48%)	9(31%)	6(21%)	29(100%)
Highlighting the answer in the text	7(59%)	3(25%)	2(16%)	12(100%)
Answer the question while reading	2(100%)	0(100%)	0(100%)	2(100%)
Total	461(52%)	238(26%)	200(22%)	899(100%)
Mean frequency per person	46	47	33	

To examine the observed differences in the frequencies of the processes and strategies used in “searching theme”, Chi-square goodness of fit was applied to the mean frequencies. Result of chi-square goodness of fit for the mean frequency of the search processes is presented in Table 6.9.

Table 6.9.

Result of chi-square goodness of fit for the mean frequency of the searching theme

	Observed	Expected	Difference	Difference Sq.	Diff. Sq./Exp. Fr.
L1 test takers	46	42	4.00	16.00	0.38
L2: more successful TTs	47	42	5.00	25.00	0.60
L2: less successful TTs	33	42	-9.00	81.00	1.93
					2.905

As shown in Table 6.9, the Chi Square value obtained was 2.34 which is not significant at $p < .05$. with the p -value of .73. This simply means that there is no significant difference between the three groups of test takers in the frequency use of answering processes and strategies.

6.4.3.2. Association of the answering processes and test tasks

Furthermore, association between each process and strategy used in answering the test tasks was further analyzed. To this end, based on the frequencies observed, a scale consisting of five-level categories was established. It indicated degree of associations between the processes and each test task. The scale ranged from very weak association to very strong association. Table 6.10 presents scale for classifying the degrees of association.

Table 6.10.

Scale for classification of degree of association between test tasks and answering theme.

Frequency band	>50-40	30-39	21-29	11-20 weak	0-10
Intensity of association	Very strong	Strong	Moderate	Weak	Very weak

A scale was developed for assessing degree of association between test asks and answering processes. As Table 6.10 shows, frequency band >50-40 indicated the highest degree of association and frequency band 0-10 was the lowest degree of association. Other degrees of association were strong, moderate, weak. Table 6. 11 reports the results of association between test tasks and processes and strategies used in answering theme.

Table 6.11.
Association of processes and strategies used in answering theme with test tasks

Answering process	<i>Task type</i>	<i>Strength of association</i>
<i>Processing levels</i>		
	True, False, Not given-----	Moderate
	Diagram Completion-----	Strong
	Multiple Choice-----	Very strong
Sentence level answer	Yes, No, Not given -----	Moderate
Paragraph level answer	Matching Headings -----	very strong
Several paragraphs level	Very weak for all tasks	
Phrase level answer	Very weak for all tasks	
	Diagram Completion -----	Very strong
	Matching Headings -----	Moderate
	Summary Task I -----	Moderate
Use of vocabulary knowledge	Multiple Choice -----	Moderate
Literal meaning	Weak or very weak with all tasks	
Careful reading of the relevant info	Weak or very weak with all tasks	
Inferencing	Yes, No, Not given -----	Strong
Paraphrasing	Weak or very weak for all tasks	
Recall info from the text (no-re-reading)	Weak or very weak for most tasks	
Re-reading paragraph or the relevant info.	Weak or very weak for most tasks	
Lexical inferencing	Weak or very weak for all tasks	
Highlighting the answer in the text	Weak or very weak for all tasks	
Answer the question while reading	Very weak for all tasks	

As shown in Table 6.11, for most processes and strategies very weak association could be established. Only few processes and strategies strongly associated with test tasks. For example, “Sentence level answer” strongly associated with the *Diagram Completion Task* and very strongly associated with the *Multiple-Choice Task*. It also moderately associated with the *True, False, Not given Task* and the *Yes, No, Not given Task*. “Paragraph level answer” also associated very strongly with the *Diagram Completion Task*. “Use of vocabulary knowledge” very strongly associated with the *Diagram Completion Task* and moderately associated with the *Matching Headings Task*, the *Summary Completion Task 1*, and the *Multiple-Choice Task*. Finally, “inferencing” was strongly associated with the *Yes, No, Not given Task*. All other answering - processes and strategies did not associate with the test tasks or they had a very weak association. Overall, these association could show some specific features of the test task. For example, strong association of “inferencing” with the *Yes, No, Not given Task* clearly showed the interpretive nature of the test task which asked the test takers to judge the inferences made in the *Yes, No, Not given* statements. Or “paragraph level answer” did not randomly associate very strongly with the *Matching Headings Task*. The task tapped the main idea of the paragraph which could not be answered unless test takers read the whole paragraph.

6.4.4. Results of Strategies Used in IELTS RCM Test Performance

In addition to the three main themes, i.e., Reading theme, Searching Theme, and Answering Theme, that emerged from the first and second cycles of coding, test takers also made use of some metacognitive strategies in their test performance. These strategies were used in different phases of reading, searching, and answering processes. Due to the importance of the metacognitive strategies in planning, managing, and assessing different phases of test performance, they were also separately tabulated and analyzed. In this section, results of strategies used by test takers are reported. It includes distribution of the strategies across different groups of test takers and across test tasks and association of strategies with different test tasks. Definition of each metacognitive strategy along with examples from test takers' transcripts are presented in Appendix M.

Results of the first cycle of data coding indicated that test takers made use of a wide variety of metacognitive strategies in their test performance. Use of strategies reflected conscious efforts of test takers to accomplish the test tasks. They used these strategies while processing the text and the test task. These strategies helped them decode the text and retrieve some information from the text and activate background knowledge to facilitate choosing or generating an answer to the test tasks. They used a variety of strategies such as delay answering some test items, highlighting the answer in the text, tagging each paragraph with a word or phrase, double checking the answer, rephrasing the test item, noticing the order of the questions in the text, narrowing down the options, noticing grammatical points.

To get a better picture of how test takers used these strategies, NVivo crosstab query was used for further analysis of the strategies among different groups of test takers. Table 6. 29 reports the frequency and percentage of each strategy across three groups of test takers.

It is worth mentioning that these frequencies indicate the number of participants who used each strategy not the number of times they used them. For instance, "use text structure knowledge" strategy was used by 10 L1 test takers, 5 more successful L2 test takers, and 5 less successful L2 test takers. In fact, these frequencies show the number of participants who reported use of this strategy in their retrospection verbal report. Practically, it was technically too difficult to assess the exact number of times these strategies were used in doing the 40 test tasks. Participants reported the use of these strategies in their test performance but not the number of times they used them. No test taker specifically indicated how many times s/he used them for

one test task or another. In some cases, frequencies of their use could be inferred but not counted. For instance, “tagging each paragraph with a word or phrase” which was used by only one L1 test taker was repeatedly used while she was reading each paragraph. She indicated that she repeatedly used this strategy while she was reading each text. As each of the three texts in the sample IELTS RCM text consisted of some 7-9 paragraphs, one can assume that she used this strategy at least 20 times which is a highly frequent strategy.

Table 6.12 shows the frequency and percentage of each strategy across the three groups of test takers. Given that the sample size for each group differed 10 (L1), 5 (more successful L2 test takers), and 6 (less successful L2 test takers), simple frequencies tend to be higher for larger groups.

Table 6.12.
Crosstab of strategies used across groups

Strategies used	L1	L2: More successful	L2: Less successful	Total
1. Use text structure knowledge	10(100%)	5(100%)	5(100%)	20
2. Eliminating/ narrowing down the options	9(90%)	4(80%)	5(83%)	18
3. Develop a gist of the text	10(100%)	4(80%)	2(34%)	16
4. Delay answering	8(80%)	4(80%)	4(66%)	16
5. Use test-wiseness strategy	6(60%)	5(100%)	5(83%)	16
6. Use of background knowledge	9(90%)	4(80%)	3(50%)	16
7. Attention to details	9(90%)	4(80%)	0(0%)	13
8. Double checking the answer	6(60%)	3(60%)	4(6%)	13
9. Guess an answer	4(40%)	3(60%)	6(100%)	13
10. Re-phrasing or re-reading the question	7(70%)	3(60%)	2(34%)	12
11. Plan a strategy	6(60%)	4(80%)	1(17%)	11
12. Developing and using a representation of the text	7(70%)	2(40%)	0(0%)	9
13. Highlight the answer in the text	4(40%)	3(60%)	1(17%)	8
14. Careful analysis of the options	6(60%)	1(20%)	0(0%)	7
15. Attention to certain features of the text	5(50%)	1(20%)	1(17%)	7
16. Recalling information from previous reading	3(30%)	0(0%)	2 (34%)	5
17. Noticing grammar	3(30%)	1(20%)	0(0%)	4
18. Read the text with some questions in mind	4(40%)	0(0%)	0(0%)	4
19. Attention to the repetition of inter-related words (lexical cohesion)	2(20%)	2(20%)	0(0%)	4
20. Attention to the topic sentence and concluding sentences	1(10%)	1(20%)	1(17%)	3
21. Recalling a question while reading the text	2(20%)	1(10%)	0(0%)	3
22. Answer the test task without moving back to the text	3(30%)	0(0%)	0(0%)	3
23. Noticing the order of the questions in the text	2(20%)	4(80%)	1(17%)	1
24. Re-read the last sentence in each paragraph	0(0%)	0(0%)	1(17%)	1
25. Tagging paragraphs	1(10%)	0(0%)	0(0%)	1
26. Use the heading to comprehend the paragraphs	0(0%)	0(0%)	1(17%)	1
Total	10	5	6	21

As mentioned earlier, in further analysis of the strategies, they were categorized in terms of the number of test takers who used them. Some strategies were used by more test takers while some were used by only few. For example, “Tagging the paragraph with a word/phrase” was used only by one test taker. Table 6.12 shows that most L1 test takers used a variety of strategies in their test performance. They used more types of strategies in their test performance while L2 test takers used fewer strategies.

Strategies were variably used by different groups of test takers. Some strategies such as “use text structure knowledge” were used by almost all the test takers irrespective of their language background (L1 or L2) and language proficiency level (high or low). They were used by the L1 test takers, more successful L2 test takers, and less successful L2 test takers. Some strategies such as “using test-wise strategies” and “attending to the topic sentence and concluding sentence” were equally used across the two groups of L2 test takers. Yet, other strategies were used more frequently by L1 test takers but not the L2 test takers. Some other strategies were uniquely used by only one group of test takers. For example, “using the headings to comprehend the paragraphs” was used only by less successful L2 test takers.

The first 15 (55%) strategies listed in Table 6.12 were used by all the test takers while the other 12 (45%) strategies were used by fewer test takers. L1 test takers used 27 (92%) of the strategies in the list while the more successful L2 test takers used only 22 (78%) of the strategies. Moreover, 6 (22%) other strategies were used by just one participant. As to the less successful L2 test takers, there were a number of strategies they did not use at all. They used only 9 (33%) the strategies. Based on the number of test takers who used these strategies, the top 10 strategies seemed to play a more significant role in the test processes. They formed more than 2/3 (68%) of all the strategies used by the test takers. The other 17 remaining strategies made just 1/3 (33%) of the strategy use.

These frequencies can show the weight and importance of these strategies across different groups of test takers. The variation observed in strategy use across test takers and test tasks can also reveal some information about the nature of the reading construct in IELTS RCM and how it might be influenced or alter across test takers. The more frequently used strategies had the potential to provide the most salient aspects of IELTS RCM construct while the less-frequently used can indicate individual variations among the test takers. For instance, “developing a gist of the text” was used by a total of 17 test takers across the three groups while “using the headings to comprehend the paragraph” which was used by only one less successful L2 test takers. Evidently, comprehending a text and developing its gist is core to the reading comprehension construct while using several headings to comprehend a paragraph shows the test taker’s inability to read a paragraph and comprehend it. In fact, the less frequently used strategies did not tell much about the core components of the construct of reading in IELTS RCM. They could tell something about some individual characteristics of the test taker. For instance, it indicated

that the test taker struggled to comprehend the main idea of some paragraphs and tried to make up for it by looking at the heading options.

As shown in these tables, different strategies were variably used by different groups of test takers across different test tasks. But the strategies were also variably used across different test tasks which can show how test task features, test takers' characteristics, and the strategies used interact. Closer analysis of the interaction helped explore some of these intricacies involved in test performance of each group of test takers. For example, further analysis of the results showed that certain strategies (10 strategies) were more frequently used by L1 test takers. The mean frequency of these strategies were much higher for the L1 test takers, indicating the influence of language background (L1-L2) on strategy use. Also, results showed that the less successful L2 test takers used “guessing an answer”, “Noting grammar”, “attention to the topic sentence”, “re-reading the last sentence in the paragraph”, and “using the headings to comprehend the paragraph”, more frequently than other test takers. It is worth noting that all these strategies were used to make for problems they faced in comprehending the text, showing that they faced more challenges in comprehension of the text and they appealed to these strategies to help them with answering the test tasks.

6.4.4.1 Summary of strategy used across test tasks and test takers

To have an overall picture of the number of strategies used by test takers in their performance, these strategies were tabulated and summarized. Table 6.13 presents summary of total frequency and the mean frequency of the strategies used by each group of test takers.

Table 6.13.
Summary of frequency of strategies used across groups

Strategies used	L1 test takers	L2: Successful test takers	L2: Less successful test takers	Total
1. Use of paragraph/text structure knowledge	57(60%)	24(25%)	14(15%)	95(100%)
2. Eliminating/narrowing the options	38(62%)	9(14%)	14(20%)	61(100%)
3. Develop a gist of the text	34(55%)	25(40%)	3(5%)	62(100%)
4. Delay answering	20(46%)	15(36%)	8(18%)	43(100%)
5. Test-wiseness strategy	12(30%)	17(44%)	10(26%)	39(100%)
6. Use of background knowledge	11(41%)	13(48%)	3(11%)	27(100%)
7. Attention to details	27(73%)	11(27%)	4(10%)	42(100%)
8. Double checking the answer	21(61%)	7(21%)	5(18%)	33(100%)
9. Guessing an answer	6(31%)	4(22%)	9(47%)	19(100%)
10. Re-phrasing or rereading the question	18(62%)	8(27%)	7(11%)	33(100%)
11. Planning a strategy	7(54%)	4(38%)	1(8%)	12(100%)
12. Developing and using a text representation	16(72%)	4(19%)	2(9%)	22(100%)
13. Highlighting the answer in the text	5(42%)	6(50%)	2(8%)	13(100%)
14. Careful analysis of the options	16(54%)	4(13%)	10(33%)	30(100%)
15. Attention to certain features of the text	10(78%)	3(11%)	4(11%)	17(100%)
16. Recalling information from previous reading	7(46%)	6(40%)	2(14%)	15(100%)
17. Noticing grammar	4(66%)	1(0%)	11(34%)	6(100%)
18. Read the text with the questions in mind	3(75%)	1(25%)	0(0%)	4(100%)
19. Attention to the repetition of inter-related words (lexical cohesion)	3(34%)	6(66%)	0(0%)	9(100%)
20. Attention to the topic sentence and concluding sentences	5(28%)	5(28%)	8(44%)	18(100%)
21. Recall the questions while reading	3(100)	0(0%)	0(0%)	3(100%)
22. Answer the test task without moving back to the text	8(100%)	0(0%)	0(0%)	8(100%)
23. Noticing the order of the questions in the text	3(18%)	7(41%)	7(41%)	17(100%)
24. Re-reading the last few sentences in each paragraph	1(20%)	2(40%)	4(40%)	7(100%)
25. Tagging paragraphs	10(100%)	0(0%)	0(0%)	10(100%)
26. Using the headings to comprehend the paragraphs	0(0%)	0(0%)	5(100%)	5(100%)
Total	323	143	88	554 (100%)
Mean frequency per person	32	28	15	

To examine the observed differences in the frequencies of the strategies used, chi-square goodness of fit was applied to the mean frequencies. Result of chi-square goodness of fit for the mean frequency of the strategies used is presented in Table 6.14.

Table 6.14.
Result of Chi-square goodness of fit for strategies used across test takers

	Observed	Expected	Difference	Difference Sq.	Diff. Sq./Exp. Fr.
L1 test takers	32	25	7.00	49.00	1.96
L2: more successful TTs	28	25	3.00	9.00	0.36
L2: less successful TTs	15	25	-10.00	100.00	4.00
					6.320

The Chi-square value obtained was 6.32. Compared with the p -value is .042 at $p < .05$, the observe difference in the use of strategies by the test takers was significant at $p < .05$. This indicates that L1 test takers systematically used more metacognitive strategies in their test performance than the L2 test takers, suggesting that compared to the L2 test takers, the IELTS RCM test tasks tap more into the strategic competence of the L1 test takers.

6.4.4.2 Association of strategy use with test tasks

Another aspect of the strategies used was their association with the test tasks. These strategies were not randomly used in the test takers' performance. Some of them were specifically used for specific tasks or more frequently used with a specific test task. For instance, use of background knowledge was more frequently used with the *Diagram Completion Task* (10 times), the *Yes, No, Not given Task* (6 times), and the *Matching Headings Task* (3 times). Test taker from different groups used background knowledge for answering some items on these tasks. Table 6.15 presents results of association of different strategies with different test tasks.

Table 6.15.
Association of the strategies used with test tasks

Strategies used	Test tasks	Strength of association
1. Use text structure knowledge (search)	-Diagram Completion----- -Multiple Choice (two answers) -----	Weak Weak
2. Eliminating/ narrowing down the options	-Multiple Choice----- -Matching Headings -----	Moderate Weak
3. Develop a gist of the text	-Matching Headings----- -Multiple Choice -----	Strong Weak
4. Delaying answering	-Matching Headings----- -Diagram Completion ----- -Yes, No, Not given-----	Weak Weak Weak
5. Test-wiseness strategy	-True, False, Not given-----	Weak
6. Use of background knowledge	-Diagram Completion Task----- -Yes, No, Not given ----- -Matching Headings -----	Weak Weak Weak
7. Attention to details	-Matching Headings -----	Strong
8. Double checking the answer	-	
9. Guessing the answer	-	
10. Re-phrasing or rereading the question	-Diagram Completion-----	Weak
11. Planning a strategy	-	
12. Developing and using a mental representation of the text	-Matching Headings-----	Weak
13. Highlighting the answer in the text	-	
14. Careful analysis of the options (answer)	-Multiple Choice (two answer) -----	Weak
15. Attention to certain features of the text	-Matching Features-----	Very weak
16. Recalling information from previous reading	-True, False, Not given-----	Very weak
17. Noticing grammar	-Summary Completion----- - Diagram Completion -----	Very weak Very weak
18. Read the text with a question in mind	-	
19. Attention to the semantically inter-related words	-Matching Headings -----	Weak
20. Attention to the topic or concluding sentences	-Matching Headings-----	Weak
21. Recall the questions while reading	-True, false, not given-----	Very weak
22. Answering the task without moving back to the text	-	
23. Noticing the order of the questions in the text	True, False, Not given-----	Very weak
24. Re-reading the last sentence in each paragraph	Matching Headings -----	Very weak
25. Tagging paragraphs	-True, False, Not given----- -Matching Headings-----	Very weak Very weak
26. Using the heading to comprehend the paragraphs	-Matching Headings-----	Very weak

Based on the rate of strategy use for each type of test task, a table of association was developed. These associations can better show what specific skills and processes are tapped by each test task. For instance, association of “attention to details” and “developing a gist of the text” were strongly associated with the *Matching Headings Task*. This clearly indicates the *Matching Headings Task* involved attending to both the main idea of the paragraph and its details. These associations can also show how difficult some test tasks or items are. For instance, “delay strategy” were associated with the *Diagram Completion Task*, the *Matching Headings*

Task, and the *Yes, No, Not given Task*. Compared to other test tasks, these tasks were relatively more difficult. Had they not posed a challenge to the test takers, test takers would not have delayed answering them. The observed associations between test tasks and strategies indicated that use of strategies is influenced by characteristics of test tasks and possibly reading ability of the test taker.

An interesting observation related to the use of strategies had to do with the *Matching Headings Task*. Results showed that it triggered use of 11 different metacognitive strategies. This indicates the task construct is multidimensional. In fact, the *Matching Headings Task* was one of the few tasks that measured comprehension at paragraph level and required developing text representation to choose the right heading. Based on this observation one can argue that some test tasks seem to require more processing and use of different knowledge sources, processes and activities. Not all test tasks seem to trigger strategy use. Among the 9 test tasks, *the Matching Headings Task* stood out and inviting and involving more strategy use. It enjoyed features that necessitate more strategy use. Other test tasks included the *Diagram Completion Task* and the *True, False, Not given Task* which were associated with 6 and 5 strategies, respectively.

6.4.4.3. Strategy use in different phases of test performance

In further analysis of verbal report transcripts, the strategies used by the test takers were also analyzed in terms of their use in different phases of test performance. These phases dealt with processing different components of the test, i.e. 1) processing the texts, 2) processing the test tasks, and 3) processes used for connecting the text to the test task. Some of the strategies were used to read, process, and comprehend the text while some other strategies were used in processing the text task. The third category of strategies were related to examining the test task with the ideas and information in the text. For instance, strategies such as “developing a gist of the text”, “paying attention to details”, or “reading the first and last sentence of the paragraph” were all related to reading and processing the text itself while “careful analysis of the test task” or “careful analysis of the options” were used in processing the test task. Finally, strategies such as “use of background knowledge”, “guessing an answer”, “narrowing down the options”, “double checking the answer” were used when test takers wanted to relate the text content to the test task. Table 6.16 presents categorization of different strategies based on their use in different phases of test performance.

Table 6.16.

Strategies used in different phases of test performance

Text processing strategies	Task processing strategies	Text-task processing strategies
<ul style="list-style-type: none"> -Developing a mental representation of the text -Develop a gist of the text -Attention to details -Attention to certain features of the text - Attention to the topic and concluding sentences - Re-reading the last sentence in each paragraph - Tagging paragraphs - Noticing grammar - Read the text with the questions in mind - Strategic reading 	<ul style="list-style-type: none"> -Eliminating and narrowing the options -Careful analysis of the options -Re-phrasing or rereading the question - Answering the test task without moving back to the text 	<ul style="list-style-type: none"> -Use of text structure knowledge -Delay answering - Use of background knowledge - Noticing the order of the questions in the text - Highlighting the answer in the text - Using the headings to comprehend the paragraphs - Using previous reading - Double checking the answer - Guessing the answer -Use Test-wiseness strategy - Recall the questions while reading -Planning a strategy

As shown in Table 6.16, most strategies were related to either reading and comprehending the text and relating the test task to the text. Only few strategies were related to comprehending or analysis of the test task. Based on the number of strategies used in different phases of test performance, it seems that the main challenge test takers faced was comprehending the text and finding an answer to the test task which required relating the text and the test task. The test tasks were clear and straightforward in posing the problem to the test takers.

Closer examination of these strategies in the context of the three phases of test performance revealed some valuable details that can shed some light on the construct of the IELTS RCM. Some strategies were unique to the L1 test takers and some to the L2 test takers. For example, “answering questions without looking back at the text” was used just by L1 test takers. Or “re-reading the last sentence in the paragraphs” was used by only less successful L2 test takers. There seems to be a systematic relationship between these associations. These strategies were not randomly used by test takers. For instance, the reason why L1 test takers did not look back at the text and just used their comprehension of the text to choose an answer was that they had read the whole text and had developed a mental representation of the text which could help them choose an answer while L2 test takers had not read the whole text and had to go back to the text to search for an answer. Or in the case of “re-reading the last sentence of the paragraph”, the less successful L2 test takers faced problems and challenges understanding the gist of the paragraph, therefore, they had to read the last sentence to see if they can find some concluding statements to get the answer.

6.5. Results of Test Performance Observation Scheme (TPOS)

As mentioned in Chapter Four, as test takers were taking the test, their test performance was observed for the amount of time they spent on each test task, and the number of times they moved back and forth between the text and item in a task- (B&FMBTI (hereafter, B&FM). These dimensions of reading behavior were recorded in the *Test Performance Observation Scheme (TPOS)*. Results of these observations are reported below. In addition, it is worth mentioning that to improve the reliability and validity of the observations the recorded observations were double checked with the test takers' immediate retrospective accounts to make sure the observations were consistent with what the test takers said they did during test performance.

6.5.1. Frequency of back and forth movements between the text and test items (B&FM)

A very meaningful feature observed in the reading behaviour of the test takers was their back and forth movement between the test items and the texts (B&FM). B&FM is a rich source of information that can help understand the construct of different test tasks. It can indicate use of several processes and strategies for comprehending the text and the test task where test takers had to check more points and features of the test item and the text. It can also show the intensity of processing where more processes had to be conducted due to the difficulty of the test task or unique characteristic of the test task. So, (B&FM) can be a sign of task difficulty. In addition, it can also reveal how test takers with different L1s and levels of language proficiency may differ in their test performance.

All test takers moved back and forth between the text and the test task items to search for an answer or to double check their guesses. This behaviour was common to L1 and L2 test takers and for almost all test items. However, the frequency of back and forth movements between the test task items and the text (B&FM) differed across both test tasks and test takers. Frequency of these back and forth movements for each group of test takers was tallied first. However, as the sample size of the test taker groups differed, it made no sense to compare the simple frequencies of the groups. Therefore, mean frequencies for each group and test task were separately calculated. Mean frequency could take care of the variable group sizes. As shown in Table 6.17, mean frequency per person was calculated by dividing the simple frequency of (B&FM) to the

sample size of each group. Mean frequency per task was calculated by dividing the simple frequency of (B&FM) for each task to the number of items in each test task.

Table 6.17

Calculation of mean frequency of (B&FM) per task and per item

Mean frequency of (B&FM) per person = Total frequency (B&FM)/sample size

Mean frequency of (B&FM) per item = Total frequency of (B&FM)/ No. of items in the task)

Table 6.18 presents result of the simple frequencies and mean frequencies of back and forth movements between the text and the test tasks for different groups of test takers and test tasks.

Table 6.18.

Frequency of (B&FM) for different test tasks and test takers

	L1 Test takers			L2: Successful L2 Test takers			L2: Less successful Test takers		
	Total F.	F/ person	F/Task	Total F	F/ person	F/task	Total F	F/ person	F/task
<i>True, False, Not given</i>	58	5.8	<u>14.5</u>	50	10	<u>12.5</u>	74	12.3	<u>18.5</u>
<i>Matching Features</i>	79	7.9	<u>19.7</u>	58	11.6	<u>14.5</u>	87	14.5	<u>21.7</u>
<i>Diagram Completion</i>	158	15.8	<u>31.6</u>	119	23.8	<u>23.8</u>	152	25.3	<u>30.4</u>
<i>Matching Headings</i>	214	21.4	<u>30.5</u>	76	15.2	<u>10.8</u>	90	15	<u>12.8</u>
<i>Summary 1</i>	57	5.7	<u>14.2</u>	42	8.4	<u>10.5</u>	35	5.8	<u>8.75</u>
<i>MC (Two answers)</i>	41	4.1	<u>20.5</u>	17	3.4	<u>8.5</u>	31	5.1	<u>15.5</u>
<i>Multiple Choice</i>	117	11.7	<u>23.4</u>	56	11.2	<u>11.2</u>	74	12.3	<u>14.8</u>
<i>Summary 2</i>	47	4.7	<u>15.6</u>	41	8.2	<u>13.6</u>	40	6.6	<u>13.3</u>
<i>Yes, No, Not given</i>	94	9.4	<u>15.6</u>	44	8.8	<u>7.3</u>	55	9.1	<u>9.1</u>
Total	866	86.6		503	100.6		638	106.6	

As shown in Table 6.18, the mean frequencies of B&FM for the L1 test takers, more successful L2 test takers, and less successful L2 test takers were 86.5, 100.6, and 106.3, respectively. This indicates that L1 test takers could rely more on their memory of the text and make less reference to the text. On average, the less successful L2 test takers had the highest frequency of B&FM and L1 test takers had the lowest rate of B&FM. The more successful L2 test takers fell in between. It seemed that before they could answer the test items, the L2 test takers had much work to do and process different features of the text and the test items.

The mean frequencies of B&FM also differed for different test tasks. Some test tasks involved more back and forth movements which can indicate the nature and intensity of the

processes involved. Except for three test tasks (the *Summary Completion Task 1*, the *Matching Headings Task*, and the *Yes, No, Not given Task*), the less successful L2 test takers had the highest frequencies of back and forth movements which indicates that they had more to do while performing on the test. For the *Matching Headings Task*, the trend was quite the opposite with the L1 speakers having the highest frequency and the less successful the least frequency. The *Yes, No, Not given Task* was mostly the same and the frequencies did not show significant difference. Finally, for the *Summary Completion Task 1 and 2*, the more successful L2 participants had the highest frequency and L1 test takers and less successful L2 test takers had almost the same frequency.

For L1 test takers the highest frequency of (B&FM) belonged to the *Diagram Completion Task* (31.6 times per task), followed by the *Matching Headings Task* (30.5 per task), the *Multiple-Choice Task* (23.4 times per task), and the *Matching Features Task* (1.97 times per task). This means that for the *Diagram Completion Task*, for instance, each participant moved back and forth 30.1 times between the test task and the text for the whole task. For more successful L2 test takers, the highest mean frequency of (B&FM) was for the *Diagram Completion Task* (23.8 times per task) followed by the *Matching Features Task* (14.5 times per task) and the *Summary Completion Task 2* (13.6 times per task). For less successful L2 test takers, the highest frequency of (B&FM) belonged to the *Diagram Completion Task* (30.4 times per task), followed by the *Matching Features Task* (21.7 times per task), and the *True, False, Not given Task* (18.5 times per task).

Results also showed that the mean frequency of B&FM for each of the L2 test taker participant groups varied and was dependent on the test takers' proficiency and type of test task. English L1 speakers differed from the L2 test takers in the frequency of B&FM; and, the more successful L2 test takers differed from the less successful L2 participants. To examine the observed differences between the (B&FM) of different groups of test takers, Chi Square significance tests was conducted. Table 6.19 presents the result.

Table 6.19.
Results of the Chi Square significance tests of B&FM across proficiency groups

	Observed	Expected	Difference	Difference Sq.	Diff. Sq./Exp. Fr.
L1 test takers	87	98	-11.00	121.00	1.23
L2: more successful TTs	100	98	2.00	4.00	0.04
L2: less successful TTs	107	98	9.00	81.00	0.83
					2.102

As shown in Table 6.7, the chi square value was 2.102 and the p -value was .35. The result was *not* significant, with alpha set at $p < .05$. This indicated that in spite of the observed differences, there was no significant difference between the number of times test takers moved back and forth between the text and the test task.

Table 6.20 rank orders tasks with the most and least frequent (B&FM). The ranking can be used as an indication of task difficulty. More B&FM can indicate more processing and difficulty. Test tasks at the top of the table required the more (B&FM) and tasks at the bottom of the table required the less (B&FM).

Table 6.20.
Rank order of (B&FM) for different test tasks and test takers*

	L1 Test takers	L2: Successful Test takers	L2: Less successful Test takers
1	<i>Diagram Completion</i>	<i>Diagram Completion</i>	<i>Diagram Completion</i>
2	<i>Matching Headings</i>	<i>Matching Features</i>	<i>Matching Features</i>
3	<i>Multiple Choice</i>	<i>Summary Task 2</i>	<i>True, False, Not given</i>
4	<i>Multiple Choice (two answers)</i>	<i>True, False, Not given</i>	<i>Multiple Choice (two answers)</i>
5	<i>Matching Features</i>	<i>Multiple Choice</i>	<i>Multiple Choice</i>
6	<i>Summary Task 2</i>	<i>Matching Headings</i>	<i>Summary 2</i>
7	<i>Yes, No, Not given</i>	<i>Summary Task 1</i>	<i>Matching Headings</i>
8	<i>True, False, Not given</i>	<i>Multiple Choice (two answers)</i>	<i>Yes, No, Not given</i>
9	<i>Summary Task 1</i>	<i>Yes, No, Not given</i>	<i>Summary Task 1</i>

*NOTE: (B&FMBTT, abbreviated to B&FM): Back and forth movements between the test item and the text

It is interesting to note that for all the three groups of test takers, the *Diagram Completion Task* involved the most (B&FM). The *Matching Features Task* was also among the top three test tasks that required high frequency (B&FM) for the L2 test takers. results also showed that tasks that required the least number of (B&FM) were very different across the three groups of test takers. The *Summary Completion Task 1* and the *Yes, No, Not given Task* were at the bottom of

the ranking for all test takers. *The True, False, Not given Task* was at the bottom of the ranking for the L1 test takers, while for the L2 test takers it was among the top tasks.

6.5.2. Time spent on the test tasks

Another aspect of test takers' behaviour that was observed and recorded during their test performance was the amount of time they spent on each test task. For each participant the amount of time spent on each test task was separately recorded. Amount of time spent on a task cannot not show what processes were being used but it can show if and how the processes and strategies used were affected by characteristics of the test task and/or the test taker. Table 6.21 reports results of the overall time spent for each group of test takers for each test task.

Table 6.21.

Time spent on the test tasks

Test takers	L1 Test takers			L2: Successful			L2: Less successful		
	Task	Person	Item	Task	Person	Item	Tasks	Person	Item
Test tasks									
<i>True, False, Not given</i>	58m	5.8 m	87s	30m	6m	90s	58m	9.6m	145s
<i>Matching Features</i>	26m	2.6m	39s	34m	6.8m	102s	35m	5.8m	87s
<i>Diagram Completion</i>	39m	3.9m	47s	35m	7m	84s	61m	10.16m	122s
<i>Matching Headings</i>	107m	10.7m	92s	70m	14m	120s	98m	16.33m	140s
<i>Summary Completion 1</i>	24m	2.4m	36s	22m	4.4m	66s	37m	6.13m	92s
<i>MC (two answers)</i>	19m	1.9m	57s	11m	2.2m	66s	22m	3.66m	110s
<i>Multiple Choice</i>	72m	7.2m	86s	53m	10.58m	127s	71m	11.83	142s
<i>Summary Completion 2</i>	21m	2.1m	42s	18m	3.6m	72s	20m	3.30m	66s
<i>Yes, No, not given</i>	47m	4.7m	47s	31m	6.2m	62s	50m	8.3m	83s
<i>Total amount of time spent</i>	413m			304m			452m		
<i>Average time spent by each TT</i>	41.3m		61.95	60.8m		91.25	75.6m		113.75

As shown in Table 6.21, all the L1 test takers and the more successful L2 test takers could finish all the test tasks within the time limit of 60 minutes. On average, L1 test takers spent 41.3 minutes and more successful test takers spent 60.8 minutes to finish all the test tasks. None of the less successful L2 test takers, however, could finish the test tasks within the time limit of 60 minutes and they failed to do the last two test tasks of the third passage. It is worth mentioning that as one purpose of this research was to examine the construct of each type of test task across different levels of language proficiency, they were given some more time to finish the tasks and provide their accounts of the test task. Without such a provision, the last two test tasks,

i.e., The *Summary Completion Task 2* and The *Yes, No, Not given Task* could not be examined with this group of test takers. The average amount of time they needed to finish the tasks was 75.6 minutes. On average, L1 test takers, more successful L2 test takers, and less successful L2 test takers spent about 1 minute, 1.5 minutes, and 2 minutes for each test item, respectively.

L1 test takers answered the test items within the range of 36-92 seconds per item while for more successful and less successful L2 test takers the test items were answered within the range of 66-127 seconds, and 66-145 seconds, respectively. For L1 test takers, six tasks were done in less than one minute while for more successful and less successful L2 test takers no task was done in less than 60 seconds. These values clearly showed that L2 test takers needed more time to do the tasks while the L1 test takers could finish the test tasks even less than the given time (60 minutes).

Based on the average amount of time spent on each test task by the test takers, the test tasks were placed into five time-bands ranging from more than two minutes (>120s) to less than one minute (< 60s). Other times included 2 minutes, 1.5 minutes, 1 minute and less than 1 minute. Table 6.22 presents the time bands of each type of test task for the three groups of test takers.

Table 6.22.
Time bands for different test tasks across groups of test takers

Time spent (seconds)	L1 test takers	L2: More successful	L2: Less successful
>120s	-	-Multiple Choice (127s)	-True, False, Not given(145s) -Multiple Choice (142s) -Matching Headings (140s)
120s	-	-Matching Headings (120s)	-Diagram Completion Task(122s) Multiple Choice (110s)
90s	-True, False, Not given(87s)	-Matching Features Task(102s) -True, False, Not given(90s) -Diagram Completion Task(84s)	-Summary 1 (92s) -Matching Features Task(87s) -Yes, No, Not given (83s)
60s	-Matching Headings (70s) -MC (two answers) (57s) -Multiple choice (53s)	-Summary 2 (72s) -Summary 1 (66s) -MC (two answers) (66s) -Yes, No, Not given (62s)	-Summary 2 (66s)
<60s	-Diagram Completion Task(47s) -Summary 2 (42s) -Matching Features Task(39s) -Summary 1 (36s)	-	-

As Table 6.22 suggests, for L1 test takers, tasks that needed the most amount of time were, the *True, False, Not given Task* (87s), the *Matching Headings Task* (70s), the *Multiple-Choice Task*, the *Multiple Choice (two answers) Task* (53s). On average, it took almost 90 seconds for each item in these tasks to be answered. For the *Multiple-Choice Task* the time spent was almost 60 seconds and for all other tasks it took less than one minute to do. For more successful L2 test takers, on the other hand, the *Multiple-Choice Task* took the most time (>120s). Next was the *Matching Headings Task* (120s) and the *True, False, Not given Task* (90s). For less successful L2 test takers, three tasks took more than 120 s to do (the *True, False, Not given Task*/the *Matching Headings Task*, and the *Multiple-Choice Task*). Two tasks took 120s to do (the *Diagram Completion Task*, the *Multiple Choice (two answers) Task*, and the *Multiple-Choice Task*). The *Matching Features Task*, the *Summary Completion Task 1*, and the *Yes, no, Not given Task* were done within the range of 90s. Only the *Summary Completion Task 2* was done within the range of 60s and no task was done in less than 60 seconds (<60s).

In sum, the amount of time spent on the task can indicate the difficulty of the task or the intensity of the processes involved in doing the task. For some tasks and items, test takers spent more time reading and processing the text and the items, guessing an answer or searching for it, and examining it. For each one of these processes, they needed to spend some time. More time indicates more processes and the speed which the test takers could efficiently go through the processes. It is an indication of how the three groups of test takers differed in their reading fluency (reading speed). It differentiated one group from another. The amount of time spent could be looked at as an index of task difficulty and item difficulty, an index which can be cross checked with other indicators of task difficulty such as the test takers' self-declared difficulty of the tasks or the actual test scores on the test. However, based on the time spent by the three groups, "time" seems to be influencing the test construct. It would seem that L2 test takers would perform better with more time available to them. So, time may be considered as part of the reading construct. L2 test takers could definitely perform better had they been given extra time.

6.6. Difficulty of the IELTS RCM Tasks

To assess task difficulty, a multi-faceted approach was adopted. In addition to the number of back and forth movements between the test and the test items, the amount of time spent on

each task and the test scores obtained, the overall task difficulty of each of the test tasks was examined.

To assess task difficulty of the IELTS RCM, the common indices of item difficulty in language testing (Brown, 2014) were adopted to categorize test task difficulty. Task difficulty simply refers to the degree of difficulty based on how an item is answered. It is calculated by dividing the number of wrong answers by the total number of items. Therefore, for each group of test takers, the number of wrong answers for each test task was calculated by counting the number of wrong answers divided by the total number of answers. Based on the difficult indices, they can be assessed as excellent, good, fair, and poor (Brown, 2014). Table 6.23 presents task difficulty indices that were used for categorizing test task difficulty.

Table 6.23.

Item difficulty levels

	Index	Difficulty
1.	0.40 and larger	Excellent
2.	0.30-0.39	Good
3.	0.11-0.29	Fair
4.	0.0-0.10	Poor

Based on the test scores obtained, task difficulty of the nine test tasks for each group of test takers was calculated. Table 6.24 presents result of task difficulty for each task for the three groups of test takers.

Table 6. 24.

Task difficulty for the three groups of test takers based on test scores

L1 Test takers									
	<i>True, False, Not given</i>	<i>Matching Features</i>	<i>Diagram Completion</i>	<i>Matching Headings</i>	<i>Summary Completion 1</i>	<i>Multiple Choice (Two)</i>	<i>Multiple Choice</i>	<i>Summary completion 2</i>	<i>Yes, No, Not given</i>
Right	38	37	47	64	39	20	46	30	43
Wrong	2	3	3	6	1	0	4	0	17
ID	.05	.05	.06	.09	.03	.00	.08	.00	.29
More successful L2 Test takers									
Right	18	17	23	24	19	10	13	14	28
Wrong	2	3	2	11	1	0	7	1	7
ID	.10	.15	.18	.32	.05	.00	.35	.07	.20
Less successful L2 Test takers									
Right	17	15	23	22	15	9	12	13	17
Wrong	7	9	7	20	9	3	18	5	19
ID	.30	.38	.24	.48	.21	.25	.60	.28	.53

As shown in Table 6.24, test tasks enjoyed different levels of difficulty for each group of test takers. The most difficult test task for the L1 test takers was the *Yes, No, Not given Task* (.29) which is fairly difficult. Other indices of difficulty for other test tasks ranged between 0.00 and 09 which are at poor level of difficulty which means most test tasks were fairly easy for the L1 test takers. For more successful L2 test takers, the most difficult test tasks were the *Multiple-Choice Task* (.35) and the *Matching Headings Task* (.32) which were at fair level of difficulty. For the less successful L2 test takers all test tasks were difficult. Three most difficult tasks test tasks enjoyed excellent difficulty indices were *the Multiple-Choice Task* (.60), the *Yes, No, Not given Task* (.53), and the *Matching Headings Task* (.48). The *Matching Features Task* (.38) and the *True, False, Not given* (.30) had good difficulty level. Finally, the *Summary Completion Task 1* (.21), the *Diagram Completion Task* (.24), the *Summary Completion Task 2* (.28), had fair difficulty level. Based on these indices, test tasks' difficulty was categorized into four levels of excellent, good, fair, and poor. Table 6.25 reports the result.

Table 6.25.

Difficulty of test tasks for each group of test takers

	Excellent (.40>)	Good (0.30-0.39)	Fair (0.11-0.29)	Poor (0.00-0.10)
L1s Test takers	-	-	Yes, No, Not given	All other tasks
More successful Test takers	-	- <i>Matching Headings, Multiple-choice</i>	- <i>Matching Features, Diagram completion, Yes, no, not given</i>	- <i>True, false, not given Summary 1 Summary 2</i>
Less successful Test takers	- <i>Multiple-choice Yes, No, Not given Matching Headings</i>	- <i>True, false, not given Matching Features Task</i>	- <i>Summary 1 Multiple choice (two answers) Summary 2</i>	-

In summary, for L1 test takers most test tasks were not difficult at all. Only *the Yes, No, Not Task* was fairly difficult while for more successful L2 test taker most tasks were at good-fair difficulty levels. Two tasks (*the Matching Headings Task* and the *Multiple-Choice Task*) were at good level of difficult and three tasks were fairly difficult. Two tasks (*the Matching Features Task*, the *Diagram Completion Task*, and the *Yes, No, Not given Task*) were poorly difficult for this group of test takers. For the less successful L2 test takers, most tasks enjoyed good-excellent difficulty and no test task was poorly difficult. Three tasks (*the Multiple-Choice Task*, *the Yes, no, Not given Task*, and the *Matching Headings Task*) enjoyed excellent levels of difficulty and two tasks (*the True, False, Not given Task* and the *Matching Features Task*) had good difficulty

levels. Finally, three tasks (the *Summary Completion Task 1*, the *Multiple Choice (two answers) Task*, and the *Summary Completion Task 2*) were fairly difficult for this group of test takers.

Another approach adopted to examine task difficulty was asking the test takers to identify the difficulty of the test tasks after they finished their performance. Based, on their actual experience with the test, they rank ordered difficulty of the test tasks. Table 6.27 summarizes the results of task difficulty for each group of test takers.

Table 6.26.

Self-declared task difficulty

Test takers	L1 Test takers	L2: more successful TTs	L2: less successful TTs
Task difficulty	1) <i>Diagram Completion</i> , 2) <i>Matching Headings</i> 3) <i>(True, False, Not given/ Yes, No, Not given)</i> 5) <i>Matching Features</i> 6) <i>MC (two answers)</i> 7) <i>Multiple choice</i> 8) <i>Summary Task 1</i> 9) <i>Summary Task 2</i>	1) <i>Matching Headings</i> 2) <i>Diagram Completion</i> 3) <i>Yes, no, not given</i> 4) <i>MC (two answers)</i> 5) <i>True, False, Not given</i> 6) <i>Multiple Choice</i> 7) <i>Matching Features Task</i> 8) <i>Summary Task 1</i> 9) <i>Summary Task 2</i>	1) <i>(Yes, no, not given, MC (Two answers)</i> 3) <i>Matching Headings</i> 4) <i>True, False, Not given</i> 5) <i>Multiple Choice</i> 6) <i>Matching Features</i> 7) <i>Diagram Completion)</i> 8) <i>Summary Task 1</i> 9) <i>Summary Task 2</i>

As shown in Table 6.26, for the majority of L1 test takers, the *Diagram Completion Task* was the most difficult test task to do. They could not comprehend the processes being discussed in the text and the diagram. Next was the *Matching Headings Task*. They thought the headings were very implicit and needed inferencing which made the test task challenging. The *True, False, Not given Task* and the *Yes, No, Not given tasks* were equally difficult for them. They thought these two test tasks involved test items with “not given” as their correct answer. such items were difficult to process and decide. Other tasks were moderately difficult. Finally, they all agreed that the *Summary Completion Task 1 and 2* were the easiest task especially the *Summary Completion Task 2* which asked for choosing the correct word to fill the blanks.

For the more successful L2 test takers, the most difficult tasks with equal degree of difficulty were the *Matching Headings Task* and the *Yes, No, Not given Task*. This was followed by the *Diagram Completion Task* and the *Yes, No, Not given Task* and the *True, False, Not given Tasks*. They argued that unlike other tasks which addressed ideas that were explicitly mentioned in the text, the *Matching Headings Task* and the *Yes, No, Not given Task* targeted comprehension of ideas that were implicitly presented in the text. Other tasks such as the *Multiple-Choice Task* were moderately difficult for them. To them the easiest tasks were the *Summary Completion Tasks 1 and 2*. For the less successful L2 test takers, no particular pattern of task difficulty

emerged. Every test taker had a different ranking of the difficulty, but they all agreed that tasks such as the *Matching Headings Task*; the *Yes, No, Not given Task*, the *Matching Features Task*, the *Multiple-Choice Task (two answers)* and the *Multiple-Choice Task* are difficult. They did not agree on the specific difficulty ranking of these tasks. Like the other two groups of test takers, they all agreed that the *Summary Completion Task 1 and 2* were the easiest tasks.

In addition to the difficulty ranking, all test takers agreed that the *Matching Headings Task*, the *Matching Features Task*, the *Diagram Completion Task*, and the *Yes, No, Not given Tasks* involved more reading than others and they had to read more text and process more to get to the answers.

One final point that is very relevant to task difficulty in the context of the current study has to do with lexical profile of the texts used in the sample IELTS RCM. As mentioned earlier in the content analysis of each test task, almost all test items contained some form of lexical clues such as synonyms which could be used by the test takers to get the answer to different test items. However, for the less successful L2 test takers who had limited vocabulary knowledge, this caused comprehension problems and added to the challenge of understanding the text. As they did not know several the words used in the text, they failed to relate it to the relevant information in the text. The lexical clues were very helpful to the L1 and the successful L2 test takers. They could use these clues to locate the relevant information in the text more efficiently.

To examine the degree to which the sample IELTS RCM text included some new words for the test takers, they were asked to identify those words after they finished doing the test.

Table 6.27 presents list of the new words test takers faced while doing the test.

Table 6.27.

List of new words in the IELTS RCM for each group of test takers

Native TTs	More successful TTs	Less successful test takers
Amygdala	Starboard-grooves-hauled	wreck -mishandled- degradation-starboard-marine- fleet-overladen- sunken- sonar-sediment- catastrophe- groove-meticulously- replica- starboard- neuroaesthetics- hauled-accelerated- toppling-sledges-hauled-infertile-canoes-descended-windbreaks-grooves-shrivelled-manoeuvre-fuzzy-Amygdala-haphazard- shed-inclination-chimp-scrutiny-intricacy-precise-fractal- perceptual-evolve

As the results in Table 6.27 show, the IELTS RCM texts included several new vocabulary items for both more successful and less successful L2 test takers. However, few native

test takers claimed not knowing the word “Amygdala” which was the only unfamiliar word for them. The number of new words in the text can explain some of the difference observed in the test takers’ processes and strategies they used in their test performance.

In summary, task difficulty analysis based on different approaches and sources of evidence, including actual test scores, the amount of time spent on each test task, the number of times test takers moved back and forth between the text and the items in a task, and the self-declared ranking of task difficulty allowed delving into different dimensions of test task difficulty. Unlike the unimodal approach which identifies task difficulty based on actual test score only, the multi-modal approach adopted for the analysis of task difficulty of the IELTS RCM mixed product-based and process-based approaches to task difficulty. It is rich in information for understanding the construct of each test task. It helps test developers make informed decision for inclusion or exclusion of test a test task or item in the test. The main advantage of the multi-modal approach was integration of both quantitative and qualitative dimensions of task difficulty. Figure 6.20 presents the multi-modal triangulated approach adopted in the examination of task difficulty of the IELTS RCM.

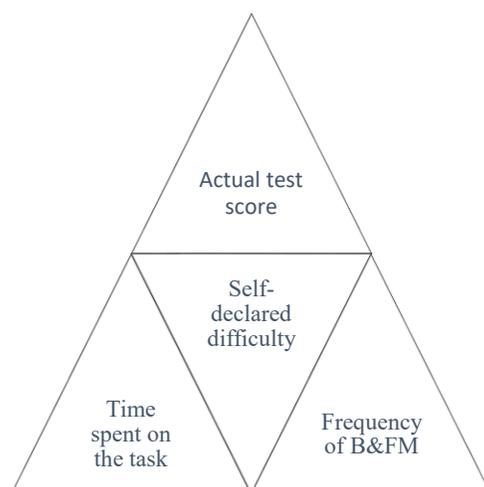


Figure 6.20. Multi-modal triangulated approach to task difficulty

6.7. Summary of the Chapter

This chapter reported on the results of the construct of each of the nine test tasks in IELTS RRCM based on their verbatim retrospective verbal reports. It provided a window to look at the skills, knowledge sources, processes, and strategies (SKSPs) used by the three groups of test takers during their test performance on the IELTS RCM. In addition, based on the first and

second cycles of coding three main themes that emerged from the first and second cycles of coding were reported. This included 1) Reading Theme, 2) Searching Theme and 3) Answering Theme. Details of these three themes of test performance across test takers and test tasks were also presented. Additionally, the chapter included association of these components with different test tasks. Finally, results of the metacognitive strategies used during test performance were reported. In the next chapter, findings of the experts' accounts will be reported.

CHAPTER SEVEN: RESULTS OF EXPERTS' ACCOUNTS AND JUDGEMENTS

7.1. Introduction

As discussed in Chapter Two, one viable construct validity source of evidence is accounts and judgements of testing experts. Their perception of what each test task measures is based on well-informed knowledge and practice in language testing and provide a different angle in the study of construct validity of the test construct. In the second phase of the study a group of 10 testing experts with several years of language testing experience were asked to take the sample IELTS RCM test, do the test tasks, and provide their accounts and judgements of the test tasks by filling in the “*Construct of the IELTS RCM Questionnaire*” for each test task (See Appendix I, for a copy of the questionnaire). After completing the questionnaire, the experts also added further elaborations on their judgements, if there were any. This chapter reports on the results of testing experts' accounts and judgements of the test tasks which are presented under the following classifications (See Chapter Four, Section 4.5.7); level of reading (e.g., word, phrase, sentence, paragraph, text), knowledge involved (e.g., lexical, grammatical, textual, and topical), type of comprehension (e.g., literal, inferential, evaluative), type of reading (e.g., careful-expeditious), text difficulty (e.g., easy-difficult), task difficulty (e.g., easy difficult), level of task processing (low-high), and time needed (e.g., less than one minute- more than 2 minutes). They classification were actually codes that were presented to the experts to judge the IELTS RCM test tasks. Their judgements of the test tasks served as a kind of protocol coding of the data, which was made up of the IELTS RCM with 9 different test tasks.

7.2. The Level of Reading

The first classification was concerned with the level of reading (e.g., word level, phrase level, sentence level, inter-sentence level, paragraph level, and text level) involved in task performance. Experts were asked to judge the level of reading they used or required in responding to an item or task. Table 7.1 reports the results of the experts' judgements for the reading level involved across different types of test tasks.

Table 7.1.

Results of the experts' judgements for the reading level

Level of reading	Word	Phrase	Sentence	Inter-sentence	Paragraph	Text	Total
<i>True, False, Not given Task</i>	0	1	3	5	4	2	15
<i>Matching Features Task</i>	0	1	4	4	4	1	14
<i>Diagram Completion Task</i>	2	3	8	2	4	0	19
<i>Matching Headings Task</i>	0	0	0	0	10	0	10
<i>Summary Task 1</i>	1	0	5	3	2	0	11
<i>Multiple Choice Task (two answers)</i>	0	0	0	0	3	7	10
<i>Multiple choice</i>	0	0	0	4	6	7	17
<i>Summary Task 2</i>	2	0	5	6	1	0	14
<i>Yes, No, Not given Task</i>	0	0	0	1	8	3	12
Total	5	5	25	25	42	20	122
Percentages	5%	5%	20%	20%	34%	16%	100%

As shown in Table 7.1, according to the experts' judgements, tasks differed in terms of what needed to be read for answering the test items. The least frequently selected levels of reading were word level reading (5%) and phrase level reading (5%). Text level reading was chosen for 16% of the test tasks. However, the most frequent level of reading was paragraph level reading (34%) followed by sentence level reading (20%) and inter-sentence level reading (20%). This simply means that experts indicated that for almost 74% of the test tasks the required readings was at sentence-paragraph levels. Closer examination of these percentages show that 85% of the test items can be answered by reading units not larger than a paragraph. This is a clear indication that majority of the RCM IELTS test tasks have little to do with reading and comprehending the whole text and the main ideas. The main finding from this classification of the experts' judgements is that test takers can answer the majority of the items by reading a few sentences or a paragraph but not the whole text.

Some test tasks such as the *Diagram Completion Task* and the *Summary Completion Tasks 1 and 2* asked for some specific information that could be answered at the sentence level whereas the *True, False, Not given Task* tended to involve the sentence and/or inter-sentence

level. Yet, the *Matching Headings Task* needed reading at the paragraph level. Textual level reading was mainly used just for three test items in the *Multiple-Choice (two answers) Task* and one of the *Multiple-Choice Task* items which asked for a subtitle for the text. These tasks were the only tasks that asked for textual understanding of a large portion of the text or the whole text.

Results also showed that for some tasks there was strong agreement among experts while for other test tasks there was great disparity among the experts. For instance, all experts (10/10) agreed that to do the *Matching Headings Task* test takers have to read the whole paragraph while for the *Matching Features Task* experts were inconsistent. They opted for phrase level (1/14), sentence level (4/14), inter-sentence level (4/14), paragraph level (4/14), and even text level (1/14).

One point mentioned by most of the experts was that reading the whole text and comprehending the gist of the text could help test takers in their test performance. They argued that without comprehension of the main point of the text, answering even a simple item will be a challenge for the test takers because they do not know where to look for it in the text and even reading the relevant information may not be enough to answer the item correctly because they have not seen the big picture of the text. One of the test takers mentioned,

I think all these items in the *True, False, Not given Task*, or the *Matching Features Task*, the *Summary Completion Tasks*, and the *Multiple-Choice Task*, and other tasks are not really very difficult to answer, but you need to read the whole text to answer them quickly. Of course, the time available puts a lot of pressure on the test taker and they may not read the whole text. (Jala)

7.3. The Type of Knowledge Needed

Another classification of test performance examined the type of knowledge needed to answer the test tasks. This included linguistic knowledge (lexical knowledge, grammar knowledge at sentence and inter-sentence levels, knowledge of paragraph structure and textual knowledge) and non-linguistic topical knowledge (Khalifa and Weir, 2009). Results of the type of knowledge needed to do the tasks are reported in Table 7.2.

Table 7.2.
Results of type of knowledge needed to do the tasks

Test tasks	Lexical K.	Sentential grammar	Inter-sentential grammar	Paragraph K.	Textual K.	Topical K.	Total
True, False Not given	8	6	4	1	1	1	21
Matching Features Task	7	7	3	1	1	0	19
Diagram Completion Task	9	6	1	3	0	3	22
Matching Headings Task	4	0	7	4	5	2	22
Summary Task 1	5	5	3	5	0	0	18
Multiple choice (two answers)	3	1	1	0	6	0	11
Multiple Choice Task	3	2	7	4	1	1	18
Summary Task 2	6	4	3	1	1	0	15
Yes, No, Not given Task	6	2	2	1	2	0	13
Total	51	33	31	20	17	7	159
Percentages	32%	21%	19%	14%	10%	4%	100%

As Table 7.2 shows, lexical knowledge (32%) and grammar knowledge (21%) were key to the knowledge required for doing the tasks. This finding is self-evident as vocabulary knowledge and grammar knowledge are core reading components that are needed for reading and comprehending the text and doing all the test tasks (Khalifa, and Weir, 2009, Grabe, 2009; Grabe and Stroller, 2011). The least frequent type of knowledge sources used were textual knowledge (10%) and topical knowledge (7%). Results also showed that the type of knowledge had a direct association with some types of test tasks. Textual knowledge, for example, was found to be relevant to only two test tasks, i.e., the *Matching Headings Task*, and the *Multiple-choice (with two answers) Task*. This was consistent with results of the content analysis which indicated these tasks tapped high-level processes of comprehension.

Results also indicated that there was no consensus over the type of knowledge involved in each test tasks. For most test tasks, experts had different judgements which can be explained

in terms of the judgemental nature of the task. Judging and rating are essentially subject to variation unless it is supported by continuous training and discussion of the criterion used for judgement. In the context of the current study the judgement criteria were orally described to the experts and briefly discussed with them. Oral description may not produce the same results as regular training. So, it is not surprising the results for some judgements varied across the experts. Further, testing experts differed in their judgements simply because they thought differently. For example, one of the experts mentioned,

The *Summary Completion Task 2* can be answered without going back to the text and it taps comprehension of simple sentences. (Rahi)

Another expert thought that the task could not be answered without going back to the text and re-reading relevant parts of the text. He said,

For this task [the *Summary Completion Task*], I think one should go back to the text and re-read two paragraphs. It seems to be a simple task, but you need more reading of the relevant sections. (Zare)

7.4. The Types of Comprehension

Another key classification of the reading construct had to do with the type of comprehension involved in each type of test task. Experts were asked to decide what kind of comprehension was involved in each test task: literal meaning, inferential meaning, or evaluative meaning. (As noted in Chapter Four, all categories were orally explained to the experts, and examples of each classification were provided before they evaluated the types of comprehension involved.) Since each test task consisted of several items any of which could tap into different dimension of comprehension, experts could choose more than one option for each test task. Some items could tap literal meaning while other items in the same test task could call for inferential or evaluative meaning. Results of experts' judgements on the type of comprehension tapped by each type of test task are presented in Table 7.3.

Table 7.3.
Results of experts' judgements on type of comprehension

Test tasks	Type of Comprehension			
	Literal	Inferential	Evaluative	Total
<i>True, False, Not given</i>	8	6	1	15
<i>Matching Features Task</i>	7	3	0	10
<i>Diagram Completion Task</i>	8	0	0	8
<i>Matching Headings Task</i>	3	7	0	10
<i>Summary completion 1</i>	9	1	0	10
<i>Multiple Choice Task (two answers)</i>	4	6	2	12
<i>Multiple Choice Task</i>	7	7	2	16
<i>Summary Completion 2</i>	8	1	0	9
<i>Yes, No, Not given Task</i>	4	6	5	15
Total	58	37	10	105
Percentage	55%	35%	10%	100%

As Table 7.3 shows, testing experts' judgements of the type of comprehension involved in test performance produced a total of 100% (n=105) of cases, out of which 55% (n=58) were assigned to literal comprehension while for inferential comprehension this rate was 37 (35%). The least chosen option was evaluative comprehension with only 10% (n=10) choosing this. Some experts also opted for lexical inferencing as another option. They believed that some test tasks involved lexical inferencing. They assigned lexical inferencing to 7% (n=11) of cases. This was then included under inferential comprehension. Testing experts also believed that in most types of test tasks (6/9), both literal and inferential meaning are involved. However, in all test tasks the literal meaning dominated the inferential meaning. For 5/9 test tasks the dominant type of comprehension was literal. Among different test tasks, the *Summary Completion Task 1 and 2* were the most literal and the *Matching Headings Task* and the *Multiple-Choice Task* were the most inferential. The *Yes, No, Not given Task* and the *Multiple-Choice Task* were the most diverse test tasks in terms of the type of meaning they tapped into. Experts indicated these tasks involved all three types of comprehension. Some items in these test tasks involved literal meaning and some tapped inferential and/or evaluative meaning.

According to the testing experts, the *Multiple-Choice Task (two answers)*, the *Multiple-Choice Task*, and the *Yes, No, Not given Task* involved some evaluative comprehension. They reported that the *Yes, No, Not given Task* required some judgements and the use of background knowledge, and the *Multiple-Choice Task (two answers)* required synthesizing different pieces of information from the text to arrive at an answer. They indicated that the answer to some items in these tasks were not explicitly stated in the text or the answer required nuanced understanding across several paragraphs. For the last multiple-choice item, which asked for a subtitle for the whole text, 3 testing experts believed it involved evaluative comprehension.

One of the experts reported,

When I read item 30, I was kind of surprised seeing an item asking for comprehension of the whole text. I knew won't be easy to answer and I had to read several paragraphs and compare them to choose the right option which I think was A. (Esha)

As mentioned, experts opted for literal meaning more than inferential meaning. To compare the observed frequencies in the experts' judgements of the type of comprehension involved in test tasks, Chi Square goodness of fit was run. Results of the test are reported in Table 7.4.

Table 7.4.

Results of Chi Square for types of comprehension

	Observed	Expected	Difference	Difference Sq.	Diff. Sq./Exp. Fr
Literal	58	35	23.00	529.00	15.11
Inferential	37	35	2.00	4.00	0.11
Evaluative	10	35	-25.00	625.00	17.86
					*33.086

Based on the results of Chi Square test (33.86, p-value <.00001) the difference between the types of comprehension was statistically significant. It can be concluded that according to the experts' judgements, the RCM IELTS basically tapped into literal meaning and to a lesser extent inferential meaning.

7.5. The Type of Reading (careful-expeditious)

Another classification judged by the experts was related to the type of reading involved. Tasks could involve careful reading of different parts of the text to get the main ideas and relevant details or expeditious reading of selective reading of certain sections of the text. Testing experts were asked to judge the degree to which the tasks needed to read carefully or expeditiously on a scale of 1-10, where 1 referred to very careful reading at one end of the continuum and 10 referred to expeditious reading at the other end. The values assigned by the experts were then analyzed to calculate the mean score for each test task. Results of their judgements are reported in Table 7.5.

Table 7.5.

Results of experts' judgements of type of reading (careful-expeditious)

Test tasks	Type of reading: Total score	Type of reading: Mean score	Type of Reading
<i>True, False Not given Task</i>	40	(4.0)	Careful
<i>Matching Features Task</i>	40	(4.0)	Careful
<i>Diagram Completion Task</i>	28	(2.8)	Very Careful
<i>Matching Headings Task</i>	42	(4.2)	Careful
<i>Summary Completion Task 1</i>	44	(4.4)	Careful
<i>Multiple Choice Task (two answers)</i>	46	(4.6)	Careful
<i>Multiple Choice Task</i>	40	(4.0)	Careful
<i>Summary Completion Task 2</i>	47	(4.7)	Careful
<i>Yes, No, Not given Task</i>	34	(3.4)	Very Careful

Based on the experts' judgements, the range of reading for different test tasks was 2.8-4.7, meaning that some test tasks required more careful reading (2.8) while some other test tasks involved less careful reading tending towards expeditious reading (4.7). According to the experts, all test tasks fell on the careful end of the continuum and no test task passed the midpoint (5) of the continuum which can be assumed representing *normal* reading. (NOTE: "Normal reading" was a term used by test takers especially the L1 test takers. They said they read the texts *normally*, meaning they read the whole text as they normally read their academic texts.

Two tasks fell on the very careful end of the continuum and required “very careful” reading; the *Diagram Completion Task* (2.8) and the *Yes, No, Not given Task* (3.4). All the other test tasks were somewhere in the middle of the continuum and were judged to require “careful reading”; the *Multiple Choice Task* (4.0), the *True, False, Not given task* (4.0), the *Matching Features Task*, (4.0) the *Summary Completion Task 1* (4.4), the *Multiple Choice Task* (two answers) (4.6), and the *Summary Completion Task 2* (4.7). It is interesting to note that all test tasks were on the careful end of the continuum, i.e., all had <5 (less than 5) values. Putting this finding in the context of the limited time of 60 minutes for the 40 test items, one can argue that in reality the test performance becomes a race against time and every test taker has to process the text and the test tasks as fast as possible. On the one hand, as indicated by testing experts, the test tasks involved careful reading which takes more time, on the other hand, test taker have only limited time of 20 minutes for each text. This gap was also observed in the test performance of the L2 test takers especially the less successful L2 test takers who struggled to finish the test tasks within the time limit. This is why all the L2 test takers found time as their main challenge and obstacle in test performance.

One of the experts argued answering test tasks requires careful reading of all the text within the very time limits of 20 minutes for each text but L2 readers cannot manage answering all the items without being test-wise. He reported,

I doubt if L2 people can really do this in 20 minutes. They should be very familiar with the test or they fail answering all items. (Zare)

Another expert argued that scanning and skimming can help read selectively, but without reading the relevant information in the scanned section, readers cannot answer the items. She said,

We can be very quick and scan the key words in the text, yet we need to read the relevant information very carefully to answer the question. (Esha)

In brief, the main findings related to the experts’ judgements of the type of reading involved in the IELTS RCM indicated that all test task involved careful and very careful reading and no task can be expeditiously read.

7.6. Text Difficulty

Experts were also asked to judge the degree they thought are easy-difficult to read on a scale of 1-10 where one represented the very easy to read and 10 represented very difficult to read. No specific criterion of difficulty was assigned to them for judgements. They were just asked to decide how difficult each text is based on their experience. The values they assigned were then analyzed to calculate the mean score of difficulty for each text. Results of experts' judgements of the difficulty of three RCM IELTS texts are reported in table 7.6.

Table 7.6.

Results of text difficulty judgements

Experts	1	2	3	4	5	6	7	8	9	10	Mean score
The Mary Rose	6	6	5	6	5	7	5	7	6	6	5.9/10
Easter Island	6	6	7	7	6	6	6	6	7	8	6.5/10
Neuraesthetics	7	8	6	7	7	7	8	8	6	7	7.1/10

As shown in Table 7.6, experts' judgements indicated that the texts were not easy and on a scale of 1-10. All the texts were above the midpoint of the scale, i.e., (5.0) which suggest they were not very difficult, rather moderately difficult. Results also showed that experts saw the "Neuroaesthetics" text as the most difficult with the mean difficulty of 7.1 and "the Mary Rose" text as the easiest with the mean difficulty of 5.9. Experts differed in their judgements of the texts. Some of them argued that "Neuraesthetics" text was more abstract and requires more careful reading and processing. Few experts, however, defined difficulty not just in terms of abstractness. They argued that the first text was the most difficult because it was full of facts and details which are difficult to absorb. Yet some other experts judged Easter Island the most difficult because it was full of names which made the text confusing.

7.7. Task Difficulty

Furthermore, testing experts also judged difficulty of each test task. Using the same scale of difficulty 1-10, they assigned different difficulty values to each test task. The values were then analyzed to calculate average difficulty for each test task. Results of task difficulty judgements are reported in Table 7.7

As shown in Table 7.7, according to the experts' judgements, among the nine test tasks, four tasks were above the midpoint of difficulty (5); the *Matching Headings Task* (6.2), the *Yes, No, Not given* (5.6), the *Diagram Completion Task* (5.3), and the *Matching Features Task* (5.2).

Except for the *Multiple-Choice Task* which fell on the midpoint (5.) other test tasks ranged between 3.6- 4.7 with the *Summary Completion Tasks 1 and 2* as the easiest tasks especially the *Summary Completion Task 2* (3.6) which involved choosing the answer from a table of answers.

Table 7.7.

Results of experts' judgements of the task difficulty

Test tasks	Level of Processing: Total score	Level of processing: Mean score	Low-High
<i>True, False Not given</i>	45	(4.5)	Easy
<i>Matching Features Task</i>	52	(5.2)	Moderate
<i>Diagram Completion Task</i>	53	(5.3)	Moderate
<i>Matching Headings Task</i>	62	(6.2)	Difficult
<i>Summary Completion Task 1</i>	47	(4.7)	Easy
<i>Multiple Choice (two answers)</i>	50	(5.0)	Difficult
<i>Multiple Choice Task</i>	50	(5.0)	Moderate
<i>Summary Completion Task 2</i>	36	(3.6)	Easy
<i>Yes, No, Not given Task</i>	56	(5.6)	Moderate

Based on the indices of difficulty assigned by the experts, the tasks were rank ordered from the easiest to the most difficult. Figure. 7.1 presents result of experts' judgements of task difficulty of the RCM IELTS test tasks.

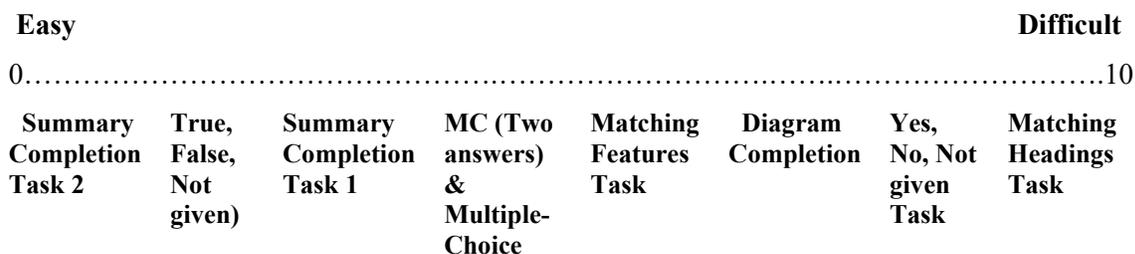


Figure. 7. 1. Experts' judgements of task difficulty of the RCM IELTS test tasks

As shown in Figure 7.1, the most difficult task was the *Matching Headings Tasks 1 and 2* which required careful reading of all the paragraphs and the text, while the *Summary Completion Task 1* which involved fill in the blanks with a word from the list of words was judged as the

easiest task of the test. The second most difficult task were the *Yes, No, Not given Task*, the *True, False, Not given Task*, and the *Diagram Completion Task* which were equally difficult. The *Yes, No, Not given Task* fell in the middle of the difficulty. Interestingly, the *Multiple-Choice Task* was also judged as an easy task. Finally, the *Matching Features Task*, the *Summary Completion Task 1*, and the *Multiple-Choice Task (two answers)* were at the same level of difficulty.

Task difficulty can be a function of the level of processing needed to do the test tasks where high level discourse-driven processes of comprehension can contribute to task difficulty. However, close comparison of the difficult level with level of processing indicated that some test tasks such as the *Multiple-Choice Task (two answers)* and the *Multiple-Choice Task* which involved high level of processing, were ranked among the moderate difficulty.

7.8. The Level of Processing

Another main classification addressed in the *Construct of the RECM IELTS Questionnaire* was level of processing involved in each type of test task. Experts were asked to judge level of processing on a scale of 1-10 (low-high) where 1 represented the form-based low level of processing at one end of the continuum and 10 represented the meaning and discourse based high level processing at the other end. Table 7.8 presents result experts' judgements of the level of processing.

Table 7.8.

Results of experts' judgements of the level of processing

Test Tasks	Level of Processing: Total score	Level of processing: Mean score	Low-High
<i>True, False Not given</i>	47	(4.7)	Low
<i>Matching Features Task</i>	44	(4.4)	Low
<i>Diagram Completion Task</i>	41	(4.1)	Low
<i>Matching Headings Task</i>	71	(7.1)	High
<i>Summary Completion Task 1</i>	41	(4.1)	Low
<i>Multiple Choice (two answers)</i>	63	(6.3)	High
<i>Multiple Choice Task</i>	45	(4.5)	Low
<i>Summary Completion Task 2</i>	39	(3.9)	Low
<i>Yes, No, Not given Task</i>	65	(6.5)	High

As results in Table 7.8 show, testing experts believed that level of processing for most tasks required low level of processing (6/9). The average level of processing, on a scale 1-10, for six test tasks was below 5 (the *True False, Not Given Task*, the *Matching Features Task*, the *Diagram Completion Task*, the *Summary Task 1*, the *Multiple-Choice Task*, and the *Summary Task 2*). This means that on average most test tasks involved low levels of processing. Only three test tasks had values >5; the *Matching Headings Task*, the *Yes, No, Not given Task*, and the *Multiple-Choice (two answers) Task*.

Level of processing values ranged from 3.9-4.7 for these tasks. Only three tasks required higher levels of discourse and text processing by attending to textual features such as coherence and cohesion. For the *Matching Headings Task* the average was 7.1 and for the *Yes, No, Not given Task* and the *Multiple-Choice Task (with two answers) Task* the average level of processing were 6.5 and 6.3, respectively. Looking back at these three tasks in the context of the test, the *Matching Heading* and the *Yes, No, Not given Task* asked for some kind of textual comprehension and evaluation of the headings and inferencing. In other words, these tasks required attending to overall meaning and structure of the text and moving beyond the literal meanings expressed in the text. Each heading and statement had to be judged against the context

in which these statements were made by relating what was stated in the prompt to some textual representation and/or background knowledge.

Based on the values assigned to the level of processes required they were rank ordered from low-high level processes. Results are presented in Figure 7.2. (See Appendix G for a copy of the Sample Test. All tasks are labelled there).

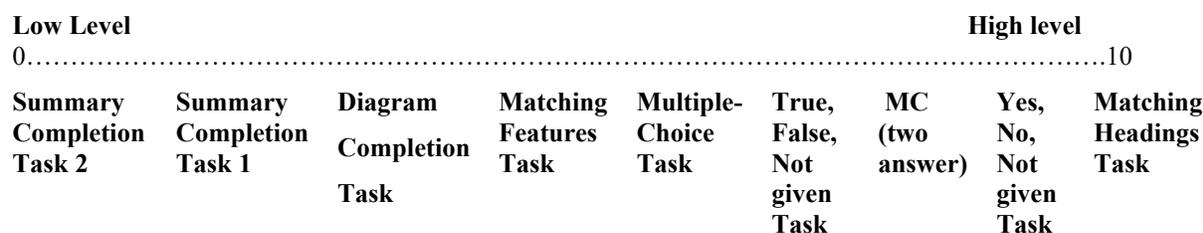


Figure. 7.2. Experts' judgements of level of processing for IELTS RCM test tasks

As shown in Figure 7.2, the *Summary Completion Tasks* were ranked at the lowest levels of processing. These two tasks involved filling in few blanks in a very short summary text. Test takers had to look for a few words in few sentences. Next, there was the *Diagram Completion Task* which was not very different from the *Summary Completion Task*. It required filling in the blanks in short sentences with two-word phrases. The only difference was that the short sentences were part of a diagram. Interestingly, tasks such as the *Multiple-Choice (with two answers)*, The *Yes, No, Not given Task*, and the *Matching Headings Task* involved reading and comprehension of larger sections from the text and involved higher levels of processing.

For the *Matching Headings*, the *Multiple-Choice (two answers) Task*, and the *Yes, No, Not given Task* there was consensus among the judges that they tapped into high level of processing. Likewise, they all agreed that the *Summary Completion Task 1* and the *Summary Completion Task 2*, involved low levels of processing. For other test tasks, however, results showed a high degree of inconsistency in the experts' judgements. For instance, for the *True, False, Not given Task*, 4 experts judged it as a low-level processing task while 6 other judges rated it as a high-level processing task. The same disagreement was also observed for the *Matching Features Task*, the *Multiple-Choice Tasks*. Judgement consistency was observed for few test tasks and for most test tasks judges had diverse opinions. Consistency in the experts' judgements can be a sign real disagreement, fuzzy nature of level of processing, difficulty of assessing it, or the need for more precise definition and more training of the judges.

7.9. Time Needed

Another classification evaluated by the testing experts was the amount of time they assumed test takers need to process each test task. They were given a list of different time options ranging from less than one minute to more than two minutes and were asked to assign any amount of time they thought is needed to do each test task. Results of experts' judgments of how much time each test item in each test task needs, is presented in Table 7.9.

Table 7.9.

Experts' Assessment of time needed for test takers' task performance

Experts	1	2	3	4	5	6	7	8	9	10	Total: Minutes
Test tasks	Time suggested										
<i>True, False Not given Task</i>	90s	2m+	90s	90s- 90s+	90s	2m	2m	2+	1m	2m	19m
<i>Matching Features Task</i>	90	90	1	1	2+	90	1	2	2	1	14m
<i>Diagram Completion Task</i>	2	1+	1	1	1	1	1+	2+	1	90	14m
<i>Matching Headings Task</i>	90	2	90	90+	2	90	1	2+	2+	2+	19m
<i>Summary Completion Task 1</i>	9	1	1	1-90	90	1	1	-1	1		12m
<i>Multiple Choice (two answers)</i>	90	2	2	1	2+	1	1	-1	1	90	15m
<i>Multiple Choice Task</i>	90- 120	1	2	1	90	1	2+	2	1	90	15m
<i>Summary Completion Task 2</i>	90	90	1	1	2	1	1	-1	1	1	13m
<i>Yes, No, Not given Task</i>	90	2+	90	90	2+	90	90	2	90	1	11m
Total											132m

As shown in Table 7.9, the time allocated fell within the range of 11 minutes to 19 minutes per task. the *True, False, Not given Task* and the *Matching Headings Task* were assessed to require the most amount of time, 19 minutes. Results indicated that the amount of time experts assigned for doing the test tasks was related to the order of the tasks. For instance, for the first tasks after each text, experts assigned more time while for the second and third tasks they assigned less time for doing the test task. They included the *True, False, Not given Task*, the *Matching Headings Task*, and the *Multiple-Choice Task*. The time assigned to these tasks was 19

minutes, 19 minutes, and 15 minutes, respectively. Experts argued that for the first tasks test takers need some more time to read the text before answering the items while for the second and third tasks after each text test takers do not need to read the text anymore. Except for these first test tasks, other test tasks fell within the range of 11 minutes to 15 minutes. A surprising finding for the time needed to do the test tasks was the total amount of time experts assigned for the whole test which was 132 minutes (twice as much as what is allowed), which is far more than 60 minutes allowed for test performance. These judgements were consistent with their judgements of the test involving “very careful reading” and “careful reading”. In other words, they were arguing that test performance requires more time because for most test tasks careful reading and very careful reading is required. This speaks to the serious challenges actual test takers face in successfully doing the test tasks within 60 minutes.

7.10. The Reading Skills Measured

Finally, testing experts judged the reading skills tapped by each type of test task. They were given a list of 18 reading skills and an open option to add any other skill they thought relevant to the test tasks. The reading skills in the list were all familiar to the experts and were used in different reading comprehension tests. Results of testing experts’ judgements of the reading skills tapped by each test task are presented in Table 7.10.

Table 7.10.
Experts' judgments of skills measured by RCM IELTS

	<i>True, False, Not given</i>	<i>Matching Features</i>	<i>Diagram Completion</i>	<i>Matching Headings</i>	<i>Summary Task 1</i>	<i>MC (two answers)</i>	<i>Multiple choice</i>	<i>Summary Task 2</i>	<i>Yes, no, not give</i>	Total	Total
1. Understanding the main idea	0	2	0	10	2	10	0	0	0	24	3%
2. Understanding details	10	8	9	8	10	0	10	10	10	75	9%
3. Sentence comprehension	5	0	8	0	2	0	5	5	0	25	3%
4. Inter-sentential comprehension	7	7	8	6	8	0	8	5	6	55	6%
5. Paragraph comprehension	2	6	3	10	8	2	2	3	8	44	5%
6. Paragraph structure knowledge	6	5	6	8	2	2	10	7	2	48	7%
7. Text structure knowledge	0	4	0	8	2	10	2	2	7	35	4%
8. Vocabulary knowledge	10	7	10	6	10	10	5	10	7	75	8%
9. Background knowledge	0	0	8	2	0	0	0	0	3	13	1%
10. Grammar knowledge	8	10	10	8	10	8	10	8	8	80	9%
11. Inferencing	8	2	0	10	0	8	4	0	8	40	4%
12. Lexical inferencing	8	5	5	2	0	2	3	0	7	32	3.5%
13. Reading speed	10	10	10	10	10	10	10	10	10	90	10%
14. Careful reading	10	10	10	10	10	10	10	10	10	90	10%
14. Skimming	5	9	4	10	7	2	5	5	6	53	6%
15. Scanning	10	6	10	10	10	10	10	10	10	86	10%
16. Text type knowledge	1	0	7	6	0	0	0	0	0	14	1%
17. Writers' attitude	0	0	0	0	0	0	0	0	7	7	1%
Total										886	100%

As shown in Table 7.10, based on the total frequencies, the most frequently identified skills were 1) reading speed 90 (10%), 2) scanning 86 (10%), 3) grammar knowledge 80 (9%), 3) understanding details 75 (9%), and vocabulary knowledge (8%). As to the least frequently selected skills, testing experts chose the following skills; understanding the main idea 24 (3%),

text type knowledge 14 (1%), background knowledge 13 (1%), and attitude 7 (1%). These skills were among the least chosen skills for the test tasks. These percentages clearly demonstrate that the form-based micro level elements and skills are dominant in the test tasks while the macro elements of meaning and text are secondary.

Two main skills that are usually targeted in almost all reading comprehension tests include comprehension of the main idea and comprehension of details. According to the experts' judgements, comprehension of the main idea was required for only (n=24) 3% of the test tasks while comprehension of details was measured by (n=75) 9% of the test tasks. This finding indicates that the focus of the sample RCM IELTS is basically on comprehension of details not the main ideas which is consistent with results of test content analysis (See Chapter Five, Section 5.3). Comprehension of details clearly relies on sentence and inter-sentence levels comprehension and does not require macro-level text comprehension. Closer examination of the findings also showed that test tasks targeted different levels of comprehension. Sentence level comprehension (n=25) 3%, inter-sentence comprehension (n=55) 6% and paragraph level comprehension (n=44) 5% together accounted for 14% of the skills measured by the test tasks while text level comprehension was reported to be (n=35) 4%. These percentages are evidence of the dominance of the lower level of sentence and inter-sentence comprehension over higher level text comprehension.

In terms of type of reading (careful- expeditious), results showed that scanning which requires a searching the paragraph or text to locate specific words and phrases accounted for 86 (10%) of the reported reading skills while skimming accounted for 53 (6%) of reading skills. These two reading skills played a key role in comprehension processes to help test takers scan specific information (e.g., words, names, and dates) in the text and skimming main ideas at paragraph level. In terms of knowledge sources involved in task performance, experts reported involvement of vocabulary knowledge 75 (8%), grammar knowledge 80 (10%), paragraph structure knowledge (7%), textual structure knowledge 35 (4%), knowledge of text type 14 (1%), and topical knowledge (1.5%). Again, low level form-based components of vocabulary and grammar accounted for 18% of the skills in the RCM IELTS, while high level discourse and meaning-based components of text and topical knowledge together accounted for 7.5% of the skills.

Interestingly, consistent with their views on the amount of time needed to do the test tasks, testing experts judged the test requiring speed reading. The experts agreed that all of the test tasks involved an element of *speededness* (Henning, 1987). Henning discusses speededness as part of a test construct in opposition to *power*, when tests allow more than sufficient time for test takers to complete them (e.g., a take home essay test). As noted earlier, the amount of time the experts identified that was required for test performance, was more than twice as much as the time officially allowed by the test.

In terms of the type of comprehension involved, experts' judgements indicated that inferencing account for 40 (5%) of the skills, and. Lexical inferencing 32 (4%), suggesting that the test tasks were tapping into inferencing at lexical and comprehension levels. Inferencing was limited to four test tasks; the *True, False, Not given Task*, the *Matching Headings Task*, the *Multiple-Choice (two answers)*, and the *Yes, No, Not given Task*, while lexical inferencing was involved in almost all test tasks (7/9).

Testing experts also commented on the skills measured by each test task and argued that a great majority of the skills listed in the questionnaire are tapped by the test takers and it may not be easy to say which ones are not measured by the task. Some of the test takers indicated that the test basically asks for low-level reading skills and high-level discourse processes are not much tapped by the test. One of them reported,

What stood out for me is that most of these questions highlight details in the text. You need to look at dates and names and specific words to answer them. Very few questions asked for the main points. (Sam)

7.11. Summary of Experts' Judgements

Finally, to have a more precise measure of the experts' judgements about different dimensions of the test tasks, their judgements were related to the number of test items in the test. Table 7.11 reports the number of items for these classifications: level of Reading level of reading (word, sentence, paragraph, text), Knowledge used (linguist and topical knowledge), type of comprehension (literal, inferential, evaluative, level of processing (low-high), type of reading (careful-expeditious) and task difficulty (easy-difficult),

Table 7.11.
Summary of experts' judgements for different classifications

	No. of items	Percentage
Level of reading		
-Word/phrase	-	10%
-Sentence and inter-sentence	-	40%
-Paragraph	-	34%
-Text	-	16%
Type of Knowledge		
-Lexical	-	32%
- Grammar: Sentential	-	21%
Grammar: inter-sentential	-	19%
-Paragraph structure	-	14%
-Textual	-	10%
-Topical	-	4%
Type of reading		
-Literal	-	55%
-Inferential	-	35%
-Evaluative	-	10%
Level of Processing		
-High level	15	38%
-Low-level	25	62%
Type of Reading		
- Very careful	11	28%
- Careful	29	72%
- Expeditious	0	0%
Task Difficulty		
-Easy	13	32%
-Moderate	20	50%
-Difficult	7	18%

As the results in Table 7.11 show, (n=25) 62% of the test tasks tapped into low level processes whereas high level processes were tapped by only 38% (n=15) of the items. In terms of task difficulty, experts judged (n=20) 50% of the items as moderately difficult and 32% (n=13) as easy and only 8% (n=7) as difficult, suggesting that the test tended toward the easy end of the continuum in their view. Finally, for type of reading, results indicated that no items were answered by expeditious reading. More than ¼ (28%) of the items involved very careful reading and the remainder involved careful reading.

7.12. Results of experts' accounts and judgements of the test tasks

To get a clearer picture of experts' judgements of each of the 9 test tasks, this section reports results of their response to the Construct of the IELTS RCM Questionnaire for each of the 9 test tasks.

7.12.1. Construct of The True, False, Not Given Tasks

The majority of the 10 testing experts agreed on the construct of the *True, False, Not given Task*. The numbers in parentheses show how many experts (out of ten) shared the same idea. The experts' accounts and judgements of the *True, False, Not given Task* also indicated that the task targets literal (10/10) and inferential (8/10) comprehension of specific details (10/10) which require sentence level (5/10) inter-sentence level (7/10) comprehension. Task performance was facilitated by using lexical clues and vocabulary knowledge (10/10). Experts also believed that scanning skill (10/10) and awareness of paragraph structure (6/10) can facilitate performance. The task also measures speed reading (10/10). Again, it is worth noting that the numbers in parentheses indicate the number of testing experts who opted for these skills. Ten testing participants had participated in this phase of the study.

7.12.2. Construct of The Matching Features Task

Testing experts believed that the *Matching Feature Task* measures literal comprehension of specific details (8/10) at inter-sentential (7/10) and paragraph levels (6/10). They also thought grammatical knowledge (10/10) vocabulary knowledge (7/10), awareness of paragraph structure (5/10), and text structure (4/10) can help processing the task. Finally, they indicated that the task can be achieved by careful reading (10/10), skimming (9/10) and some scanning skill (6/10).

7.12.3. Construct of The Diagram Completion Task

Based on the accounts and judgements of the testing experts, the *Diagram Completion Task* measures vocabulary knowledge (10/10) in the context of some specific details (9/10) at sentence (8/10) and inter-sentential (8/10) levels. Background knowledge, awareness of paragraph structure (6/10), and vocabulary knowledge (10/10) are also tapped by the task. The task can be achieved by careful reading (10/10), scanning (10/10) and skimming skill (4/10). Testing experts also indicated that reading speed skills are tapped by the task.

7.12.4. Construct of The Matching Headings Task

Experts indicated that the *Matching Headings Task* taps inferential comprehension (10/10) of both the main idea (10/10) and details (8/10) at inter-sentential (6/10) and paragraph levels (10/10). The task also involves knowledge of paragraph structure (8/10) and text structure (8/10). They also indicated other skills such as vocabulary knowledge (6/10) and grammar knowledge (8/10) are conducive in task performance. They also indicated that the task can be better achieved by skimming (10/10), scanning (10/10), and careful reading (10/10).

7.12.5. Construct of The Summary Completion Task 1

Experts' accounts and judgements of the *summary completion task 1* indicated that it involves literal comprehension of specific details (10/10) at inter-sentential (8/10) and paragraph (8/10) levels. Test performance also involve use of vocabulary and grammar knowledge. Task performance can be achieved by skimming (7/10), scanning, and careful reading (10/10). The test also taps speed reading (10/10). Most of the experts' accounts and judgements of *the Summary Completion Task 1* matched the processes and strategies observed in the test takers' performance.

7.12.6. Construct of The Multiple-Choice (two answers) Task

Experts' accounts and judgements of the *Multiple-Choice (two answers) Task* indicated that the task taps into literal (10/10) and to a lesser degree inferential (2/8) comprehension the main idea (10/10) of the text. Knowledge of text structure (10/10) in addition to vocabulary (10/10) and grammatical knowledge (10/10) were believed to be conducive in processing the test task. According to the experts the task does not tap comprehension of details while test performance of the test takers showed that comprehension of details played a major role in choosing the right options.

7.12.7. Construct of The Multiple-Choice Task

Experts' accounts and judgements showed that the *Multiple-Choice Task* measured specific details (10/10) at sentence (5/10), inter-sentential (8/10), and paragraph (10/10) levels which means different levels of comprehension (sentence-paragraph) and types of comprehension (literal-inferential) were tapped by *the Multiple-Choice Task*. The test task also involves use of vocabulary (5/10) and grammar knowledge (10/10). The test task, they indicated, can be achieved by careful reading (10/10) skimming (5/10) scanning (10/10). All experts also

agreed that the last item measured comprehension of the main idea at text level and called for textual knowledge and awareness.

7.12.8. Construct of The Summary Completion Task 2

Based on the experts' account and judgements, the *Summary Completion Task 2* measured vocabulary knowledge at inter-sentential (6/10) and paragraph (8/10) levels. Paragraph structure knowledge (7/10), scanning (10/10), careful reading (10/10), and skimming (5/10) skills were involved in test processing. The task is basically concerned with literal comprehension which involves matching two similar sentences one in the summary text and one in the main text.

7.12.9. Construct of The Yes, No, Not given Task

Testing experts believed the *Yes, No, Not given Task* mainly measured specific details (10/10) at inter-sentential (6/10) and paragraph (8/10) levels. The task also measured inferential meaning (8/10). Moreover, experts believed that test items include several lexical clues (7/10) and answering them requires lexical inferencing. Skimming (6/10), scanning (10/10), and careful reading (10/10) was used in processing test items. In doing the task, knowledge of text structure (7/10) was also tapped by the test task. Like all other test tasks, speed reading was key in doing the task within the allowed time limit

7.13. Is IELTS RCM a measure of academic reading?

IELTS In conclusion, based on the accounts and judgements of the experts, the IELTS RCM does not represent academic reading skills and cannot be regarded as a measure of academic reading. In addition to the disciplinary variation in terms of the way academic reading is practiced across disciplines such as chemistry, engineering, medicine, and social sciences, the multitude of texts and tasks in the IELTS RCM can be very confusing to the test takers. One of the experts argued that

There are several texts and test tasks, each with a different format. This is really confusing and tough if you have not attended the IELTS preparation classes and courses. You cannot make it. (Sam)

Another expert raised the same issue and said,

I think the multitude of task formats increases the potential error due to misreading of the instructions for those who can hardly afford to pay for the preparation programs. (Fanj)

When asked what the IELTS RCM measure and if it measure academic reading, most expert (8/10) said, “no”. One of them commented,

If I am a second language reader and I can read and do well on these tasks, I do not think I am a good academic reader. It just shows I am test-wise reader. (Jala)

The same point was reiterated by another expert who said,

If non-native test takers get 8 or 9 out of this test, it indicates they are not reading. They are exercising strategies. They are test-wise skills not reading skills, not actual reading. If a student has this level of reading, they can probably navigate and do what they need to read in their academic work, but this is not a measure of academic reading. It is just a measure of reading. (Fanj)

She also indicated the test tasks are really challenging and very different from actual academic reading. She reported,

I am in a position of confidence. I am a good academic reader and I am a native speaker of English and I may not get 7/9 on this? I am above proficient in English but struggle doing this task. (Fanj)

Some testing experts also rejected the test as a measure of academic reading because the texts and test tasks had no context. They were isolated from other language activities and practices. They indicated that reading is part of a network of academic activities and language-related activities. One of them reported’

The test is just reading few texts followed by some factual questions. This might be what we do in ESL classes. In Academia, we read to and write. We read and present. We read and ask. Reading is integrated with other skills and activities. (Rahi)

Another expert also mentioned this point and reported,

Your writing draws on what you hear in the lecture and what you read and that is why I say academic reading academic writing, academic listening and academic speaking are fully integrated. Once you begin to tease them apart into micro elements you lose the purpose.
(Fanj)

7.14. Summary of the Chapter

This chapter presented results of testing experts' judgements and accounts of the IELTS RCM test tasks. They assessed each task in terms of several classifications including the level of reading, the type of knowledge needed, the type of comprehension involved, the type of reading, the level of processing, text and task difficulty, and the skills measured. Results indicated that the sample RCM IELTS basically tapped low-level literal comprehension at sentence-paragraph levels. High-level textual comprehension and inferential meaning were secondary. The chapter also presented results of experts' responses to each of the nine test tasks.

Chapter Eight discusses results of the three sources of evidence in light of Khalifa and Weir's (2009) model of reading comprehension and the relevant literature.

CHAPTER EIGHT: DISCUSSION

8.1. Introduction

This exploratory case study was motivated by personal experience and the existing gap in research related to the construct validity of the Academic IELTS RCM as a high-stakes English language proficiency test. The study adopted unified validity theory (Messick, 1989) and explored cognitive evidence for the validity of inferences drawn from the IELTS RCM based on the reading comprehension construct model proposed by Khalifa and Weir (2009). Evidence was drawn from three different sources: 1) test content, 2) test takers' processes and strategies, and 3) experts' accounts and judgements. Four research questions guided analysis of the validity evidence. As illustrated in Figure 8.1, research question one, which was addressed in Chapter Five, examined content validity evidence for the RCM. The second research question explored the cognitive processes used by test takers during RCM test performance by examining retrospective verbal accounts of the skills, knowledge sources, processes, and strategies (SKSPs) (Gorin, 2006) observed or reported by three different groups of test takers during RCM test performance. Results of the second research questions were presented in Chapter Six. The third research question examined testing experts' judgements and accounts of the SKSPs and other aspects of the RCM test construct. Research findings addressing this research question were discussed in Chapter Seven. The fourth research question looked for congruence of findings across the three sources of evidence and in relation to the existing literature on the RCM. This is what is discussed in the present chapter.

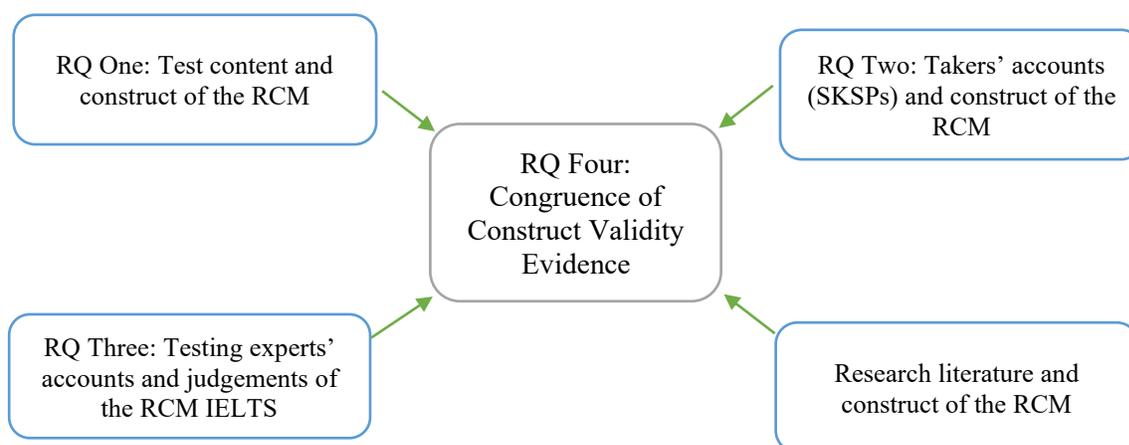


Figure 8.1. Schematic representation of the research questions

The current study adopted Messick's (1989) unified validity theory which hypothesized construct validity as the unified concept that integrates empirical evidence and theoretical rationales to support the adequacy and appropriateness of the inferences made, and actions taken based on test scores. The theory allows integration of multiple sources of evidence under a generalized notion of construct validity as one single validity argument with multiple lines of evidence. The theory recognized multiple sources of evidence as "complementary forms of evidence to be integrated into an overall judgment of construct validity" (p.37). Unified validity proved helpful for studying the construct validity of the IELTS RCM. The current study probed into three independent but inter-related sources of evidence and provided detailed evidence from these sources to argue for the validity of the IELTSRCM. There has been no study of the IELTS RCM that integrates three different sources of evidence into a unified argument to examine its test construct. Unified validity theory justified the use of these three sources of evidence. Multiplicity of the sources used in the study allowed their cross examination and triangulation, which in turn resulted in more robust research findings.

The study also adopted Gorin's (2006) process-oriented model of construct validation which suggested a cognitive approach to test validation and recommended probing into the skills, knowledge sources, processes, and strategies (SKSPs) that are used during test performance. Drawing on her model, this study used retrospective verbal reports for examining the IELTS RCM construct. This provided information about the internal dynamic of the IELTS RCM and how test takers read and performed on each type of test task. The cognitive approach adopted in Gorin's (2006) model also contributed to the study by providing a framework for examining validity evidence that could not have been accessed otherwise.

8.2. Discussion

In this section, results of the study are discussed in light of Khalifa and Weir's (2009) comprehension model, cognitive reading theories, and empirical findings related to the construct of the IELTS RCM. Khalifa and Weir (2009) suggested three components for reading processes: base component (cognitive processes), metacognitive component (metacognitive strategies), and knowledge base (knowledge sources) component (See Sections 8.3. for discussion of the model). First, results of content analysis of the IELTS RCM test content are discussed. Next, results of

the study are discussed in relation to the three components of Khalifa and Weir's (2009) model. Then, findings are discussed in relation to other reading theories and empirical findings. However, before discussing the results, I would like to reiterate that results from the three sources of evidence (i.e., test content analysis, test takers' accounts of the processes and strategies they used, and experts' accounts and judgements) were generally consistent and did not show meaningful differences. Test content analysis and experts' accounts and judgements both indicated that the test tapped into local literal comprehension at low sentence and inter-sentence levels. Few test tasks tapped inferential and higher levels of comprehension at the text level. L2 test takers' accounts also showed that they mostly focused on answering the test tasks instead of comprehending the texts and selectively searched the text at sentence and inter-sentence levels to get to the relevant information and the correct answer.

Results of the test content analysis (See Chapter Five) indicated that the IELTS RCM mostly focused on literal comprehension at low-sentence and inter-sentence levels. Paragraph level comprehension was moderately measured by a few test items (e.g., the *Matching Headings Task*), but the test failed to represent some features of academic texts and academic reading activities (e.g., text length, text organization, and lexical features). In Khalifa and Weir' (2009) terms, these findings indicate that only low-level processes of central core components (i.e., cognitive processes) such as word recognition, syntactic parsing, development of semantic proposition and sentence level comprehension, are tapped by the test tasks. Higher level cognitive processes such as inferencing, developing text representation, and mental representation were only partially tapped by the test tasks. These findings resonate with Weir, Hawkey, Green, and Devi's (2009) findings that suggested gaps in the IELTS RCM test content and academic reading material. Results of content analysis of the sample IELTS RCM test call the validity of inferences drawn from the RCM into question because it provided only partial evidence of academic processes involved in academic reading comprehension.

Results of the content analysis of the test also showed that they did not tap features of "extensive reading" (Brown and Abeywickrama, 2004), which represent key features of academic reading. As noted in Chapter Three, they defined extensive reading as reading that involves high levels of processing with a focus on discoursal and topical aspects of the text in professional settings. Based on the content analysis reported in Chapter Five, almost all test tasks on the IELTS sample test used in the study represented either selective reading or interactive

reading dimensions (Brown and Abeywickrama, 2004; See also Chapter Three, Section 3.3.2), which mostly focus on low level literal comprehension based on analysis of vocabulary and grammar.

Assuming that one basic feature of academic reading skill is to “gain both a local and general understanding of the passage” (Cohen & Upton, 2006, p. 117), results of the content analysis of the test showed that the test focus was mainly on local comprehension, suggesting underrepresentation of the construct (Messick, 1989) of academic reading. At best, the test seems to measure only skills that test takers may need for learning academic reading.

Taylor (2007) argued that measuring academic reading skills can occur only in the context of the test takers’ specific discipline. This view of the academic reading construct assumes that due to variation in genres and topical knowledge across disciplines, test takers can master discipline specific academic reading skills only after they enter university and engage in their disciplinary programs. Assuming that academic reading is learned after students enter their academic programs, it seems that tests of academic reading such as IELTS RCM would need to assess test takers’ readiness to face the challenges and demands that will occur in their disciplines. It is worth noting that the results of testing experts’ judgments of the IELTS RCM test tasks (See Chapter Seven) were also consistent with findings of test content analysis and findings from the processes and strategies reported by the L2 test takers. All indicated that most test tasks and items tapped into local literal comprehension at sentence and inter-sentence levels. Altogether these findings can lend only partial support to the construct validity of the IELTS RCM as a rather limited measure of local but not global comprehension.

IELTS RCM adopts a global proficiency approach to testing language skills. Test score results are used to assess undergraduate and graduate student academic skills which are required for entering academic programs. A language proficiency approach is a global “crude and barbaric measure of language” (Burno & Zumbo, personal communication). In the last two decades other pathways to assess the required language skills to enter academic programs and university education have been developed. For example, Post Admission Language Assessment (Also known as Post Entry Language Assessment, PELA) suggests an alternative approach which is diagnostic in nature and adopts a narrow, context specific and disciplinary approach to

academic language assessment (Fox & Artemeva, 2017; Fox, Haggerty, & Artemeva, 2016; Read, 2015, 2016).

8.2.1. Khalifa and Weir's (2009) model of reading comprehension

Results of the verbal reports of the processes and strategies used by the three groups of test takers provided further evidence of the construct operationalized by the IELTS RCM in relation to Khalifa and Weir's (2009) reading comprehension model. The *central core component* of the model, which addresses cognitive processes of reading, postulated a set of low-level bottom-up and high-level top-down cognitive processes involved in reading. Results of SKSPs used by the three groups of test takers indicated L1 test takers used both low-level cognitive processes of word recognition, sentence parsing, formation of propositional meaning, and high-level processes of inferencing, developing text representation and mental models. These test takers applied the text representation they had developed during reading the text and used it as a reference for answering some test items. For the L2 test takers, however, the processes were mostly low-level bottom-up processes. Unlike the L1 test takers, L2 test takers focused on doing the test items rather than comprehending the whole text and developing a text representation. All they were concerned with was completing the test tasks by searching for segments of the text in order to locate the relevant information.

In terms of the *knowledge-base* component of Khalifa and Weir's (2009) model, results also indicated use of topical knowledge and different types of language knowledge sources. Test takers frequently reported use of these knowledge sources in their accounts. Some test tasks, such as the *Diagram Completion Task* and the *Yes, No, Not given Task*, required use of background knowledge. Test takers needed some technical knowledge of engineering for clear understanding of the information related *raising a sunken ship* from the seabed. Knowledge about engineering projects could facilitate comprehension. Further, vocabulary knowledge was key to the test performance of test takers. Most L2 test takers, especially the less successful L2 test takers, complained about the new vocabulary items they encountered in the text which stopped them from comprehending different sections of the text, while the successful test takers frequently referred to the use of lexical clues in the test items and the text as a resource for choosing the right answer. In addition, textual knowledge which was mostly used by the L1 test takers was very conducive to helping them recall some of the information they had read to answer some items. Verbal reports of the test takers also showed that some test tasks appealed to

their background knowledge as part of processing the test tasks and used it in answering test items. In a few test tasks such as the *Diagram Completion Task* and the *Yes, No, not given Task*, use of background knowledge was more frequently mentioned and used in test performance processes. Furthermore, topical knowledge of the L1 test takers was a great asset to them in reading and comprehending the text, while almost all L2 test takers were not familiar with the text topics at all.

Results also supported the metacognitive component of Khalifa and Weir's (2009) model. All test takers used a number of different strategies either for reading the texts or doing the test tasks. Some strategies such as "careful analysis of the options" or "re-phrasing or rereading the question" were specifically related to reading and processing the texts, while some other strategies, such as "paying attention to details" or "attention to certain features of the text", were used for doing the test tasks. In fact, test performance for all test takers was a strategic activity. Some test takers clearly expressed their awareness of using an appropriate strategy for reading and answering a particular test task or test items. "Planning a strategy", as indicated by some test takers in their verbal reports, was a clear indication of metacognitive awareness.

Depending on the task demands, test takers' language proficiency, and reading proficiency, they assessed a test task and planned which strategies to use, which they then implemented by using different metacognitive strategies. Findings also lent support to Khalifa and Weir's (2009) argument for the role of the metacognitive component serving goal setting, goal checking (monitoring), and remediation purposes. Test takers used a number of different strategies to manage their task performance. Depending on the task features, they used different types of reading, skimmed different sections of the text, scanned different key words and phrases, and carefully read different parts of the text to locate and process the relevant information and get the answer to each test item. Some of the strategies they used were related to comprehension of the text, while other task management strategies and test-wiseness strategies were used to enhance their test performance.

Results can also be discussed in light of the strategic component of Bachman and Palmer's (2010) model of communicative language use. They hypothesized that strategic competence serves goal setting, planning, and appraisal purposes. Test takers did not consistently use one specific strategy across test tasks. They executed each task based on the task demands and appraised their performance during test performance. In many cases, they had to "delay

answering” the items by appraising their performance, “go back to the text” and “re-read parts of the text”, or “change their answers”. They also re-read some test items that were ambiguous to them and “rephrased” them to clarify the task at hand.

Findings of the cognitive processes and metacognitive strategies reported by the test takers are consistent with Pearson (2009), who reported the use of metacognitive strategies by test takers—especially L1 test takers and more successful L2 test takers during test performance—served as a problem solver to repair their comprehension failures and maximize their test performance. Furthermore, Pearson’s results showed that L1 test takers and the successful L2 test takers used more metacognitive strategies, which is consistent with the general consensus that strategic awareness and monitoring skills distinguish good skilled readers from poor unskilled readers in their meaning-making processes (Grabe, 2009; Grabe & Stoller, 2002; Paris & Jacobs, 1984) and this affects their test performance (Carrell, 1989; Phakiti, 2003, 2008; Salehi, 2011; Saraswati, 2015).

Further, results of L1 test takers can be best justified by discourse comprehension model (Kintsch 1998; Kintsch & Rawson, 2005) that suggests text base and situation model of comprehension. Test performance processes of the L1 test takers and some of the more successful L2 test takers also lent support to Kintsch’s (1989, 1998, 2005) text base and situation model of text comprehension processes (See Chapter Three, Section 3.4.2). They proceeded through the text by careful reading of all paragraphs and sections of the text to develop a representation of the text. They also related the texts or parts of the text to their experiences and background knowledge when necessary. What they did in terms of SKSPs was consistent with Kintsch’s model which characterizes text base as a representation of those propositions that readers/test takers can directly derive from the text, while the situational model, which supplements the text base, requires background knowledge and experience that is retrieved from memory. In terms of processing, text base can be seen as bottom-up processing whereas the situation model is basically top-down processing. Results of the L2 test takers can be justified in light of Construction-Integration Theory (Kinstch, 1988; Kintsch, & van Dijk, 1978) which argues for bottom-up processes of meaning making. For some of the successful L2 test takers, the processes were basically limited to text base representation. The less successful L2 test takers struggled even with bottom-up processing of sentences and could not develop a text representation and proceed to a higher level of developing a situation model. The processes and strategies they used were isolated, text-driven and task-oriented, while the L1

test takers used inter-related processes that were comprehension-oriented. These findings, as mentioned earlier, were consistent with the results of content analysis and testing experts' accounts that indicated the test tasks focused on low level comprehension and only few test items required developing a situation model.

On a different note, regarding the use of cognitive processes and metacognitive strategies, results suggest that test takers used both the cognitive processes and the metacognitive strategies. Use of different SKSPs seemed to have helped them enhance their performance on the test. Use of the processes and strategies and different knowledge sources were not separate phases of test performance, rather, they were one unitary, inter-related, and integrated activity. However, test takers' use of the processes, strategies, and knowledge sources seemed to vary across the context of test performance. Depending on the variation of text topics, test tasks, and individual differences related to their language backgrounds and levels of language proficiency, test takers used different cognitive processes (processes, strategies, and knowledge sources).

These findings are consistent with those of researchers (e.g., Chapelle et al., 1997; Vandergrift, Goh, Mareschal, & Tafaghodtari, 2006) who found variable use of strategies in comprehension processes. In addition to the time constraints of the test, test takers' choices of processes and strategies was highly inter-related to both the reading texts and task demands. In fact, the variation and distribution of the processes, strategies, and knowledge sources used during test performance can be justified in the context of constraints of time, text, and task. This aspect of test performance is also theoretically supported by Phakiti (2008) and Purpura (1999), who argued for a continuum of processes and strategies that are variably used in reading comprehension and reading test performance. Findings related to the use of different SKSPs are also consistent with cognitive models of reading comprehension, which view reading as a multi-componential skill involving use of different processes, strategies, and knowledge sources (e.g., Kintsch, 2005; Kintsch, & Rawson, 2005). In addition, results resonate with the *Simple View of Reading*, which conceptualizes comprehension as a combination of word decoding abilities and text comprehension abilities (Gough & Tunmer, 1986).

Results from the test performances of the test takers also showed that L2 test takers focused on "doing the test tasks" not "reading for comprehension". Test performance mainly revolved around task-related strategies and not comprehension-related strategies. Their use of processes and strategies were mostly for task management purposes not comprehension

purposes. Hence, it can be argued that the IELTS RCM failed to tap the main purpose of testing academic reading comprehension, which is assumed to be about reading for comprehension rather than reading for answering specific questions. Most L2 test takers treated the task as a test-taking task not as a normal reading task where they read and attend to meaning from different parts of the text for clear comprehension. Furthermore, they did not read to learn from the text, which is what readers do in academic reading. Arguably, the processes engaged for academic reading may be underrepresented (Messick, 1989) by processes used for responding to test tasks, especially if the task focus is on literal comprehension of specific details at sentence and inter-sentence levels. This finding is consistent with Cohen and Upton (2006) who studied the reading section of the TOFEL and found that the strategies used by the participants were primarily test-management or test-wisness strategies rather than reading comprehension strategies. However, in the present study, the processes and strategies used in the IELTS RCM lend partial support to the validity claim of the test as a *few* of the successful L2 test takers (although not *all* successful L2 test takers) read the text to gain both a local and general understanding of the texts to answer the test tasks. In other words, they read the text and tried to comprehend the main points of different sections and the whole text before they answered the items.

In terms of test-wisness strategies, which can introduce construct irrelevant variance (Messick, 1989), results of the current study showed test-wisness strategies were not key in test takers' performance. L1 test takers tended to rely on comprehension strategies for answering the test tasks. L2 test takers, on the other, hand, used strategies to manage test tasks but did not rely on test-wisness strategies as the main source for test task performance. These findings are consistent with Cohen and Upton (2006) who found that test-wisness strategies played a minor role in test performance. However, other studies (Sheeham & Ginter, 2001; Rupp, Ferne, & Choi, 2006; Ozuru et al., 2008) indicated that that some test takers successfully relied on test-wisness strategies in their test performance. These inconsistencies need to be further studied to identify test features that involve test-wisness strategies.

8.2.2. Inferencing

As suggested by Khalifa and Weir's (2009) model of reading comprehension, inferencing is one of the key high-level processes of comprehension. Inferencing involves building a mental representation based on the readers' own topical knowledge and the information presented in the text (Bowyer-Crane, & Snowling, 2005; Goldman, McCarthy, & Burkett, 2015). It helps readers

use their background knowledge and fill some gaps related to information that is implicitly presented in the text (Graesser, Wiemer-Hastings, & Wiemer Hasti, 2001; Hammadou, 199, McKoon, & Ratcliff, 2017). Inferencing helps readers/test takers to bridge the gaps and explicate implicit information in the text. Using their background knowledge, readers/test takers need to move beyond comprehending the text and infer meaning based on assumptions and deductions related to the specific topic they are dealing with. In the current study, results from all the three sources of validity showed that the IELTS RCM sample tapped inferencing. Results of the test content analysis indicated that the IELTS RCM dominantly tapped literal comprehension. Inferential comprehension was also tapped by some test items in different test tasks, but it was secondary and fewer items tapped inferential comprehension. At least one item in the *True, False, Not given Task*, the *Matching features Task*, the *Matching Headings Task*, the *Multiple-Choice (two answers) Task*, the *Multiple-Choice Task*, and the *Yes, No, Not given Task* involved inferencing. Test takers' performance also indicated that L1 test takers and more successful L2 test takers frequently used inferencing in different test tasks. L1 test takers used inferencing twice as much as the successful L2 test takers and five times as much as the less successful L2 test takers. Interestingly, the less successful L2 test takers used it the least, indicating their inability to make inferences. Inferencing was more frequently used in the *Matching Headings Task*, and the *Yes, No, No given Task*. Furthermore, for some test items test takers used lexical inferencing to relate the test item to the text. Testing experts' judgements also showed that a number of test tasks and items involved inferential comprehension, supporting findings from the other two sources. Findings from these three sources of evidence showed that the IELTS RCM involves measuring inferencing, which is viewed as a high order process by Khalifa and Weir's (2009) model. In this respect, the test can reflect the test takers' ability in inference making, which is certainly key to academic reading.

Differential performance by different test takers in using inferencing is consistent with studies that find poorer performance on inferential comprehension questions compared with literal comprehension (e.g., Pearson, Hansen, & Gordon, 1979). Literal questions, which require matching information from the text with information in the question, while inferential question requires more processing, such as remembering the passage, recalling information from background knowledge, and most importantly integrating across these sources of information to derive an answer. current literature also indicates that inferencing is a much harder process

(Perfetti, Landi, & Oakhill, 2005). Inferencing processes are harder for poor readers whose reduced working memory stops them from integrating ideas (Carretti, Borella, Cornoldi, & De Beni, 2009; Catts, Adlof, & Weismer, 2006; Oakhill, 1982). Further, poor vocabulary knowledge and lack of background knowledge (Cain, Oakhill, & Elbro, 2003; Nation, Clarke, Marshall, & Durand, 2004; Nation, Cocksey, Taylor, & Bishop, 2010; Nation & Snowling, 1998).

8.2.3. The types of reading

Khalifa and Weir's (2009) model of reading comprehension suggested two distinct types of reading: careful reading and expeditious reading. Careful reading involves reading every section of the text to get an accurate comprehension of the text, while expeditious reading involves skimming and scanning the text to locate the relevant information needed to answer the test items. Khalifa and Weir (2009) argued that depending on the purpose of reading, the features of the test task, and the readers'/test takers' level of reading skills, readers may choose to read a text carefully, or they may read the text expeditiously, L1 test takers and some successful L2 test takers used both careful reading and expeditious reading in their test performance. They systematically read the whole texts carefully before answering the test items and then they read it expeditiously by skimming and scanning the text for locating the relevant information to find the answer or to double check their answer. In other words, careful reading was used for reading the whole text and expeditious reading was reserved for searching the relevant information and double checking the answers to the test items. This indicates that careful reading was the dominant type of reading. All test items were answered in light of the text structure that had developed through careful reading of the text.

Results of L2 test takers' performance, on the other hand, as indicated in their verbal accounts, suggested that the dominant type of reading involved in test performance was expeditious reading. The L2 test takers, especially the less successful ones, were quite aware that they were reading for a test. They were also very aware of the speededness of the test and the time pressure to complete the test items within the limited time. This did not allow them to opt for careful reading of the whole text as L1 test takers and some of the more successful L2 test takers did. Instead they opted for scanning the relevant sections and sentences that seemed to contain the relevant information. For them, the starting point was reading the test items—not the text. Since they lacked the skimming skill to get the gist of the text, they expeditiously scanned the text for key words and phrases to locate the relevant information. However, they carefully

read the relevant section, and in many cases, they had to re-read it several times. A key point that needs to be highlighted is that careful reading was not less important in the L2 test takers' reading, but rather it was embedded in the context of selective reading and not reading the whole text. It differed in terms of its scope and dominance. It can be concluded that both careful and expeditious reading were used but they differed in their scope and dominance. Careful reading was dominant in L1 reading, which guided test meaning-based test performance. For L2 test takers, especially the less successful ones, careful reading was used in the context of reading only specific sections of the text and not the whole text.

Results of test takers' performance also indicated that expeditious reading was essentially used to locate the relevant information in the text. This seems to differ from the way expeditious reading is practiced in academic reading. It seems that in academic reading skimming is used as a way to get the gist of the text. It is used as a shortcut to text comprehension, while in reading in a test context, test takers used skimming and scanning to locate the relevant information in the text. Once the relevant text was located, the test takers then carefully read that text. In other words, skimming and scanning—which represent expeditious reading—served different functions in academic reading versus reading in test context.

Some research findings regarding the IELTS RCM have suggested that academic reading skills involve both careful and expeditious reading (Cohen & Upton, 2006; Hawkey, 2006; Weir, Green, Hawkey, Maniski, Devi, Unaldi & Zegarac, 2007; Rosenfeld, Oltman & Sheppard, 2004). In the present study, findings related to the L1 test takers are in line with Weir, Hawkey, Green, Unaldi, and Devi (2009) who found the preponderance of careful reading (77%) over selective expeditious reading but contradict L2 test takers' performance, which was guided by expeditious reading. On the other hand, results from L1 test takers of the current study contradict Weir, Hawkey, Green, and Devi's (2009) findings that showed that selective expeditious reading played a more important role in answering the test items. However, results of the L2 test takers in the current study which indicated dominance of expeditious reading resonated with Weir, Hawkey, Green, and Devi's (2009) and Weir, Hawkey, Green, Unaldi, and Devi (2009) who reported expeditious reading was more relevant and appropriate to academic reading tasks at university but inconsistent with results of L1 test takers. Finally, testing experts' judgements also showed that the IELTS RCM is dominated by very careful and careful reading and fell short of satisfying expeditious reading as one key component of academic reading. This supports Weir,

Hawkey, Green, Unaldi, and Devi's (2009) findings that reported a dominance of careful reading.

Results from L1 test takers, but not from the L2 test takers, also resonate with other research findings related to the type of reading used during IELTS RCM test performance. For example, Sarojani (2011) found that IELTS RCM items were imbalanced in terms of the type of reading involved. Most test items in the large sample of 14 IELTS RCM required careful reading, while expeditious reading was much less used by the research participants. The same results were also found for other high stakes reading comprehension tests. For instance, Katalayi and Sivasubramaniam (2013) found that there were more items that targeted careful reading than those that targeted expeditious reading in the high-stakes multiple-choice test of reading that was used for university admission purposes in DR Congo. In brief, over-representation of items involving careful reading in the IELTS RCM seem to contradict research findings concerning the importance of expeditious reading as one of the main features of academic reading needs of undergraduates' in accomplishing their academic reading tasks at the tertiary level.

8.2.4. Local-global comprehension

With regard to another key distinction of Khalifa and Weir's (2009) model, i.e., local-global comprehension (See Chapter Three, Section, 3.2.2), the results of the content analysis of the test tasks, test takers' accounts of the test performance processes used, and testing experts' judgements were all consistent and indicated that the test focus was on local comprehension of specific details at the sentence, inter-sentential level, and paragraph level. Except for few test items such as the *Multiple-Choice (two answers) Task* and one item in the *Multiple-Choice Task* which required global comprehension, most test tasks and items such as the *True, False, Not given Task*, the *Matching Features Task*, the *Summary Completion Task 1 and 2*, etc. called for local comprehension. Global comprehension at text level was extremely underrepresented by test tasks and items. Only 3/40 items tapped text level comprehension. Furthermore, most of the metacognitive strategies test takers used were used for task management but not comprehension of the whole text as a unit. This provides further evidence for the strategic and test-specific nature of their test performance. These findings are consistent with Weir, Hawkey, Green, and Devi (2009) who found that participant test takers put information together across sentences, suggesting that sentence level comprehension dominated the test. Findings from the present study are in line with another conclusion of their study regarding validity of the test. Weir,

Hawkey, Green, and Devi (2009) concluded that the test provided only partial evidence of the academic processes involved in academic reading comprehension. Their findings indicated a fair degree of difference between requirements for tasks in the actual academic domain and test tasks on the test. To better represent the academic reading construct, the test tasks need to go beyond the local level of comprehension and expand comprehension levels to the text level by adding tasks that specifically require attention to global comprehension.

Furthermore, results from the content analysis phase and testing experts' judgements also resonate with Moore, Morton, and Price's (2009) study, who reported that the majority of the IELTS RCM test tasks involved local-literal comprehension, while a sizeable proportion of real (i.e., like those that actually happen in the classroom in the university) academic tasks require more global-interpretive engagement. In line with all the relevant studies conducted on the construct of the IELTS RCM, certain changes need to be made in the test, the most important of which is including items that tap global comprehension and inferential interpretive items.

On a different note, as most test items targeted local comprehension at sentence and inter-sentence levels, test performance involved lexico-grammatical analysis of relevant sentences which required good vocabulary knowledge. As mentioned earlier, details of test takers' processes showed that vocabulary knowledge played a key role in lexico-grammatical analysis of the relevant sentences. Test takers relied heavily on lexical clues and synonyms in the text and stems during their test performance. These results are consistent with Chen and Liu (2020) who explored the role of vocabulary breadth and depth in the IELTS RCM. They found that both vocabulary breadth and depth correlated significantly with IELTS reading test scores, more specifically with the *True, False, Not Given Task*, the *Multiple-Choice Task*, the *Matching Headings Task*, and the *Summary Completion Task*. Results from a multiple regression model also indicated large vocabulary size was necessary for improving certain IELTS band scores. Results are also consistent with previous research that has supported the role of vocabulary breadth in reading comprehension (Akbariam & Alavi, 2013; Alavi & Akbarian, 2012; Chiang, 2018; Laufer, 1992) as well as the relationship between vocabulary depth and reading comprehension (Li & Kirby, 2015).

In conclusion, the available literature on the IELTS RCM suggests that test tasks focus on measuring literal comprehension at the local level of sentence and inter-sentence. Findings from the current study also supported these findings. However, based on the "searching theme"—one

of the three main themes identified as a result of the coding of test taker and expert accounts—findings from the current study expand previous findings by arguing that the “search process” used by the test takers went beyond local sentence level processing. As indicated in Chapter Six (See section, 6.5), most test takers did not directly or immediately answer the test items. They had to go back to the text and search for the relevant information and get the answer. The dominant search level the test takers used was paragraph level search, followed by searching several paragraphs. Sentence level search was only third in ranking. This clearly indicates that for most test items that targeted local levels of comprehension, test takers had to read larger stretches of text to locate the relevant information. In fact, searching for and locating the relevant information for these local items that were dominant in the test tasks required reading a paragraph or a couple of sentences. During the search processes, L1 test takers and some of the successful L2 test takers also used skimming which is basically a global strategy to search for relevant information. Results of search processes clearly indicated that searching for relevant information tended to be more global and was not limited to reading one or few sentences to get the answer. Test takers had to process a couple of sentences, a whole paragraph, or a few paragraphs to locate the relevant information. Therefore, it is important to consider the global nature of searching as part of the main cognitive processes of test performance across local and global test tasks.

This has bearing on the methodology of testing researchers in examining the level of comprehension involved in each test item. In examining the level of comprehension and level of processing for different test items, testing researchers need to consider the actual test performance processes and strategies used. Relying on content analysis of the test items does not provide enough context to examine the level of processing involved. Content analysis looks at the test items in abstract, while examining the actual processes can provide information on the dynamic interaction of the test taker with the test item and has more potential to show the actual level of comprehension and level of processing.

Results of testing experts’ accounts regarding levels of processing indicated that 6/9 test tasks required low levels of processing that tapped comprehension of specific details at sentence and inter-sentence levels. Testing experts also believed the type of reading required for task performance was “very careful reading” and “careful reading”. None of the experts thought the test tasks could be performed solely by expeditious reading. Comprehension of literal meaning

dominated the test, while inferential and evaluative meaning were only marginally tapped by the test tasks. Additionally, the experts remarked that the test heavily relied on vocabulary knowledge and the lexical clues provided in the test tasks and the texts. They were aware that the test was very time constrained and tapped reading speed. These findings indicate that test items basically targeted low level processes of comprehension (Khalifa and Weir, 2009) which involved processing lexico-grammatical features and semantic propositions at sentence and inter-sentence levels. Details from the findings also indicated that the test tapped perceptive and interactive reading (Brown and Abeywickrama, 2004). Extensive reading which taps text representation and represents key features of academic reading construct are disproportionately tapped by the test.

8.2.5. Task difficulty

In the present study another feature of test performance that was analyzed at length was the task difficulty of each of the nine test tasks. Task difficulty analysis was done to provide further details about the nature of each test task and identify potential sources of construct irrelevant variables. The analysis was conducted on the assumption that task difficulty is a function of several variables. Therefore, the study adopted a multi-modal approach to assessing task difficulty by incorporating information related to 1) the amount of time spent on each test task, 2) actual test scores of the test takers, 3) the number of times test takers moved back and forth between the test task and the test (B&FM), 4) test takers' judgements of task difficulty based on their test experience, and 5) testing experts' judgements of task difficulty and the time needed to do the test tasks. The traditional approach to item difficulty is product-oriented and score-based which relies just on the binary assessment of test performance in terms of correct or incorrect, whereas the multi-componential approach adopted in the current study integrated both the product and processes of test performance and provided a richer cognitive perspective to the analysis of task difficulty.

First, one of the key issues observed in the performance of the test takers was the sharp contrast between the amount of time allowed and the amount of time test takers needed for test performance. This can be a reliable indicator of task difficulty. Test takers used time very variably for test performance. All L1 test takers used only 40 minutes (two-thirds) of the total 60 minutes allowed to accomplish the test tasks. The more successful L2 test takers took all the time allowed and could successfully finish the test tasks in the 60 minutes. The less successful L2 test

takers, however, spent much more time completing the test tasks and took 76 minutes to finish the test. Variable use of time by different groups of test takers indicated that test task difficulty varied for different groups of test takers. Based on the amount of time spent by the test takers it can be concluded that test tasks were easy for the L1 test takers, moderately difficulty for the successful L2 test takers, and difficult for the less successful L2 test takers.

Second, actual test scores showed that the three most difficult test tasks for all the three groups of test takers were the same (the *Multiple-choice Task*, the *Matching Heading Task*, and the *Yes, No, Not given Task*). This shows that task demands of these three tasks were consistently higher for all test takers, suggesting that the difficulty of these tasks is inherent to the characteristics of the tasks and not test takers characteristics. It also scores a point for the IELTS RCM which presents full range of difficulty of the test tasks.

Third, results from the test takers' judgements of task difficulty painted a different picture and were not consistent with task difficulty based on actual test scores. This means that perception of task difficulty differed from the actual task difficulty. When asked to judge and rank order the most difficult test tasks, participants opted for test tasks that were not actually that difficult – based on their test scores. This could also be indicative of different concepts of task difficulty for the test takers. In other words, task difficulty seemed to be more than just correctly answering the test item. It is possible that the test takers looked at the challenges they faced in processing the test tasks and judged test task difficulty in terms of difficulty of processing. L1 test takers and the more successful L2 test takers were not consistent in their judgements of the test tasks, but interestingly, the less successful L2 test takers were more accurate in identifying the most difficult test tasks. Arguably, this indicates that less successful L2 test takers could better experience challenges of task performance while L1 test takers faced no serious challenges in doing the test tasks, hence they were not as accurate as the less successful L2 test takers.

Fourth, in terms of the back and forth movements between the text and the test items (B&FM), results were inconsistent and differed across test takers and test tasks. Except for the *Diagram Completion Task* which involved the most frequent back and forth movements for all the test takers, for other test tasks the frequency differed for each test task, suggesting that test performance was a dynamic process and required different kinds and amounts of processing. Depending on the task demands and reading skills of the test takers, they needed to move back and forth between the text and the test tasks. Results also showed that, on average, L1 test takers

had the least B&FM while the less successful L2 test takers had the most B&FM. This suggests that processing was affected by the test takers' language background (L1), level of language proficiency, and level of reading expertise. In light of these findings, it can be argued that the test tasks were much easier for the L1 test takers, while the less successful L2 test takers found the test tasks more difficult, did more processing, and had to move back and forth more frequently. Finally, results from the testing experts' judgements indicated that their judgements of task difficulty differed from the actual test scores obtained by the test takers. Except for *the Matching Headings Task*, their estimates of difficulty differed from actual task difficulty. With regard to the amount of time needed to do the test tasks, testing experts' judgements were surprisingly very different from the actual performance of test takers. The experts were extremely liberal in estimating and judging the amount of time required for test performance. As indicated in Chapter Six, they indicated that for accomplishing the test, 135 minutes was required. This is double the time actually spent by the test takers and far more than the amount of time allowed by the test. This can indicate a high degree of subjectivity in estimating the amount of time needed for test completion.

In concluding this section on task difficulty analysis, I argue that the multi-modal approach adopted in the analysis of task difficulty of the nine test tasks in the IELTS RCM showed that task difficulty is multi-dimensional and needs to be assessed in the wider context of test performance by integrating information from different stakeholder. The multi-componential multi-sources of task difficulty analysis could identify some variables that influence test construct. For instance, test takers' judgements indicated the *Diagram Completion Task* was very difficult for them. Findings from SKSPs used during test performance could provide a valid explanation as to why the task was difficult. Almost all test takers indicated that for successful performance of the *Diagram Completion Task* background knowledge was very helpful. Without background knowledge which it was too difficult to do the task. Using multiple dimensions in the analysis of task difficulty could help identify the role of background in the difficulty of each type of test task.

8.2.6. Reading fluency (speededness)

Another perspective that can explain the amount of time used by different test takers has to do with speededness or reading fluency. Results showed that reading fluency is one of the key components of the IELTS RCM construct. As reported in the previous section, different groups

of test takers spent different amounts of time on the test. Interestingly, more proficient test takers (i.e., L1 test takers and more successful L2 test takers) spent less time doing the tasks while the less proficient L2 test takers needed much more time than allowed. One way to explain these differences is reading fluency. L1 test takers were fluent readers, so they did not need the whole time and used only two-thirds (40 minutes) of the time allowed. The successful L2 test takers spent less time (60 minutes) doing the tasks than the less successful L2 test takers and finished the test tasks on time, but all the successful L2 test takers wished they had more time to do better on the test. L1 test takers and the more successful L2 test takers were fluent enough to process the test tasks within the time limit. The less successful L2 test takers who spent 76 minutes (on average) lacked the fluency to finish the tasks within the time limit. All the less successful L2 test takers needed more time than allowed to finish the nine test tasks. For them, time seemed to have the greatest impact on their performance. They spent 15 more minutes on the tasks, yet they failed to successfully do the tasks. They believed had they more time, they would have performed better. They all claimed had they had more time, they would have done much better on the test. It seemed that the amount of time needed for test performance depended on the level of language proficiency; the higher the level of language proficiency the less time was needed to spend on the tasks and vice versa. In brief, the actual amount of time used by different test takers indicates that reading fluency is one key component of the IELTS RCM reading construct.

As reading fluency (speededness) has not been included in any of the three components of reading comprehension model of Khalifa and Weir (2009), it raises serious questions about the operationalization of the IELTS RCM construct. Speededness can change the nature of reading, taking it from a skill and power, to reading as speed. As indicated in the performance of the test takers, speededness of the IELTS RCM with the constrained time limit influenced different dimensions of test performance of the L2 test takers including the processes and metacognitive strategies they used. Speededness can be also be discussed as a construct irrelevant factor which poses a threat to test construct. It also has bearing on the representation of the academic reading construct which is basically discussed as a set of skills and abilities that readers manifest in their studies with no mention of any time constraints.

These findings resonate with most of the studies related to the impact of time allotted to test performance on reading comprehension scores. For example, Miller (2014) examined the effects of extended time as a test accommodation on a timed reading comprehension test for ESL

students and non-ESL peers under three time-conditions: standard time, time and one half, and double time conditions. Results revealed that all three groups improved reading their comprehension performance under extended time conditions, especially those with lower levels of English language proficiency. Low proficiency students were able to surpass the performance of non-ESL peers at standard time when allotted 50% to 100% extra time. This finding was also supported by Alshammari (2012), who adopted three time-conditions for the reading section of TOFEL IBT and concluded that time given to the reading task significantly impacted overall reading comprehension scores, but he also suggested that the effects varied in relation to the types of questions. Additionally, Colbert (1983) administered a test battery consisting of different language skills under two time-conditions. Interestingly, he found that time was a major factor influencing test performance only for the reading component of the test. In his study, on all subtests the extended time limit group surpassed the standard time limit group, but significant levels of variation were achieved only on the reading subtest. Yet, in another study, Armagan and Genc (2017) conducted an experiment on the impact of timed reading practices on the comprehension level and reading speed of EFL learners in Turkey. Results illustrated that timed reading intervention activities positively affected EFL students' comprehension level and reading speed. The findings also resonate with Runyan (1991), who examined the effect of extra time on reading test performance of normal achieving students and students with reading problems. Results showed significant differences between scores obtained by students with reading problems and by normally achieving students under timed conditions. However, there were no significant differences in test performance between students with reading problems and normally achieving university students when students with reading problems were provided extra time. Normally achieving students did not perform significantly better with extra time, which seems to support the conclusion drawn in the current study that argued for the positive effect of extra time for the less successful test takers and not the native test takers.

Moreover, testing experts' estimate of the amount of time needed to do the test was twice as much as the amount of time officially allowed by the test. This suggests that they thought the test could not be done within the time limit of 60 minutes and that test takers needed much more time to do the test. Based on the amount of time they assigned for the test, it can be argued that experts viewed the test as a speed-reading test, which measures reading within a limited amount of time. This explains why they estimated the test task needed much more time to accomplish.

Furthermore, frequent use of skimming and scanning on the one hand, and high frequency of careful reading on the other, showed that the test is a competition against time. Test takers were quite aware of the time limit and chose to do a lot of skimming and scanning to locate the relevant information, and at the same time they needed to read the relevant information carefully to get the correct answer from it. Given that skimming and scanning represent expeditious reading, test takers needed to strike a balance between the two opposing forces of careful reading and expeditious reading.

In conclusion, the results of the time used by different test takers on the test tasks and experts' judgements of the amount of time needed to do the tasks indicated that reading fluency (speededness) is a key component of the IELTS RCM. The RCM measures reading skills under a very limited time condition. One can assume that results of the L2 test takers' performance (scores) might change under different time constraints, which has implications for the validity of the inferences drawn from the test as a measure of academic reading. Academic reading is usually not time constrained. Academic readers read at their own pace. In the absence of empirical evidence to the contrary, it would seem that the speededness which puts time constraint on a test takers' performance (Hennings, 1987) may be introducing construct irrelevant variance (Messick, 1989) – which, as discussed in Chapter Two, is a threat to the validity of inferences drawn from the test.

8.2.7. Use of background knowledge

Another important feature of the RCM that needs discussion has to do with the use of background knowledge in test performance. It seems that the IELTS RCM assumed topical knowledge as part of the test construct. This is consistent with Khalifa and Weir's (2009) model, which suggests different sources of knowledge, including topical world knowledge, used for inferencing and developing mental representations of a text. However, results from the study indicated that only L1 test takers were familiar with the text topics. They found the topics familiar and relevant to their cultural and background knowledge. Most of the L2 test takers, however, had no familiarity with the topics. They had to rely more on processing the text to make meaning; topical knowledge did not help them at all. The difference can be looked at in terms of test-bias where some items favor one group of test takers over the others.

L2 test takers' lack of topical knowledge can be also looked at through the lens of intercultural competence which can provide L2 learners/ test takers the targeted knowledge,

skills, and attitudes that are needed for effective inter-cultural interactions. In fact, intercultural competence can explain miscommunication not only among students and between students and teachers in EAP classroom (Douglas & Rosvold, 2018), but also in a test performance context.

The observed difference between topical knowledge of L1 test takers and L2 test takers can be also interpreted in light of cultural knowledge or what Hirsch (1983) called *cultural literacy*, which is defined as what an average member of a certain culture is expected to know. The theory argues that people cannot learn reading, writing, and other communication skills separate from the culturally assumed knowledge that shapes what they communicate about. Based on cultural literacy theory, meanings of many words are culture-specific, and people cannot understand or use those words before they get familiar with them. Most people are culturally literate and can fluently communicate in their own culture of origin. This can explain why the topics used in the sample IELTS RCM texts were quite familiar to most L1 test takers but not the L2 participants who mostly claimed to have no background about the topics. L1 test takers were all English L1 speakers living and studying in a western English-speaking society, while the L2 test takers were all non-English speaking learners living and studying in a developing country. It is not surprising they had different topical knowledge to bring to bear on the texts used in the sample test. L2 test takers were culturally literate in their own language but not in English and could not engage with the text as fluently as the L1 participants did.

In conclusion, based on the findings of the test takers' use of skills, knowledge sources, processes, and strategies during test performance, I can conclude that most components and features of Khalifa and Weir's (2009) model of reading comprehension were operationalized and tapped by the IELTS RCM as indicated in the test performance of the L1 test takers, but only some features of the model were tapped in the test performance of L2 test takers. With regard to the *central core component* of the model, only some low-level processes such as word recognition, syntactic parsing, development of semantic propositions, sentence level and inter-sentence level comprehension, and to some extent inferencing were more prominently tapped by the test tasks. However, high-level processes such as inferencing, developing textual representation, and creating a mental model, were involved only in the performance of the L1 and some (2/5) L2 test takers. So, it can be concluded that they were only marginally measured. In terms of the *knowledge base component* of the model, some knowledge sources such as vocabulary knowledge and grammar knowledge were more relevant to successful test task

performance of L2 test takers, while textual knowledge and background knowledge were more relevant to the L1 test takers. However, there were only a few items that required textual and background knowledge. Results also showed that vocabulary knowledge played a crucial role in the test performance of all test takers and should be counted as the most important linguistic knowledge relevant to successful test performance on the RCM. Additionally, regarding the *metacognitive component* of the model, results from the test takers' performance showed that a wide range of strategies related to their comprehension of the text and answering the test task were used, which facilitated the performance of the test takers. However, L2 test takers used metacognitive strategies mainly for task management rather than comprehension purposes. For them, reading was essentially strategic (test-based) rather than reading for typical academic purposes (meaning-based).

The test seems to capture some key components of reading comprehension, but there is doubt if it represents components of academic reading. It might be measuring some aspects of the reading construct but not academic reading. For instance, the test lacks the integrative context of academic reading where reading is integrated with a host of other skills such as writing, speaking, and listening as well as academic reading activities such as read to write, read to ask, read to present, read to summarize, read to learn, etc. Finally, as Bachman and Palmer (2010) argued, tests and assessment procedures should provide context to enhance and “encourage and enable test takers to perform at their highest level of language ability” (p. 13). The IELTS RCM does not seem to provide the best context for EFL and ESL test takers who have learned and practiced English in very diverse contexts. The unrealistic expectations in terms of speededness and topical knowledge may make the test inappropriate for EFL test takers. Additionally, the test does not seem to have built in fairness—which is a technical quality of the test—into the test design. Underrepresentation of some dimensions of academic construct, test bias due to the irrelevance of background knowledge of the texts to the EFL test takers who attend the test with different cultural literacies, and the threat of construct irrelevance due to the speededness of the test can have consequences for the test takers and deprive them of gaining university admission. Discussing these limitations that represent test injustice (McNamara & Ryan, 2011) for high stakes language tests such as IELTS need further consideration and discussion. Finally, findings related to the irrelevance of the text topics to the EFL readers also highlighted one main

shortcoming in Khalifa and Weir's (2009) model, which has no provision for the role and influence of cultural literacy and cultural background in reading comprehension.

8.3. Summary of the Chapter

This chapter merged the main findings of the three sources of evidence considered in the study (the content analysis of the test, test takers' accounts of their cognitive processes used during test performance, and experts' accounts) in light of Messick's unified validity theory, Gorin's (2006) processes-oriented model of test validity and, most importantly, Khalifa and Weir's (2009) model of reading comprehension. The next chapter concludes the study by discussing the implications and applications of the findings for different stakeholders and suggesting some areas for further research.

CHAPTER NINE: CONCLUSION, IMPLICATIONS, AND SUGGESTIONS FOR FUTURE RESEARCH

9.1. Introduction

In the last two decades, global education has steadily increased demand for high-stakes language proficiency tests such as (IELTS) (Farrugia, 2014; Sinclair, Larson, & Rajendram, 2019). High stakes tests function as gatekeepers for entry into target countries, institutes, and universities and exert high impact on the personal lives of test takers (Shohamy, 2001). Results of these tests define a turning point in the life of the test takers and deserve thorough research into their meaning, use, and consequences. Given the impact of high-stakes language proficiency tests such as TOFEL and IELTS on the test takers' lives as well as on other stakeholders—including test users, academic departments, content instructors, language teachers, and parents—these tests deserve more research and understanding. To address the gap in this under-researched area, the current study sought to examine the construct validity of the reading module of the IELTS test and provide cognitive evidence for the construct validity with a focus on studying the Skills, Knowledge sources, Processes, and Strategies (SKSPs) (Gorin, 2006) tapped by the test tasks. To this end, a sample of the IELTS RCM test was content analyzed. Further, three groups of test takers with different language backgrounds and levels of language proficiency took the test and provided their accounts of the processes and strategies they used during their test performance. In addition, a sample of testing experts provided their accounts and judgements of the test tasks. Chapter Eight discussed the findings of these three sources of validity in light of Messick's (1989) unified validity theory, Gorin's (2006) process-oriented validity model, and more importantly Khalifa and Weir's (2009) cognitive model of reading comprehension. This chapter concludes the research findings and presents the implications and applications of the study for different stakeholders and suggests some areas for further research.

9.2. Conclusion

The first research question sought to examine construct validity evidence that can emerge from the linguistic, textual, topical, and cultural dimensions of the test content and the test tasks of the Reading Comprehension Module (RCM) of IELTS. To this end, a sample RCM IELTS

test consisting of three reading texts and nine test tasks (40 test items) was content analyzed.

Results showed that

- The IELTS RCM texts differed from a typical academic text in terms of language, length, topic, tone, argument, degree of abstractness, use of academic words, off-list words, and genre features. These features indicated gaps and mismatches between the texts used in the IELTS RCM and academic reading texts. They underrepresented some of the main textual and topical features of academic texts.
- Comprehension level for a large majority (75%) of test tasks and items was low level and local at sentence and inter-sentential levels. In terms of knowledge sources, vocabulary knowledge was key to understanding the text and answering the test items. Almost all test items included some lexical clues for matching with the relevant information in the text. In brief, the test tapped into selective reading which deals with comprehension of a few features at a low-level of comprehension, targeting the lexical and grammatical features used in short texts (Brown and Abeywickrama, 2004).

The second research question explored the cognitive processes reported in the retrospective verbal reports of three groups of test takers (English L1s (10), more successful (5) and less successful (6) L2 test takers) during their test performance. The focus of this question was on the Skills, Knowledge sources, Processes, and Strategies (SKSPs) used by the test takers. Results were as follows:

- Three main themes of test performance emerged from test takers' accounts of their test performance: 1) Reading which involved different types of reading, 2) Searching for and locating the relevant information in the text, and 3) Answering test items. These themes and the categories involved varied across the three groups of participants, suggesting the test construct altered across L1 and L2 test takers. L1 readers attended to the main ideas of the text and developed a text representation, while L2 test takers mostly relied on the lexico-grammatical analysis at the sentence and inter-sentence levels. Results also showed that the nine test tasks required different SKSPs mostly due to test method formats (e.g., diagram reading, summary, true false not give), and triggered different responses on the part of the test takers. However, low-level comprehension of specific details was emphasized as opposed to high-level inference and text representation.
- Interestingly, all L1 test takers adopted a top-down approach and read the whole text before answering the items, while most L2 test takers adopted a bottom-up approach and focused on doing the test tasks without reading the whole text. For the L1 test takers the reading construct was basically "*reading to comprehend*" the main ideas and some details in order to answer the items. They developed a text representation and mental representation of the text, which helped them to be more fluent and efficient in searching and answering. For L2 test takers, on the other hand, the construct of the RCM was "*reading to search for an answer*". The main purpose of reading, for them, was doing the test tasks not developing a gist of the text and building a text representation. (See Chapter Six, Figures 6.11 and 6.12 for a comparison of models for L1 and L2 test takers).

- An important difference was observed with regard to the Searching Theme. For the L1 test takers, searching for the relevant information involved frequent use of skimming the text to locate the relevant information. They used a search process either to get an answer from the text or double check their guesses. The L2 test takers, on the other hand, scanned the text to search for and locate key words and phrases relevant to the test items. Differences were also observed with regard to the Answering Theme across L1 and L2 test takers. L1 test takers could get to the answer by skimming the relevant information while L2 test takers undertook more careful reading of the relevant information to get the answer.
- Although both L1 and L2 test takers used both careful and expeditious reading, they differed greatly in the dominance of one over the other. As noted above, the L1 test takers read first for meaning (i.e., text representation), whereas the L2 test takers first read the test items and then scanned the text to find the answers. They only read carefully those section of the text (i.e., sentences, paragraphs) that pertained to the item they were answering.
- The Reading Comprehension Model of Khalifa and Weir (2009) does not include cultural literacy as part of any of the three components, and yet, based on the findings of the study, it is an important aspect in reading.

The third research question examined testing experts' accounts and judgements of the test tasks and items in the IELTS RCM. Ten experts took the sample IELTS RCM and judged, coded, and classified the test tasks and test items based on a set of pre-established codes that characterized some key components of the IELTS RCM construct, and the cognitive processes tapped by each test task. Results were as follows:

- Testing experts indicated that the majority (6/9) of the test tasks required low levels of processing that measured literal comprehension of specific details at sentence and inter-sentence levels, while inferential and evaluative meaning were only marginally tapped by a few test items. They also reported that all test tasks required *careful reading* with a limited role for *expeditious reading*. Additionally, they remarked that the test required good vocabulary knowledge in order to use the lexical clues provided in the text and the test tasks. Finally, they were aware that the test was very time constrained and tapped reading speed.
- In terms of the reading skills measured by the test tasks, experts indicated the most frequently used skills included speededness, grammar knowledge, vocabulary knowledge, careful reading, scanning, sentence level comprehension, and comprehension of details. The second set of skills included inter-sentence comprehension, paragraph comprehension, paragraph structure knowledge, inferencing, lexical inferencing, and skimming. The least frequently mentioned skills included knowledge of text type, attending to the writer's attitude, use of background knowledge, and comprehension of the main ideas. These findings were mostly consistent with the results from the content analysis of the test and test takers' accounts of test performance.

- It is worth noting that the main findings from each source of evidence shared many features in common and supported the findings from other sources.

In conclusion, based on the findings from the three sources of evidence, which generated consistent results, it can be concluded that the main reading skill tapped by the IELTS RCM was literal, low-level local comprehension of specific details at sentence and inter-sentence levels. Text level comprehension, and global-evaluative comprehension were marginally measured. Analysis of text features and task features also indicated an underrepresentation of the academic reading construct based on Khalifa and Weir's (2009) Model of Reading Comprehension. Results also showed that the test construct alternated across different test takers. For L2 test takers, test performance was dominated by expeditious reading of test items, while careful reading was limited to those sections of the text that related to the items. For the L1 test takers, the whole task was mainly "*reading for comprehension*" of the text. Task performance was achieved in light of the text representation and mental representation developed during first reading, while for majority of the L2 test takers, test performance essentially focused on "*searching the text and answering the test items*". They read just those parts of the text that were required by the test tasks. Text comprehension was not the main concern of the L2 test takers, especially the less successful L2 test takers. These findings indicate that the test, at best, can measure *readiness for academic reading* rather than academic reading itself.

9.3. Pedagogical Implications and applications

The findings of this research carry a number of implications for the use of the IELTS RCM in the context of measuring academic reading. In this section, some theoretical implications and practical applications of the findings for different test users and stakeholders are presented.

9.3.1. Implications for IELTS RCM preparation course instructors and reading instructors

Examining the cognitive processes involved in reading the IELTS RCM, with a focus on examination of the SKSPs, provided further evidence for the validity of the academic reading tests. Results showed that the cognitive processes and metacognitive strategies used by the test takers were central to different aspects of test performance including text comprehension processes and test task management. Some processes and strategies were used for reading and comprehending the texts, some for searching for the relevant information in the text, yet some

others for answering the test tasks. The multiple functions of the processes and strategies have pedagogical applications for IELTS RCM course preparation instructors and reading instructors.

For most L2 test takers (especially the less successful L2 test takers) most of these processes and strategies were used for processing test items that measured low level comprehension of specific details such as the *Summary Completion Tasks* and the *True, False, Not given Task*. These items could be answered by processing some lexico-grammatical features at sentence and inter-sentence levels. The same low-level processes and strategies, however, did not work successfully for items that tapped higher levels of discourse processing such as the *Matching Headings Task*, the *Multiple-choice Task (two answers)* and the *Yes, No, Not given Task*, which involved developing a text representation and/or mental representation (situation model). These findings indicate that in addition to test management strategies and test wiseness strategies, reading instructors—more specifically IELTS RCM instructors—need to practice and improve the processes, strategies, and reading skills of their students and avoid substituting them with task management strategies and test wiseness strategies. Rumelhart (2004) and Stanovich (1980) argued for the limited role of test task management strategies and test wiseness strategies as compensatory strategies where reading processes may fail to achieve comprehension or successful task performance. Put simply, reading processes and strategies are core to reading performance and cannot be relegated to lower levels of importance. Task management strategies and test wiseness strategies can enhance test performance only if test takers are capable of going through the basic comprehension processes. To achieve higher levels of performance in test tasks, reading instructors and test takers need to focus on the main cognitive processes that are recognized in reading literature. For example, Khalifa and Weir (2009) identified reading processes as word recognition, syntactic parsing, developing semantic representation, inferencing, development of text representation, and mental representation. As these processes all depend on the diverse knowledge sources relevant to each process, reading instructors and test takers need to improve their learners' vocabulary knowledge, grammar knowledge, topical knowledge, and cognitive capacity to make inferences in order to read and comprehend texts. For less successful L2 test takers who showed reading deficiencies stemming from their poor vocabulary knowledge, expanding their vocabulary knowledge can be highly effective in improving their skills and abilities for processing reading texts.

9.3.2. Implications for test takers

These findings also have implications for L2 learners who plan to prepare and take the IELTS RCM test. In light of the findings, potential test takers should be cognizant that they cannot successfully take the test by relying on the test wiseness strategies that are extensively recommended in the IELTS RCM preparation classes and courses. They need to develop the skills, abilities, and knowledge sources required for the main comprehension processes such as inferencing and developing text representation. They should be aware that test wiseness strategies can help them enhance their performance but cannot substitute for the cognitive processes and metacognitive strategies they need for successful test performance. Another implication for the test takers especially those with lower levels of language proficiency is the need to develop good vocabulary knowledge. Results showed that the less-successful L2 test takers struggled with comprehending the text and doing the test tasks. One main reason as expressed by them was a lack of good vocabulary knowledge. They encountered dozens of new words in each text that hindered their comprehension. Further, results showed that most test takers had topical knowledge deficiency and were not familiar with the topics. One advice for the L2 test takers is to read diverse texts and topics and expand their background knowledge to facilitate fluent processing of the texts and test tasks. These improvements can also indirectly contribute to test takers' fluency in processing the text and address one of the main challenges L2 test takers had in their test performance, that is, time needed to do the task.

9.3.3. Implications for test developers

Another application of the findings has to do with IELTS test developers. Results of the current study showed that the texts and the test tasks used in the test did not represent some of the main features of academic reading material and academic reading practices and activities. Depending on the academic discipline, there exist different genres in academic reading. Each reading material is generally a mix of concepts, theories, arguments, and sources of evidence that are organized and presented in specific sets of sections and subsections. The layout, text length, linguistic features, and discourse features of academic material are discipline specific. The texts used in the IELTS RCM, as shown in the content analysis of the test (See Chapter Five), failed to represent many of these academic reading features found in academic reading material. Therefore, it is hard to accept texts that do not represent the main features of academic reading material as a reliable and valid instrument for measuring academic reading comprehension.

Additionally, test tasks that focus on measuring comprehension of specific details at the sentence and inter-sentence level as a representation of academic reading materials and skills, or selective random reading of different sections of a short text can be hardly accepted as a representation of the normal linear process of reading in a non-test context. In normal reading, readers do not usually rely heavily on skimming and scanning for text comprehension; they use them strategically when needed. Most importantly, all reading practices in an academic context are integrated with a set of other skills in a meaningful context. They read to write, read to talk and present, and they read-listen, and they read to learn. The IELTS RCM test tasks, though, were basically, isolated sets of questions asking for some sentence level comprehension which required lexical and grammatical analysis. They had nothing to do with the communicative context of reading in an academic context. Based on these findings, which indicated that there are meaningful differences between the IELTS RCM texts and tasks and the real content of academic reading texts and tasks, test developers need to re-consider their conceptualization of the academic reading construct. They need to reconsider the operational definitions of the IELTS RCM test tasks and make them more compatible with the real academic material and reading practices. This can be addresses at text level by choosing texts that are more representative of academic reading. Further, a more integrative approach to academic reading can better represent of academic reading.

The multi-componential approach adopted in the present study allows test developers to make informed decision for inclusion or exclusion of test tasks and items in the test. The main advantage of the approach was its multimodality in integrating both quantitative and qualitative dimensions of task difficulty. This would allow item writers and test developers to decide how appropriate a given test item and task is to a given testing context and choose the most appropriate test tasks and items.

Furthermore, results from the three sources of evidence indicated the prominence of low-level comprehension in the test tasks which suggest underrepresentation of more global-evaluative comprehension. So, these test tasks may partially address some of the reading skills at undergraduate level where students may read for information and detail at the outset of their programs, and develop critical reading over time, as they engage with their disciplines, or may not need to read critically or may not read much for their courses. However, for graduate students at their outset of their graduate programs, evaluative and critical reading define key

dimensions of academic reading and reading long and complex larger texts is quint essential for academic success. Test developers, then, need to think of developing reading proficiency measures specifically for assessing graduate students' reading proficiency and differentiate academic reading required for undergraduate versus graduate levels.

Results of the study also revealed that the test performance of most L2 test takers was influenced by their vocabulary knowledge. Given the key role of vocabulary knowledge (or lack thereof) in test performance, it is necessary that the IELTS RCM which claims to measure academic reading comprehension—seriously considers the index of general words, off-list words, and more specifically the academic words used in the selected texts. As indicated in Chapter Five, the sample IELTS RCM underrepresented academic words. The number and type of academic words used in the sample academic article and book chapter exceeded the academic words used in IELTS RCM texts. Moreover, in IELTS RCM texts that were not more than 1000 words long, some 10%-17% of the words were off list, meaning they were not part of the general word list. These findings suggest that the text selected needs further examination to better represent academic vocabulary by comparing it with similar indices in actual academic material. Some successful L2 test takers, and all the less successful L2 test takers, complained about the number of new words they encountered in the texts. This can explain why they struggled with comprehending the text. One implication for test developers is to adopt an objective approach in assessing the index of different types of words (general words, academic words, and off-list words used in the text). Results of such an objective approach can improve the quality of the test and contribute to the construct validity of the test. Based on these research findings, improvements need to be made to both text level and task level. Texts serve as the stimulus for test takers' engagement, while test tasks direct the activities and processes used during test performance. Improvements in these two components of the test can influence task performance processes and lead to a better interaction between the test and the test taker.

As discussed earlier, text features influence reading processes (Gorin, 2005; Graesser et al., 1991; Graesser et al., 2004; Kintsch, 1988, 1998, 2004; van Dijk & Kintsch, 1983). One practical suggestion for controlling possible inconsistency and variability in test performance due to text features is developing a more objective and systematic diagnostic assessment approach for evaluating variation at lexical, sentential, and discourse levels of the texts used in the test. A systematic and objective approach can assess the variation of dimensions in lexical and syntactic

features of the texts as well as topical knowledge across different text genres. This may help test developers exert some control over the texts they choose for testing. Such a diagnostic system of text assessment has been used in previous studies (e.g., Biber, 1988; Reppen, 2001). More recently, Deane et al. (2006) were able to identify lexical variation in terms of non-specific lexical items and academic items in texts used in addition to some other variability dimensions among 3rd- to 6th-grade texts. The same or similar approaches can be used for assessing texts that are used in reading comprehension tests. Use of diagnostic approaches for evaluating variation across different texts can help test developers better understand the differential demands on readers, hence it can help them decide on the most appropriate texts.

9.3.4. Implications for policy makers and test users

The findings of this study are also relevant to policy makers, decision makers, and other test users. In light of the details provided by the test content analysis, they can decide if the IELTS RCM is compatible with what they expect the test to do for them. Results of the SKSPs used by test takers can also provide further evidence that decision makers and content teachers can examine. This will provide them with a clearer idea about the compatibility of the SKSPs used in IELTS and the real academic reading activities of the TLU. However, the findings of the current study can only be used if policy makers, decision makers, and other test users have clearly and precisely defined what they expect from results of measures of academic proficiency or language proficiency tests. Clear statements of admission requirements can provide a standard to compare to the findings of the current study and assess the degree to which these tests are useful for the intended standards and purposes. Results from the current study showed that IELTS RCM under-represented several features of academic reading material and academic reading activities. The test did not tap L2 test takers' high-level processes such as developing text representation and creating a situation model, which are two key processes of reading construct. These findings provide a context that is rich in information and that can help them make more informed decisions about using the IELTS RCM test score.

Finally, the multi-componential approach adopted in the analysis of task difficulty of the nine test tasks in the IELTS RCM showed that task difficulty is multi-dimensional and needs to be assessed in a wider context of test performance that integrates information from different stakeholders. The approach has relevance for testing researchers, test developers, item writers, and other test users. It provides a richer approach to the study of test task difficulty and/or item

difficulty. Testing researchers can adopt the methodology used here to gain a deeper and richer understanding of the test task features instead of relying on a dry test score. They can examine task difficulty and item difficulty from the internal dynamic of test processes. By adopting this approach, test developers and item writers can be more aware of the multidimensionality of task difficulty and consider that during item planning or revising of test tasks.

9.4. Limitations of the Study

While the present exploratory study was an attempt to address some gaps in research relating to the validity of the IELTS RCM as a high-stakes English language proficiency test, it was not without limitations. The study examined multiple sources of evidence in the study of construct validity of the IELTS RCM. These sources included test content, the cognitive processes (SKSPs) used by the test takers during test performance, and experts' accounts and judgements of the test construct. Yet, several issues were not addressed in this study that warrant further examination. The study adopted a case study design which is limited in terms of the number of cases investigated. Only a few cases from each bound system (language background and level of language proficiency) participated in the study. Each was unique in terms of their characteristics, and constrained in time, place, individual characteristics, etc. L2 participants were all EFL learners who had learned English in a foreign context. They were all Farsi speakers. The L1 test takers were all freshmen and sophomore students in the humanities and social sciences. These unique features could have influenced their performance and subsequent research findings. Test takers from different linguistic, proficiency, and educational backgrounds, such as ESL language learners and participants from engineering and science backgrounds, could contribute to the transferability of the findings. The data was collected through immediate retrospective verbal reports, which has great potential to provide construct validity evidence. Yet, other methods of data collection might produce different findings. The sample IELTS RCM used in the study was also limited to one single sample. More sample tests could provide further content evidence for test validity. Furthermore, the sample test was administered under non-test conditions. No matter how hard researchers try to create the best condition for test takers' performance, it is still a non-test condition which produces different results from performance under real conditions. Based on these limitations, some areas for further research are suggested and discussed.

9.5. Suggestions for Future Research

This study adopted a case study design to provide detailed descriptions of the cognitive dimension of the IELTS RCM construct. The findings are interesting and informative but limited to the unique characteristics of the research participants and the sample IELTS RCM test used in the study. The findings provided rich sources of evidence for understanding how the construct is operationalized. However, examining the test construct with different and larger samples of English L2 test takers from different language (L1), cultural, and educational backgrounds can enrich the evidence for the validity of the findings. This study recruited three groups of participants: English L1 speakers, more successful English L2 test takers, and less successful test takers. The sample sizes for the English L2 test takers were limited to 5 and 6 participants for each group. During the research process, I thought more participants from these two groups were needed to reach saturation in terms of the variation among test takers. However, for practical reasons this was not possible. Larger sample of English L2 learners would enrich the source of evidence for the test construct. Additionally, the study was limited to only Iranian (English L2) learners who spoke Farsi as their L1. However, ESL participants differ from EFL participants in many ways, such as more exposure and authentic interaction with L1 speakers. Future research can include larger groups and more divergent groups of ESL test takers and compare the results with ESL context. The study can be extended with ESL test takers to provide further evidence for construct validity. More diverse research participants can better represent the world of test takers who actually take the test for university admission purposes. The study also demonstrated the importance of expanding validity evidence by including L1 test takers and examining the potential of the test for measuring academic reading comprehension. Further research can provide further transferability of the findings from the current study.

Further research into the gender differences in using SKSPs may also be very illuminating. During data collection for the current study, test performance of the male and female participants showed different tendencies in terms of the processes and strategies they used. However, as gender differences were not part of the research objectives, these possible differences were not analyzed. Studying gender differences in test performance processes can provide further information about other readers' characteristics that influence test performance.

Moreover, in the present study, only one IELTS RCM sample test was used for data collection and data analysis which limits transferability of the results. To improve transferability of the findings, future research with larger and different samples of the test is suggested. Different samples of the test with different texts and topics can also add to the understanding of the test construct. Results of these studies can hopefully provide more insights from a broader perspective of test samples and test takers.

Furthermore, one factor that profoundly influences test performance is test anxiety (Amiri & Ghonsooly, 2015; Cizek and Burg, 2006; Horwitz, 1986). Examining real test takers' accounts of their feelings and anxieties right before they take the test and immediately after they finish the real test could be very illuminating and shed some light on another dimension of test performance. In the context of this study, the L2 test takers expressed a number of different feelings such as anxiety over the difficulty of some items, anguish at failing to get the answer, stress because of time pressure, confusion as a result of not clearly understanding the text or the test task, frustration at not being able to comprehend parts of the text, excitement and satisfaction when finding the right answer, and other feelings that arose during test performance. These feelings can exert some influence on the test takers' performance. To examine how test takers' feelings may influence their test performance and test construct, further research in this area is encouraged.

In the present study attempts were made to mimic a real testing situation. However, actual test performance is definitely a horse of different color and differs from performance under pseudo conditions. It is important to take the results with a grain of salt. Future research should investigate the SKSPs of real test takers as they finish their performance on the actual test. Practically and logistically this might not be easy, but the results could be much more informative and shed more light on some hidden aspects of test performance.

9.5.1. The Academic Reading Construct

This study was just one step in assessing the construct validity of the RCM. However, findings from the study will be increasingly interpretable if they can be compared against the actual academic reading activities and processes used in academic reading. A complementary area of research that can elaborate on the findings of this study is examining the construct of academic reading as it is practiced in the actual domain. The diversity academic disciplines, and the reading activities and practices involved in each, can provide a rich source of information for

test developers to develop measures of academic reading and evaluate the meaningfulness, appropriateness and usefulness of the existing academic reading tests. The SKSPs used during the IELTS RCM can be best interpreted only if they are compared with findings from research into the academic reading construct. Studying the diverse academic reading experiences of international students and more importantly, reading instructors and content teachers, can be a good area for further research. They can provide some evidence in terms of the reading construct across disciplines. In fact, academic reading should not mean the same thing across disciplines and stakeholders. Further research into academic reading construct can show the extent to which IELTS RCM and other academic reading tests are applicable and relevant to the measurement of the academic reading construct. Results of the current study joined with results from studies of academic reading construct can provide a fuller picture and offer further detailed information to decision makers about the minimum admission score for reading comprehension.

Another area worth studying has to do with online reading. With emergence of *new literacies* due to widespread use of the worldwide web, digital technologies, medias, and electronic gadgets, reading in academic contexts has drastically changed. It has shifted from paper-based reading to digital online/off-line reading. Reading online texts is becoming an increasingly important part of the general construct of reading ability. Currently, in English speaking countries most academic reading materials are made digitally available to the students and most of them seem to do their readings in digital format. The construct of online academic reading may share some common features with the paper-based reading construct, or it might call for a set of different skills, processes, and strategies. There is little research conducted in this area. Coiro & Dobler (2007) reported low correlation between students who are effective print readers versus students who are effective online readers. Therefore, more attention needs to be paid to issues of reading assessment that are tied to reading of online texts. Given the dominance of digital reading material and their importance in the practice of reading in the academic world, future research is encouraged to incorporate digital reading practices, activities, and tasks in the study of academic reading. Further research is recommended to examine different aspects of computer-based reading assessments and reading assessments involving new media. Future research is also encouraged to study how the construct of academic reading might be similar or different across media. Finally, examining the construct of academic reading in the context of the actual practice of digital reading can render more realistic results about reading constructs.

9.5.2. Sub-construct of the IELTS RCM

As the present study illustrated, the construct of the IELTS RCM consisted of several sub-constructs, each represented by one or more test tasks. Each test task served as a sub-construct of the IELTS RCM. However, there is huge gap in the study of the sub-constructs of the IELTS RCM. For every single test task, there are uncertainties and unanswered questions that need further research. The current study examined the construct of each test task by using only one single sample of the task. One single sample is too little to give a comprehensive picture of its construct. A good area of future research would be to focus on examining the validity of every single IELTS RCM test task by using a larger sample of the same test task. Results from these studies can provide a fuller picture of the construct of the test tasks. A good starting point is to consider the results of the current study for each of the nine test tasks and the pattern of test performance that emerged from test performance of the research participants (See Chapter Six).

9.5.3. Test task features

In line with the necessity to research the sub-construct of each test task, the results of the current study revealed some unique features of these test tasks that are worth studying. These features raise some serious questions about each task. For example, the *Matching Headings Task*, which proved to be a very versatile test task, raised some questions about the extent to which the suggested headings best represented the paragraphs. Several L1 test takers rejected the headings as the best headings for the paragraphs. They criticized the heading options for being inappropriate to the paragraphs and argued that the headings were very subjective. They claimed that had they been asked to develop a heading for each paragraph, they would not have come up with these headings. This showed a high degree of subjectivity in the headings that can contribute to construct irrelevant variance in the test. Hence, a good research question could involve delving into the quality and appropriacy of the headings used for each paragraph and examine the subjectivity/objectivity of the heading options. One candidate is use of experimental design that examines testing experts' and test developers' judgement about the appropriacy of headings to the paragraphs. Results could provide further evidence for the validity of the *Matching Headings Task* sub-construct.

Another example that demonstrates how unique features of each test task may influence test construct is related to the *Matching Headings Task*. This time the question is examining the

extent to which the paragraphs used in the *Matching Headings Task* represent normal academic paragraphs. Academic paragraphs usually include a topic sentence that explicitly or implicitly presents the main idea of the paragraph. As discussed in Chapter Five, almost all the paragraphs used in the *Matching Heading Task* did not include any topic sentences. This is in sharp contrast with the common practice in academic paragraphs and texts and raises questions about authenticity and academic representation of the texts used in the IELTS RCM test. Comparative research into the structure of academic paragraphs and paragraphs used in the *Matching Headings Task* can provide further information about the match between domain of the test tasks and domain of actual academic texts. In brief, examining detailed issues related to each test task is worth researching and can identify possible threats to the test construct. In light of identifying the impact of such features, a more precise and clearer picture of the IELTS RCM construct can emerge.

9.5.4. Test bias and item bias

To make sure test items are fair to test takers from different cultural, linguistic, and educational background, test developers study item bias (Abbott, 2007; Fox, 2003). This can help to better understand test method and issues that lead to item bias. As the IELTS RCM is taken by diverse groups of test takers, it is likely that some tasks or items are biased. Research into item bias of the test is highly recommended. It may be necessary to conduct studies across cultural groups, topics to identify biased items. In the context of the present study, results showed the influence of topical knowledge on successful test performance of the test takers. The *Diagram Completion Task* in the sample IELTS RCM, in particular, seemed biased against test takers who were not familiar with the topic. Furthermore, most L1 test takers were quite familiar with the topic and could easily and quickly engage with the text and the points discussed while the L2 test takers were mostly not familiar with the text at all and needed to read more and do more to find out what each paragraph discussed. The participants' successful test performance was influenced by the text topics. This influence was not limited just to the text topic. Some technical knowledge and background knowledge helped some test takers successfully perform on *the Diagram Completion Task*. Therefore, it is important to examine text bias and item bias in a larger sample of texts and tasks used in the IELTS RCM. Examining bias at text and item level can help test developers identify possible source of construct irrelevant variance.

9.5.5. General IELTS RCM

Another area for further research is examining the General IELTS RCM. The questions asked in the current study related to validity are just as relevant and significant for the General RCM IELTS. General IELTS RCM differs from the academic IELTS RCM in many ways. The number of texts and their length, linguistic and textual features, and more importantly the type of tasks used in the General IELTS RCM differ drastically from the academic module. These differences make the General RCM worth studying. In the General Module, test takers are also more diverse and come from different educational, professional, and cultural backgrounds. Future research can look into the SKSPs used in processing the General IELTS RCM and provide further evidence for the construct of reading in high stakes language proficiency tests.

9.5.6. Research designs

In terms of research design, the current exploratory study used a parallel multiple case study to explore the construct validity of the IELTS RCM. Quantitative research designs with larger groups of test takers can enrich the transferability of the findings. Also, other research designs including quantitative experimental designs and mixed methods designs could be used for examining some specific aspects of test performance and variables that influence test performance. Manipulation of different features of the text or test tasks can highlight factors and variables that influence test performance and test construct. For instance, in the current study, results showed that the amount of time used by different test takers significantly differed. Examining the effect of time on test performance and test construct can be a very illuminating research question. The test can be administered with differing timescales to examine the extent to which time influences test scores. The amount of time needed for test performance is relevant to the construct of academic reading because in real life, unlike on a test, students have plenty of time available to read the text. In the present study, results of time needed for test performance was controversial. Testing experts suggested almost twice as much time be allowed for test performance. Hence, future research can monitor changes in test score under different time limits. This could help test developers decide the extent to which time (speed reading) should be defined and included in the test construct.

Further research is also suggested to examine the validity of the test both quantitatively and qualitatively. It is hoped that a mixed methods design could provide a more comprehensive picture of test construct. A large sample of test takers who take the test, followed by a small sample of participants who present their accounts of test performance, may provide

complementary evidence for test construct. Findings from the retrospective verbal reports could provide detailed internal information about the construct of the test, which can be complemented with external test scores gathered from a larger sample of participants.

Additionally, given that the focus of this study was on collecting verbal reports of test takers for studying the cognitive processes of the test, future research is encouraged to incorporate other methods of data collection such as eye-tracking technology to observe the reading behaviors of each group of test takers. Eye movement data can provide more precise evidence to observe details of reading behavior of the test takers and reveal some of the less evident aspects of the cognitive activities and processes used during test performance. Matched with verbal reports of the test takers, eye-movements data collected by eye tracking technologies may further our understanding of the IELTS RCM construct.

Different academic reading tests are more likely to measure students differently. Therefore, it is crucial to examine if and how other academic reading measures vary. An interesting topic worth further research is related to the research into the conceptualization and operationalization of the academic reading comprehension construct in other high stakes tests such as ACEL and TOFEL. To further investigate the theoretical conceptualization of the construct and its operationalization, one very productive approach would be to compare it with the existing academic reading comprehension tests. For a better understanding of the reading construct measured by different academic reading tests, there needs to be a systematic investigation of the similarities and differences among academic reading comprehension tests. Reading assessment researchers have long compared two or more reading comprehension tests (Bowey, 1986; Francis et al., 2005; Keenan, Betjemann, & Roth, 2005; Nation & Snowling, 1997, Spear-Swerling, 2004) and have concluded that commonly used tests of reading comprehension do not necessarily tap the same array of cognitive processes and may be influenced to different degrees by particular skills. Furthermore, as Cutting & Scarborough (2006) have discussed, different tests may provide discrepant information about the comprehension difficulties of the test takers and the component skills that need to be targeted for remediation. It is recommended that future research look at different academic reading comprehension tests that have conceptually defined academic reading from a different theoretical perspective and have operationalized it differently. As Liao (2009) argued, to provide a more comprehensive and balanced view of the construct of L2 reading comprehension, different

reading tests need to be examined. Results from these complementary studies can help decision makers and test users in choosing the most appropriate reading comprehension measure. Decision makers and test users need to examine multiple reading comprehension measures to decide the most fitting measure for their educational program.

Results of such comparative studies can help academic reading assessment in many important ways. First, they can provide further evidence for the construct of academic reading comprehension and show how it is conceptualized and operationalized in different tests. Second, possible similarities and differences in the theoretical and operational dimensions of each of these tests can shed light on the intricacies involved in testing academic reading comprehension skills. These studies can highlight the main advantages and disadvantages of the tests and help identify and define the complex construct of academic reading. Third, they can lead to the development of instruments that correspond more closely to a particular conceptual model of the construct and are more appropriate to a given testing context. Finally, they may show more details about the relationship of reading test performance with text features, test format, task features, and assessment context. The more these intricate relationships are identified and disentangled, the better the construct can be operationalized and measured. This recommendation is in line with research findings that have indicated that reading comprehension tests differ in the processing domains, which makes them non-equivalent (Andreassen and Braten, 2010; Bowyer-Crane and Snowling, 2005; Cain and Oakhill, 2006; Cutting and Scarborough, 2006; Kendeou, Papadopoulou, & Spanoudis, 2012). Likewise, the validity of each of the academic reading comprehension tests can then be unique to themselves and not designed as a one size fits all solution. In short, results from comparative studies of academic reading tests can help decision makers become more aware of the appropriacy of the test they use for their context.

9.5.7. Consequential basis of Validity

Language testing theories and assessment practices have mainly focused on the evidential basis of validity but have paid much less attention to the consequential basis of validity, as discussed by Messick (1989). Yet the consequential basis of test validity examining test impact (Bachman and Palmer, 2010) on society or the washback effects (McNamara, 2001; McNamara, Knoch, & Fan, 2019; Tsagari and Cheng, 2017; Wall and Alderson, 1993) of high stakes tests and their impact on teachers and learners is a critical part of test validation and evaluation of test validity. Test results impact teachers, assessors, test developers, policy makers, parents, and test

takers. A good area of research would be studying the consequences of the pass and fail results of the IELTS RCM for the actual test takers. The emotional and financial cost of pass and fail for the test takers, their family, and the institutions involved are worth examining. Furthermore, effects of test consequences on the motivation, achievement, and emotions of the test takers need further research.

9.5.8. Test improvements

Finally, on a different note regarding the level of engagement of the test processes with the test tasks, results of the study indicated that the IELTS RCM focused on measuring low level local comprehension at sentence and inter-sentence levels. One reason why the test focus was on tapping literal comprehension of details could be the testability of the details; comprehension of the main ideas might not be as testable as the details. Furthermore, tapping global comprehension such as the main ideas requires more sample texts and longer texts. The IELTS RCM texts were all short and less than 1000 words and did not provide enough room to tap the main ideas and macro structure representation of the text. If developing text representation (as discussed in reading theories) is a key component of the reading comprehension process, then it needs to be built into the test construct definition and operationalization. Test developers need to think of different formats and layouts for the test tasks and adopt techniques that engage test takers with the global discourse features of the reading material. One suggestion to make up for deficiency of the test in measuring high level comprehension of text is to scramble the order of the test tasks and test items so that test takers do not proceed through the test tasks in a linear chronological fashion. This would force test takers to either read the whole text before answering the items or skim through the text to develop a gist of the text. With scrambled test tasks, the test takers have to read the whole text and develop a text/mental representation before answering each test task or item. However, such improvements need to be tested and examined with real test takers. Additionally, longer texts which require more amounts of reading, and test tasks that recognize the important of the discourse structure of texts and recognition of main ideas, are more reflective of academic reading and can improve construct of the IELTS RCM. To reflect the multiplicity of text genres in academic reading, the test needs to make provisions for including multiple text genres instead of simple descriptive texts.

Furthermore, to enhance the global-interpretive dimension of the test, test developers can design 1) local tasks and items that demand more interpretive comprehension 2) global items that

require literal comprehension, and 3) global items that require interpretive comprehension. Finally, to improve the expeditious reading of the IELTS RCM, it is important that test developers make provisions to include test tasks that specifically tap expeditious reading. In reality, it is not possible to ask test takers to read the text carefully or expeditiously unless it is obligated by design. Longer texts and test tasks that specifically require expeditious reading can extend the construct of the test and better tap the expeditious dimension of the academic reading construct.

Put differently, the reading construct is most likely underrepresented by all well-known standardized reading assessment systems (Grabe and Jiang, 2013). Results of the current study showed that high stakes academic reading comprehension tests are not the exception. To address construct underrepresentation, it will be highly relevant that testing researchers study academic written discourses and identify their key features, activities, and practices before they decide on the construct of the test. This will enhance test construct representation by including test construct statements that specify a number of realistic and authentic reading purposes and a set of cognitive processes and knowledge bases that constitute academic reading comprehension abilities. As academic reading is very context dependent and readers read for specific purposes, construct statements that highlight the reading purpose can provide a more meaningful description of the construct for test users and decision makers. Furthermore, construct statements that move beyond abstract skill-based descriptions can achieve higher degrees of TLU domain representation and better represent the multi-componential reading construct. Clear purpose statements such as reading for general comprehension, reading to learn, reading to search, careful reading, and so on demand reading tasks and skills that define the reading construct for the specific context of its use. Therefore, purpose statements can be one best model for reading construct definition and operationalization.

9.6. Concluding Remarks

In conclusion, as discussed earlier, reading comprehension is a complex multi-componential skill and not a unitary skill (Grabe and Stoller, 2012; Pressley, 2006; Hudson, 2007). Research in reading assessment has also indicated that assessment of reading comprehension is challenging and controversial (RAND Study Groups, 2002; van den Broek et al., 2005; Nation and Snowling, 1997). Therefore, just like teaching and learning reading,

assessing reading comprehension is not a simple, straightforward matter. Success in comprehending a text and successful performance on the test tasks can be disrupted by challenges and difficulties in any of the multiple components of reading skills. Furthermore, research findings have also indicated that any problem with the component skills hinder comprehension. For instance, problems with: word reading efficiency (Adams, 1990); oral language skills related to vocabulary, linguistic memory, and language processing (Gathercole & Pickering, 2000; Hulme, Muter, Snowling, & Stevenson, 2004); and extended discourse skills (Tabors, Snow, & Dickinson, 2001) can disrupt reading processes. The list can be extended to other component skills such as inferencing skills, retrieval of relevant background information from memory, etc. To lend more credence to this argument, I argue that in the context of the current study all the less successful L2 test takers struggled with paragraph comprehension and in some cases sentence comprehension, because they faced several unfamiliar words in the text and had serious problems with the new words in the text. Lack of sufficient vocabulary knowledge hindered their comprehension. Also, most L2 test takers could not read the text fluently and efficiently, because they had no background knowledge about the texts, while the L1 test takers were more fluent in reading and doing the test. Therefore, playing in such a field where success depends on a host of component skills and abilities is not an easy game; both test developers and test users need to be aware of the complexities involved in second language reading assessment. It may be beneficial for test takers as well to understand explicitly just how complex reading actually is.

As shown in the case of the IELTS RCM, reading tests do not test all the reading skills required for a given context and every test has its own limitations and shortcomings. Test users should decide what essential components of reading comprehension—in addition to word recognition and literal comprehension—are most relevant to their context. For instance, they need to decide if reading speed or attending to the authors' attitude and tone are relevant to their actual reading practices and activities. Further, they should be aware of what a test measures and what it *does not* measure. In fact, it may be most important for test users to be aware of such comprehension deficits in the tests they use. Therefore, assuming that developing a measure of all reading processes and skills is neither possible nor necessary, it can be argued that assessing reading should be more context dependent and focus on measuring test processes and skills that are more relevant to the context of test score use. In any given context, certain components of

reading skills and relevant features have to be highlighted in defining and operationalizing the test construct.

No doubt, international standardized reading assessment faces several limitations in terms of the type and number of test tasks, time constraints, and the heterogeneity of the test takers who come from diverse backgrounds. Given that the reading construct is complex and multi-componential, several challenges lie ahead. Test takers are not homogeneous, and they bring different background knowledge to the test. They vary in many ways in several areas such as disciplinary knowledge, knowledge of genre, cultural experiences, and topic interest, to name a few. Capturing an array of relevant construct components on a test, while addressing the constraints of the standardized assessment of reading (e.g., concerns for validity, reliability, time, cost, usability, and consequence), limits the options for the types of test tasks that can be used and adds to the challenges of developing an assessment from which valid inferences can be drawn.

Controlling for any of these individual variables and factors and the unnecessary confounding effect that they bring is a very demanding task for test developers. However, addressing all these challenges should be a realistic and a long-term goal. Second language assessment research can help address some of these challenges and expand reading measures to define and operationalize the desired L2 reading constructs more accurately. Likewise, in assessing *academic* reading comprehension, the focus should be on developing new techniques to include more relevant reading skills such as reading fluency, disciplinarity, information synthesis from multiple text sources, strategic reading abilities used for academic work, and any other skills that best suit the academic test context.

References

- Abbuhl R., Mackey A. (2017). Second language acquisition research methods. In: King K., Lai YJ., May S. (Eds.) Research methods in language and education. *Encyclopedia of language and education* (3rd ed.). Springer, Cham. https://doi.org/10.1007/978-3-319-02249-9_13
- Adlof, S., Perfetti, C., & Catts, H. (2011). Developmental changes in reading comprehension: Implications for assessment and instruction. In S. Samuels & A. Farstrup (Eds.), *What research has to say about reading instruction* (4th ed.) (pp. 186–214). Newark, DE: International Reading Association.
- Afflerbach, P. (2000). Verbal reports and protocol analysis. In M Kamil, P.B. Mosenthal, RD. Pearson, & R. Barr (Eds.), *Handbook of reading research*. (Vol 3, pp. 1633-179). Mahwah, NJ: Erlbaum.
- Ackermann, K., & Chen, Y. (2013). Developing the academic collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12, 235-247.
- Akbarian, I., & Alavi, S. M. (2013). Comparing the contribution of vocabulary breadth to IELTS and TOEFL reading subtests. *Porta Linguarum: Revista Internacional de Didáctica de Las Lenguas Extranjeras*, 20, 135–151.
- Alavi, S. M., & Akbarian, I. (2012). The role of vocabulary size in predicting performance on TOEFL reading item types. *System*, 40, 376–385. <https://doi.org/10.1016/j.system.2012.07.002>.
- Alavi, S.M., Kaivanpanah, S. & Masjedlou, A.P. (2018). Validity of the listening module of international English language testing system: multiple sources of evidence. *Lang Test Asia* 8, 8. <http://doi.org/10.1186/s40468-018-0057-4>
- Alderson, J. C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing* (pp. 46–57). Arlington. VA: TESOL.
- Alderson, J. C. (1993). The Relationship between grammar and reading in English for academic purposes test battery. In D. Douglas, & C. Chappelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium*. Alexandria, VA: TESOL.
- Alderson, J. C. (2000). *Assessing reading*. London: Cambridge University Press.
- Alderson, J.C. (1991) Bands and scores. In Alderson, J.C. and North, B., (Eds.), *Language testing in the 1990s* (pp. 71-86). London: Modern English Publications, British Council, Macmillan.

- Alderson, C. J., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, 35,4, 79-113.
- Alderson, J.C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, 30,4, 535-556.
- Alderson, J.C., & Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. *Reading in a Foreign Language* 5, 253–270.
- Allami, H., & Aghajari, J. (2014). Pragmatic knowledge assessment in listening sections of IELTS tests. *Theory and Practice in Language Studies*, 4, 2 2014, 332-345.
- Al-Malki, A. S. (2014). Testing the predictive validity of the IELTS test on Omani English candidates' professional competencies. *International Journal of Applied Linguistics and English Literature*, 3, 5, 166-172. <https://doi.org/10.7575/aiac.ijalel.v.3n.5p.166>.
- Alsagoafi, A. (2018). IELTS economic washback: A case study on English major students at King Faisal University in Al-Hasa, Saudi Arabia. *Lang Test Asia* 8, 5. <https://doi.org/10.1186/s40468-018-0058-3>
- Alshammari, H. A. M. (2012). *Effects of time constraint on second language reading comprehension*. Unpublished master thesis, Southern Illinois University Carbondale.
- Al-Shaye, S. (2002). *The effectiveness of metacognitive strategies on reading comprehension and comprehension strategies of eleventh grade students in Kuwaiti high school*. In partial fulfillment for the degree Doctor of Philosophy. Ohio university
- Amiri, M., & Ghonsooly, B. (2015). The relationship between English learning anxiety and the students' achievement on examinations. *Journal of Language Teaching and Research*, 6, 855 – 865.
- Armagan, K.S., & Genc, Z.S. (2017). Impact of timed Reading on comprehension and speed: A Study on Turkish EFL learners. *Journal of Education and Learning*, 6, 204-216.

- Andrews, S. (2015). Individual differences among skilled readers: The Role of lexical quality. In A. Pollatsek and R. Treiman (Eds.). *The Oxford handbook of reading* (pp. 129-148). London: Oxford University Press.
- Artemeva, N., & Fox, J. (2010). Awareness versus production: Probing students' antecedent genre knowledge. *Journal of Business and Technical Communication*, 24,4, 476–515. <https://doi.org/10.1177/1050651910371302>
- Ascalon, M. E. Lawrence S. Meyers, Bruce W. Davis & Niels Smits (2007). Distractor similarity and item-stem structure: Effects on item difficulty, *Applied Measurement in Education*, 20,2, 153-170. <https://doi.org/10.1080/08957340701301272>
- Bachman, L.F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, 16,1, 61-70.
- Bachman, L.F. (1985). *Performance on cloze tests with fixed-ratio and rational deletions. TESOL Quarterly*. 19,3, 535-536.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, Oxford University Press.
- Bachman, L., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.
- Bachman, L.F. (2013). Ongoing challenges in language assessment. In A.J. Kunnan (Ed.) *The Companion to Language Assessment*, (p1-18). London: Wiley Blackwell.
<https://doi.org/10.1002/9781118411360.wbcla128>
- Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*, 13,2, 125–150.
- Bada, E. (2000). Culture in ELT. *Journal of Social Sciences*, 6, 100-110.
- Baker, D.L., Biancarosa, G., Park, B.J. *et al.* (2015). Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in Grades 7 and 8. *Read Writ*, 28, 57–104.

- Bakker, M., Beijaard, D., Roelofs, E., Tigelaar, D., Sanders, P., & Verloop, N. (2008). The impact of construct-irrelevant variance and construct under-representation in assessing teachers' coaching competence. In Bakker, M. *Design and evaluation of video portfolios: Reliability, generalizability, and validity of an authentic performance assessment for teacher*. Open access. leidnuiv.nl.
- Barkaoui, K. (2016). What changes and what doesn't? An examination of changes in the linguistic characteristics of IELTS repeaters' Writing Task 2 scripts. *IELTS Research Reports*, 3, 1-55.
- Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.
- Barnett, M.A. (1986), Syntactic and lexical/semantic skill in foreign language reading: Importance and interaction. *The Modern Language Journal*, 70, 343-349. <https://doi.org/10.1111/j.1540-4781.1986.tb05286.x>
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30,4, 441–465. <https://doi.org/10.1177/0265532212473244>
- Bejar, I. (1985). Speculations on the future of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 279-294). New York: Academic Press.
- Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7, 303-310.
- Belahouel, Y. (2018). *Evaluating the use of the IELTS Test to develop the speaking skill in the oral production module: Case of third year students at Tlemcen University*. Unpublished Master Thesis.
- Berends, I. E., & Reitsma, P. (2006). Remediation of fluency: Word specific or generalised training effects? *Reading and Writing: An Interdisciplinary Journal*, 19,2, 221-234. <https://doi.org/10.1007/s11145-005-5259-3>
- Bernhardt, E. (2010). *Understanding advanced second language reading*. London: Routledge.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.
- Board, C., & Whitney, D. R. (1972). The effect of selected poor item-writing practices on test difficulty, reliability and validity. *Journal of Educational Measurement*, 9,3, 225-233.

- Bodycott, P. (2006). Cultural cross-currents in second language literacy education, *Intercultural Education*, 17, 2, 207-219, <https://doi.org/10.1080/14675980600693947>
- Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (p. 85–115). Cambridge University Press.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111,4, 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Bossers, B. (1992). Reading in two languages: A study of reading comprehension in Dutch as a second language and in Turkish as a first language.
- Bowyer-Crane, & C., Snowling, M. J. (2005). Assessing children's inference generation: what do tests of reading comprehension measure? *British Journal of Educational Psychology*, 75, 189-201.
- Brantmeier, C. (2002). Second language reading strategy instruction at the secondary and university levels: Variations, disparities, ad generalizability. *The Reading Matrix*, 2,3, 1-14.
- Brindley, G. (1991). Defining language ability: The criteria for criteria. In Anivan, S., (Ed.), *Current developments in language testing*, (p. 123-138). Singapore: Regional Language Centre.
- Brisbois, J. E. (1995). Connections between first- and second-language reading. *Journal of Reading Behavior*, 27,4, 565–584. <https://doi.org/10.1080/10862969509547899>
- Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly*, 11, 353–373. <https://doi.org/10.1080/15434303.2014.947532>
- Brown, D. & Abeywickrama, P. (2004). *Language assessment: Principles and classroom practices*. London: Longman.
- Brown, D. (2000). *Teaching by principles: An interactive approach to teaching pedagogy*. London: Longman.
- Brown, J., D. (2014). *Testing in language programs: A comprehensive guide to English language assessment*. New York: JD Brown Publishing.

- Brutten, S.R., Perkins, K. and Upshur, J.A. (1991). Measuring growth in ESL reading. *Paper presented at the Thirteenth Annual Language Testing Research Colloquium*, Princeton, NJ, March.
- Cain, K., & Oakhill, J. (2006). Assessment matters: issues in the measurement of reading comprehension. *British Journal of Educational Psychology*, 76, 697-708.
- Callins, T. (2006). *Culturally responsive literacy instruction*. www.nccrest.org.
- Carrell, P. L. (1989). Metacognitive awareness and second language reading. *Modern Language Journal*, 73, 121–131.
- Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and Individual Differences*, 19, 246–251.
- Carver, R. (1997). Reading for one second, one minute, or one year from the perspective of reading theory. *Scientific Studies of Reading*, 1,1, 3-43.
- Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research*, 49, 278–293.
- Chan, S.H., Bax, S., & Weir, C.J. (2017). Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. *IELTS Research Report*.
- Chan, D., Schmitt, N., DeShon, R., Clause, C., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, 82,2, 300–309.
- Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple. *Language Testing*. 29,1, 19–27.
- Chappelle, A. (2005). *ESOL Tests and testing: A resource for teachers and program administrators offers guidance to professionals in selecting English language tests*. TESOL Publications.

- Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. London: Sage Publications.
- Chen, C. & Liu, Y. (2020). The role of vocabulary breadth and depth in IELTS academic reading tests. *Reading in a Foreign Language*, 32,1, 1–27.
- Cheng, L. (2002). The washback effect on classroom teaching of changes in public examinations. In Savignon, S. J. (Eds.), *Interpreting communicative language teaching: Contexts and concerns in teacher education* (pp. 91-111). New Haven, CT: Yale University Press.
- Cheng, L. (2005). Changing language teaching through language testing: A washback study. *Studies in Language Testing* 21. Cambridge: Cambridge University Press and Cambridge ESOL.
- Cheng, L., & Curtis, A. (2004). *Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning*. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 3–17). Lawrence Erlbaum Associates Publishers.
- Cheng, L., Klinger, D., Fox, J., Doe, C., Jin, Y. and Wu, J. (2014), Motivation and Test Anxiety in Test Performance Across Three Testing Contexts: The CAEL, CET, and GEPT. *TESOL Q*, 48, 300-330.
- Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching*, 48, 436-470.
<https://doi.org/10.1017/S0261444815000233>
- Cheng, L., Watanabe, Y., & Curtis, A. (Eds) (2004). *Washback in language testing: Research contexts and methods*. Lawrence Erlbaum and Associates.
- Chiang, H. H. (2018). English vocabulary size as a predictor of TOEIC listening and reading achievement among EFL students in Taiwan. *Theory and Practice in Language Studies*, 8, 203–212.
<https://doi.org/10.17507/tpls.0802.04>
- Chmiliar, M. (2010). Multiple-Case Designs. In A. J., Mills, G. Durepos, & E. Wiebe (Eds.), *Encyclopedia of case study research* (pp. 581-584). Thousand Oaks, CA: SAGE Publications, Inc. Doi: 10.4135/9781412957397

- Cizek, G. J., & Burg, S. S. (2006). *Addressing test anxiety in a high-stakes environment*. Thousand Oaks, California: Corwin Press.
- Clapham, C. (1996). *The development of IELTS: A study in the effect of background knowledge on reading comprehension*. *Studies in language testing*, 6. New York, NY: Cambridge University Press.
- Clobert, F. P. (1983). *The effects of time limits and reading comprehension performance on the Canadian Tests of Basic Skills*. Masters thesis, Memorial University of Newfoundland.
- Cohen, A. (2004). Arguing about how the world is or how the world should be: The role of argument in IELTS Test. *Journal of English for Academic Purposes*, 3,3, 229–246.
- Cohen, A. (2006) The coming of age of research on test-taking strategies, *Language Assessment Quarterly*, 3,4, 307-331, <https://doi.org/10.1080/15434300701333129>
- Cohen, A. (1999). Strategies and processes in test taking and SLA. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing Research* (Cambridge Applied Linguistics), (pp. 90-111). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139524711.006>
- Cohen, A. (2012). Test taking strategies and task design. In G. Fulcher, & F. Davidson (Eds.). *Routledge handbook of language testing*. (pp. 262-277). Abingdon: Routledge.
- Cohen, A. & Upton, T. (2007). I want to go back to the text: Response strategies on the reading subtest of the new TOEFL. *Language Testing*, 24,2, 209-250.
- Cohen, A. & Upton, T. (2006). Strategies in responding to the new TOEFL reading tasks. *Educational Testing Services*, 666 Rosedale Rd Princeton, NJ 08540-2218.
- Colbert, F. P. (1983) *The effects of time limits and reading comprehension performance on the Canadian Tests of Basic Skills*. Masters thesis, Memorial University of Newfoundland.
- Cook, G. (1997). Key Concepts in ELT: Schemas. *ELT Journal*, 51,1,86-97.
- Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*, 22, 1216-1234.

- Coxhead A. (2000). A new academic word list. *TESOL Quarterly*, 34, 213-238.
<https://doi.org/10.2307/3587951>
- Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. London: Sage Publication.
- Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cronbach, L.J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. *Proceedings of the 1979 ETS Invitational Conference* (pp. 99-108). San Francisco: Jossey-Bass.
- Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity*, (pp. 3-17). Hillsdale, NJ: Erlbaum.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (ed.), *Educational Measurement* (2nd Ed) (pp 443-507). American Council on Education, Washington, DC.
- Cronbach, L.J (1989). Construct validation after thirty years. In R.L. Linn (Ed.). *Intelligence Measurement, theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.
- Crossley, S. A., Allen, D., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23, 84–102.
- Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16, 89–108.
- Crossley, S. A., & McNamara, D. S. (2016). Text-based recall and extra-textual generations resulting from simplified and authentic texts. *Reading in a Foreign Language* 28, 1, 1–19.
- Crossley, S. A., Yang, H. S., & McNamara, D. S. (2014). What’s so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26,1, 92–113.

- Cutting, L. E., & Scarborough, H. S. (2006) Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured, *Scientific Studies of Reading*, 10,3, 277-299, https://doi.org/10.1207/s1532799xssr1003_5
- Deane, D., Sheehan, K. M, Sabatini, J., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers, *Scientific Studies of Reading*, 10,3, 257-275, https://doi.org/10.1207/s1532799xssr1003_4
- Dechant, E. V. (1991). *Understanding and teaching reading: An interactive mode*, New Jersey: Lawrence Erlbaum.
- Degand, L., Sanders, T. (2002). The impact of relational markers on expository text comprehension in L1 and L2. *Reading and Writing*, 15, 739–757. <https://doi.org/10.1023/A:1020932715838>
- deMarrais, K., & Stephen D. Lapan, S. D. (2003). *Foundations for research: Methods of inquiry in education and the social sciences*. London: Routledge
- Diependaele K, Lemhöfer K, Brysbaert M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*. 66,5, 843-863. <https://doi.org/10.1080/17470218.2012.720994>
- Doe, C. & Fox, J. (2011). Exploring the testing process: Three test takers' observed and reported strategy use over time and testing contexts. *Canadian Modern Language Review*, 67,1, 29–53.
- Douglas, D. (2000). *Assessing languages for specific purposes*, Cambridge: CUP.
- Douglas, S & Rosvold, M. (2018). Intercultural communicative competence and English for Academic Purposes: A synthesis review of the scholarly literature, *The Canadian Journal of Applied Linguistics*, 21,1, 23-42.
- Downing, S.M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37, 830-837.
- Downing, S. M. & Haladyna, T. M. (1995). Evaluating licensure and certification examination programs. *CLEAR Exam Review*, 6, 23-26.

- Downing, S. M., & Haladyna, T. M. (1996). Model for evaluating high-stakes testing programs: Why the fox should not guard the chicken coop. *Educational Measurement: Issues and Practice*, 15, 5-12.
- Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, 58,1, 116-121. <https://doi.org/10.1037/h0035197>
- Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing*, 19,4, 347-368.
- Elo, S. & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62,1, 107–115.
- Embretson, S. E. (1985). Introduction to the problem of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 3-17). New York: Academic Press.
- Embretson, S. E. (1986). *Component latent trait models as an information processing approach to testing*. Paper presented at the annual meeting of the American Educational Research Association, April, San Francisco.
- Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38, 343-368.
- Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension test. *Applied Psychological Measurement*, 11,2, 175-193.
<https://doi.org/10.1177/014662168701100207>
- Enright, M.K., Grabe, W., Koda, K., Mosenthal, P.B., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 Reading Framework: A Working Paper. TOFEL Monograph Series*. MS-17. RM-00-4. Princeton, Nj: Educational Testing Service
- Entwistle, N. J. & Entwistle, A. C. (1991). Contrasting forms of understanding for degree examinations: The student experience and its implications. *Higher Education*, 22, 205-227.
- Ericsson, K.A. & Simon, H.A. (1993). *Protocol Analysis: Verbal Reports as Data*, (Revised Ed.)

- Farrugia, C. A. (2014). Charting new pathways to higher education: International secondary students in the United States. New York, NY: Center for Academic Mobility Research, Institute of International Education.
- Fremer, J. (2000). Promoting high standards and the “problem” with construct validation. *NCME Newsletter*, 8,3, 1.
- Fetchkan, R. (2015). *Evaluating the sensitivity of a reading comprehension benchmark assessment as a predictor of performance on a high stakes academic assessment for Middle Schools Students*. N.p., Web.
- Flick, U. (2002). *An Introduction to qualitative research*. London. Sage Publication.
- Fodor, J. A., Bever, T. G., and Garrett, M. F. (1974). *The psychology of Language*. McGraw-Hill, New York.
- Fortus, R., Coriat, R., & Fund, S. (1998). Prediction of item difficulty in the English subtest of Israel’s inter university psychometric entrance test. In Kunnan, A. J. (Ed.), *Validation in language assessment: Selected papers from the 17th Language Research Colloquium*, Long Beach (pp. 61-87). Mahwah, NJ: Lawrence Erlbaum.
- Fox, J., & Cheng, L. (2016). Walk a mile in my shoes: Stakeholder accounts of testing experience with a computer-administered test. *TESL Canada Journal*, 32, 65-86.
<https://doi.org/10.18806/tesl.v32i0.1218>
- Fox, J., & Cheng, L. (2007) Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers, *Assessment in Education*, 14,1, 9-26, <https://doi.org/10.1080/09695940701272773>
- Fox, J. & Cheng, L. (2016) Walk a mile in my shoes: Stakeholder accounts of testing experience with a computer-administered test. *TESL Canada Journal*, 32, 9, 65-86.
<https://doi.org/10.18806/tesl.v32i0.1218>.
- Fox, J., Haggerty, J., & Artemeva, N. (2016). Mitigating Risk: The impact of a diagnostic assessment procedure on the first-year experience in engineering. In J., Read (Ed.), *Post-admission Language Assessment of University Students*, (p. 43-66) Springer.

- Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369–394). Mahwah, NJ: Erlbaum.
- Freebody, P., & Anderson, R. C. (1983). Effects on Text Comprehension of Differing Proportions and Locations of Difficult Vocabulary. *Journal of Reading Behavior*, 15,3, 19–39. <https://doi.org/10.1080/10862968309547487>
- Freedle, R. (1997). The relevance of multiple-choice reading test data in studying expository passage comprehension: The saga of a 15 years effort towards an experimental correlational merger. *Discourse Processes*, 23,3, 399-440.
- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing*, 10,2,133–170.
- Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for mini-talk passages: Implications for construct validity*. TOEFL Research Report, 56. Princeton, NJ: Educational Testing Service.
- Frost, R. (2015). Cross-linguistic perspectives on letter-order processing: Empirical findings and theoretical considerations. In A. Pollatsek and R. Treiman (Eds.). *The Oxford handbook of reading*, 88-98. Oxford: Oxford University Press.
- Fulcher, G. (2012) Assessment Literacy for the Language Classroom, *Language Assessment Quarterly*, 9,2, 113-132, <https://doi.org/10.1080/15434303.2011.642041>
- Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Lawrence Erlbaum Associates, Inc.
- Ghaith, G., and El-Sanyoura, H. (2019). Reading comprehension: The mediating role of metacognitive strategies. *Reading in a Foreign Language*, 31,1, 19–43.
- Gilhooly, K. & Green, C. (1996). Protocol analysis: Theoretical background. In J.T.E. Richardson (Ed.), *Handbook of qualitative research methods for psychology and the social sciences* (pp. Leicester: British Psychological Society Books.

- Goldman, S. R., McCarthy, K. S., & Burkett, C. (2015). *Interpretive inferences in literature*. In E. J. O'Brien, A. E. Cook, & R. F. Lorch, Jr. (Eds.), *Inferences during reading* (p. 386–410). Cambridge University Press.
- Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, 58,3, 787–814. <https://doi.org/10.1016/j.jml.2007.07.001>
- Gordon PC, Chan D. 1995. Pronouns, passives, and discourse coherence. *J. Mem. Lang.* 34, 216–31
- Goshgarian, G. (2004). *Exploring language*. Ney York. Perason.
- Gorin, J.S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42,4, 351-373.
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25, 21–35.
- Gorin, J.S. (2007). Reconsidering issues in validity. *Educational Researcher*, 36, 456-462.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7, 6-10.
- Gove, M. K. (1983). Clarifying teacher's beliefs about reading. *The Reading Teacher*, 37, 3, 261-66.
- Grabe, W. & Jiang, X. (2013). Assessing Reading. In *The Companion to Language Assessment*, A.J. Kunnan (Ed.). John Wiley & Sons, Inc. doi:[10.1002/9781118411360.wbcla060](https://doi.org/10.1002/9781118411360.wbcla060)
- Grabe, W., & Jiang, X. (2013). Assessing reading. In *The Companion to Language Assessment*. John Wiley & Sons, Inc.
- Grabe, W. (2002) Foundations for L2 reading instruction, *The Language Teacher (Online)*.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.
- Grabe, W., & Stoller, L. F. (2011). *Teaching and researching reading*. Harlow, UK: Pearson Education.

- Graesser, A., Golding, J. M., & Long, D. L. (1991). Narrative representation and comprehension. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research, Vol. 2* (p. 171–205). Lawrence Erlbaum Associates, Inc.
- Graesser, A.C., McNamara, D.S., Louwerse, M.M. *et al.* (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36, 193–202. <https://doi.org/10.3758/BF03195564>
- Graesser, A.C., Wiemer-Hastings, P., & Wiemer-Hasti, K. (2001). Constructing inferences and relations during text comprehension. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.) *Text representation: Linguistic and psycholinguistic aspects* (pp. 249-271). Philadelphia, PA: John Benjamins. <https://doi.org/101023/A:1007989901667>
- Grasby, K. L., Byrne, B., & Olson, R. K. (2014). Validity of large-scale reading tests: A phenotypic and behaviour–genetic analysis. *Australian Journal of Education*, 59,1, 5–21. <https://doi.org/10.1177/0004944114563775>
- Green, A., & Hawkey, R. (2012). An empirical investigation of the process of writing Academic Reading test items for the International English Language Testing System. *IELTS Research Report*, 11, 273-374.
- Green, C., & Gilhooly, K. (1996). Protocol Analysis: Practical Implementation. In J.T.E. Richardson (Ed.), *Handbook of qualitative research methods for psychology and the social sciences*. Leicester: British Psychological Society Books.
- Green, A., Unaldi, A., & Weir, C. (2008). The cognitive processes of second language academic readers. LLAS Pedagogic Research Fund Project.
- Green, A., Unaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27,2, 191–211.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 1, 37-50.
- 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*, 15,3, 309-334. https://doi.org/10.1207/S15324818AME1503_5

- Haladyna, T. M., & Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 1, 51-78.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*, 23,1, 17–27. <https://doi.org/10.1111/j.1745-3992.2004.tb00149.x>
- Hambleton, R., & Jirka, S. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Erlbaum.
- Hammadou, J. (1991). Interrelationships among prior knowledge, inference, and language proficiency in foreign language reading. *The Modern Language Journal*, 75, 27-38.
- Hamp-Lyons, L., & Davies, A. (2008). The Englishes of English tests: Bias revisited. *World Englishes*, 27,1, 26-39.
- Han, B., Dai, M., & Yang L. (2004). Analyzing the problems of the College English Test based on a survey. *Foreign Languages and Their Teaching*, 179,2, 17–23.
- Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and the Proget to Lingue 2000*, Cambridge: Cambridge University Press.
- Haynes, M., & Carr, T. H. (1990). Writing system background and second language reading: A component skills analysis of English reading by native speaker-readers of Chinese. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (p. 375–421). Academic Press.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, Mass: Newberry House Publishers.
- Hilden, K., & Pressley, M. (2011). Verbal protocols of reading. In N.K. Duke & M. H. Mallette (Eds.) *Literacy research methodologies* (2nd ed) (pp. 4277-440) New York: NY: Guilford.
- Hirsch, E. (1983). Cultural literacy. *The American Scholar*, 52,2, 159-169.

- Hock, T.S. (1990). The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In J.H.A.L. d. Jong and D.K. Stevenson (Eds.). *Individualizing the assessment of language abilities*. Clevedon, PA: Multilingual Matters.
- Hong-Nam, K., Leavell, A.G. A comparative study of language learning strategy use in an EFL context. (2007). Monolingual Korean and bilingual Korean-Chinese university students. *Asia Pacific Educ. Rev.* 8, 71–88 <https://doi.org/10.1007/BF03025834>
- Horton, R. J. (1975). The construct validity of cloze procedure: An exploratory factor analysis of cloze, paragraph reading, and structure-of-intellect tests. *Reading Research Quarterly*, 10,2, 248-251.
- Horwitz, E. (1986). Primary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly*, 20, 559-564.
- Hosseini, N. (2006). *On the metacognitive awareness of reading strategies and the reading comprehension of Iranian non-English major university students* (Unpublished Master's thesis). Al-Zahra University, Tehran, Iran.
- Hua, A. N., & Keenan, J. M. (2014). The Role of Text Memory in Inferencing and in Comprehension Deficits. *Scientific Studies of Reading*, 18,6, 415–431. <https://doi.org/10.1080/10888438.2014.926906>
- Huang, H., Chern, C., & Chih-cheng, L. (2009). EFL learners' use of online reading strategies and comprehension of texts: An exploratory study. *Computers & Education*, 52, 13–26
- Hudson, T. (2007). *Teaching second language reading*. New York: Oxford University Press.
- Izura, C., & Ellis, A. W. (204). Age of acquisition effects in translation judgment tasks. *Journal of Memory & Language*, 50, 165–181.
- Just, M.A. & Carpenter, P.A. (1987). *The Psychology of Reading and Language Comprehension*. Allyn and Bacon: Boston, Massachusetts.
- Khalifa, H. & Weir, C.J. (2009). Examining Reading: Research and practice in assessing second language reading, *Studies in Language Testing*, 29, Cambridge: UCLES/Cambridge University Press.

- Kane, M. T. (2002). Current concerns in validity theory. *Journal of Educational Measurement*, 38,4, 319-342.
- Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement* 2,3, 135-170.
- Kane, M. T. (2006). Validation. In R. Brennen (Ed.), *Educational measurement*, 4th ed. (pp. 17–64). Westport, CT: Greenwood.
- Kane, M.T. (2008). Terminology, emphasis and utility in validation. *Educational Researcher*, 37, 276-282.
- Kane, M. T., (2011). Validating score interpretations and use: Messick lecture, *Language Testing Research Colloquium*, SAGE: Cambridge.
- Kane, M. T., (2013). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, policy, and practice*. 23,2, 309-311.
- Katalayi, G.B. (2012). *The DR Congo English state examination: Some fundamental validity issues (Context validity evidence)*. [Unpublished Master Thesis, University of Western Cape].
- Katalayi, G. B. (2014). *Fundamental validity issues of an English as a foreign language test: A process-oriented approach to examining the reading construct as measured by the DR Congo English state examination*. [Unpublished Doctoral Dissertation, University of the Western Cape].
- Katalayi, G.B., & Sivasubramaniam, S. (2013). Careful reading versus expeditious reading: Investigating the construct validity of a multiple-choice reading test. *Theory and Practice in Language Studies*, 3,6, 877-884.
- Keenan, J, M, Betjemann R. S. & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension, *Scientific Studies of Reading*, 12,3, 281-300, <https://doi.org/10.1080/10888430802132279>
- Kendeou, p., Papadopoulos, T, & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers, *Learning and Instruction*, 22,5. 354-367.

- Kendeou, P., Smith, E. R., & O'Brien, E. J. (2013). Updating during reading comprehension: Why causality matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39,3, 854–865. <https://doi.org/10.1037/a0029468>
- Kim, M., Crossley, S., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing: An Interdisciplinary Journal* 31, 5. 1155–1180. Web.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
- Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, 49,4, 294–303. <https://doi.org/10.1037/0003-066X.49.4.294>
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95,2, 163–182. <https://doi.org/10.1037/0033-295X.95.2.163>
- Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: The CI perspective *Discourse Processes*. 39,125-128. https://doi.org/10.1207/s15326950dp3902&3_2
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95,2, 163–182. <https://doi.org/10.1037/0033-295X.95.2.163>
- Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *Blackwell handbooks of developmental psychology. The science of reading: A handbook* (p. 209–226). Blackwell Publishing. <https://doi.org/10.1002/9780470757642.ch12>
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85,5, 363–394. <https://doi.org/10.1037/0033-295X.85.5.363>
- Klein, M. Peterson, S. & Simington, L. (1991). Teaching reading in the elementary grades. Needham Heights, MA: Allyn and Bacon.
- Knoch, U., May, L., Macqueen, S., Pill, J., & Storch, N. (2016). Transitioning from university to the workplace: Stakeholder perceptions of academic and professional writing demands. *IELTS Research Reports*. 1, 1-37.

- Koda, K. (2005). *Insights into second language reading. A cross-linguistic approach*. Cambridge: Cambridge University Press.
- Koda, K., & Zehler, A. (2008). *Learning to read across languages: Cross-linguistic relationships in first- and second-language literacy development*. London, UK: Routledge.
- Kong, A. (2006). Connections between L1 and L2 readings: Reading strategies used by four Chinese adult readers. *The Reading Matrix*, 6,2, 19-45.
- Kremmel, B. & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey, *Language Assessment Quarterly*, 17,1, 100-120.
- Knoch, U., May, L., Macqueen, S., Pill, J., & Storch, N. (2016). *Transitioning from university to the workplace: Stakeholder perceptions of academic and professional writing demands*. IELTS Research Reports.
- Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (eds.), *Vocabulary and applied linguistics* (pp. 126–132). UK: Palgrave Macmillan. https://doi.org/10.1007/978-1-349-12396-4_12
- Lazaraton, A. & Frantz, R. (1997). *An analysis of the relationship between task features and candidate output for the Revised FCE Speaking Test*. Unpublished report for Cambridge ESOL.
- Lazaraton, A., & Taylor, L. (2007). Qualitative research methods in language test development and validation. In Fox J., Wesche M., Bayliss D., Cheng L., Turner C., & Doe C. (Eds.), *Language Testing Reconsidered* (pp. 113-130). Ottawa, Ontario: University of Ottawa Press.
- Lee, C. D., & Spratley, A. (2010). *Reading in the disciplines: The challenges of adolescent literacy*. New York, NY: Carnegie Corporation of New York.
- Lesaux, N.K., Pearson, M.R. & Siegel, L.S. (2006). The effects of timed and untimed testing conditions on the reading comprehension performance of adults with reading disabilities. *Read Writ* 19, 21–48 <https://doi.org/10.1007/s11145-005-4714-5>

- Li, M., & Kirby, J. R. (2015). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics*, 36, 611–634. <https://doi.org/10.1093/applin/amu007>
- Lin, L., & Yu, W. (2015). A think-aloud study of strategy use by EFL college readers reading Chinese and English texts. *Journal of Research in Reading*, 38, 286-306.
- Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437–448.
- Liu, O. L. (2011). Do major field of study and cultural familiarity affect TOEFL® iBT reading performance? A confirmatory approach to differential item functioning, *Applied Measurement in Education*, 24,3, 235-255, <https://doi.org/10.1080/08957347.2011.580645>
- Liu, H.M. (2014). *Investigating the relationships between a reading test and can-do statements of performance on reading tasks*. (Unpublished Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database.
- Lu, S. (1990). *An investigation into EFL reading processes: Reading effectiveness, inference construction, metacognitive strategy*. Unpublished MA dissertation, Zhejiang University, China.
- Lumley, T. (1995). The judgements of language-trained raters and doctors in a test of English for health professionals. *Melbourne Papers in Language Testing*, 4,1, 74–98.
- MacGregor, D., Kenyon, D., Christenson, J., & Louguit, M. (2008). *Predicting item difficulty: A rubrics-based approach*. Paper presented at the American Association of Applied Linguistics, Washington, DC.
- Madhumathi, D.P., & Ghosh, D.A. (2017). Academic reading competence of the engineering students. *International Journal of Applied Engineering Research*, 12,20, 9561-9569.
- Malcolm, D. (2009). Reading strategy awareness of Arab medical students studying in English. *System*, 37, 640 – 651.
- Mannesa, S., & Hoyesb, S. M. 1991. Reinstating knowledge during reading: A strategic process. *Discourse Processes*, 21,1, 105-130.

- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. London: Routledge.
- Marslen-Wilson, W., & Tyler, L. K. (1987). *Against modularity*. In J. L. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding* (p. 37–62). The MIT Press.
- Marvin D. Wyne, M. D., & Stuck, G. B. (1979). Time-on-task and reading performance in underachieving children. *Journal of Reading Behavior* 1979, 11,2, 119-129
- McClelland, J., Rumelhart, D., & Hinton, G. (1986). The appeal of parallel distributed processing. In D. Rumelhart, J.
- McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition (Vol,1)*, 3–44. Cambridge, MA: MIT Press. Foundations
- McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, 99,3, 440–466. <https://doi.org/10.1037/0033-295X.99.3.440>
- McKoon, G., & Ratcliff, R. (2017). Adults with poor reading skills and the inferences they make during reading. *Scientific Studies of Reading*, 21, 4, 292–309. <https://doi.org/10.1080/10888438.2017.1287188>
- McMillan, J. H. & Shumacher, S. (2001). *Research in education, Seventh Edition*. Boston, MA: Pearson
- McNamara, Q. (1964). Lost: Our intelligence? Why? *American Psychologist*, 19, 871-882.
- McNamara, T. (1999). Computer-adaptive testing: A view from outside. In *Issues in computer- adaptive testing of reading proficiency*, (Ed.). M Chaloub-Deville, Cambridge University Press, Cambridge, *Issues in computer adaptive testing* (pp 136-149). Cambridge: Cambridge University Press.
- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18,4, 333-350.
- McNamara, T., Knoch, U. & Fan, J. (2019). *Fairness Justice and Language Assessment*. Oxford University Press.

- McNamara, T. & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test, *Language Assessment Quarterly*, 8,2, 161-178, <https://doi.org/10.1080/15434303.2011.565438>
- McRae, K., & Matsuki, K. (2013). Constraint-based models of sentence processing. In R. P. G. van Gompel (Ed.), *Current issues in the psychology of language. Sentence processing* (pp. 51-77). New York, NY, US: Psychology Press.
- McVee, M. B., Dunsmore, K., & Gavelek, J. R. (2005). Schema theory revisited. *Review of Educational Research*, 75(4), 531–566. <https://doi.org/10.3102/00346543075004531>
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (p. 33–48). Lawrence Erlbaum Associates, Inc.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education, Macmillan series on higher education. Educational measurement* (p. 13–103). Macmillan Publishing Co, Inc; American Council on Education.
- Miller, R. L., & Brewer, J. D. (2003). *The A-Z of social research: A dictionary of key social science research Concepts*: Sage Publication.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications, Inc.
- Moje, E. B., Overby, M., Tysvaer, N., & Morris, K. (2008). The complex world of adolescent literacy: Myths, motivations, and mysteries. *Harvard Educational Review*, 78, 107–154.
- Mokhtari, K., & Reichard, C. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology*, 94, 249–259. <https://doi.org/10.1037//0022-0663.94.2.249>
- Mokhtari, K., & Sheorey, R. (2002). Measuring ESL students' awareness of reading strategies. *Journal of Developmental Education*, 25, 3, 2–10.
- Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, 4,1, 43-66

- Moore, T., Morton, J., & Price, S. (2012). Construct validity in the IELTS academic reading test: A comparison of reading requirements in IELTS test items and in university study. *IELTS Research Reports, 11*, 1-89.
- Munby, J. (1978). *Communicative syllabus design*. Cambridge University Press: Cambridge.
- Murray, J. C., Riazi, A. M., & Cross, J. L. (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW, Australia. *Language Testing, 29*,4, 577-595.
- Muthe'n, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*,1, 1–22.
- Nakatsuhara, F., Inoue, C., Berry, V., Galaczi, E. (2017) Exploring performance across two delivery modes for the IELTS Speaking Test: face-to-face and video-conferencing delivery (Phase 2): *IELTS Research Reports*.
- Nakatsuhara, F., Inoue, C., & Taylor, L.J. (2017). An investigation into double-marking methods: comparing live, audio and video rating of performance on the IELTS Speaking Test. *IELTS Research Reports*.
- Naqeeb, H. (2012). Promoting cultural literacy in the EFL classroom. *Global Advanced Research Journal of Educational Research and Reviews, 1*,4, 41-46.
- Nassaji, H. (2002). Schema Theory and Knowledge-Based Processes in Second Language Reading Comprehension: A Need for Alternative Perspectives. *Language Learning, 52*, 439-481.
- Nation, K., Cocksey, J., Taylor, J., & Bishop, D. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension. *Journal of Child Psychology and Psychiatry, 51*, 1031–1039.
- Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67*, 359–370.

- Nation, K., & Snowling, M. (1998). Semantic processing and the development of word-recognition-skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language*, 39, 85–101.
- Oakhill, J. (1982). Constructive processes in skilled and less skilled comprehenders' memory for sentences. *British Journal of Psychology*, 73, 13–20.
- O'Brien, E. J., & Cook, A. E. (2015). *Models of discourse comprehension*. In A. Pollatsek & R. Treiman (Eds.), *Oxford library of psychology. The Oxford handbook of reading* (p. 217–231). Oxford University Press.
- O'Brien, E. J., Cook, A. E., & Guéraud, S. (2010). Accessibility of outdated information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36,4, 979–991. <https://doi.org/10.1037/a0019763>
- O'Brien, E. J., & Cook, A. E., & Peracchi, K. A. (2004). Updating situation models: Reply to Zwaan and Madden. *Journal of Experimental Psychology: Learning, Memory, & Cognition*,30, 289–291.
- O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they?, *Language Assessment Quarterly*, 8,2, 146-160, <https://doi.org/10.1080/15434303.2011.564698>
- Ozuru, Y., Dempsey, K., & McNamara, D. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19, 228–242.
- Ozuru, Y., Rowe, M., O'Reilly, T., McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods* 40, 1001–1015. <https://doi.org/10.3758/BRM.40.4.1001>
- Pang, K. (2008). *The metacognitive expertise assessment tool: A predictive scale for academic achievement across disciplines*. Dissertation Abstracts International (UMI No. AAT 3304568).
- Paris, S. G., & Jacobs, J. (1984). The benefits of informed instruction for children's reading awareness and comprehension skills. *Child Development*, 55, 2083–2093.

- Paris, S. G., & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 15–51). Lawrence Erlbaum Associates, Inc.
- Patton, M. Q. (2001). *Qualitative Research & Evaluation Methods*. New York: Sage Publication.
- Pearson, P. D. (2009). The roots of reading comprehension instruction. In S. E. Isreal & G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 3–31). New York, NY: Routledge.
- Pearson, P. D., Hansen, J., & Gordon, C. (1979). The effect of background knowledge on young children's comprehension of literal and inferential information. *Journal of Reading Behavior*, *11*, 201–209.
- Pellegrino, J. W. (2009). *Mental Models and mental tests*. In Wainer, H., and Braun, H. (Eds.): *Test Validity*. (pp, 49-59). New York: Routledge.
- Perfetti, C. A., & Hart, L. (2001). *The lexical basis of comprehension skill*. In D. S. Gorfein (Ed.), *Decade of behavior. On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (p. 67–86). American Psychological Association. <https://doi.org/10.1037/10459-004>
- Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Malden, MA: Blackwell.
- Phakiti, A. (2003). A closer look at the relationship of cognitive and meta-cognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*, 26–56.
- Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, *25*, 237–272.
- Phipps, S. D., & Brackbill, M. L. (2009). Relationship between assessment item format and item performance characteristics. *American journal of pharmaceutical education*, *73*, 8, 146. <https://doi.org/10.5688/aj7308146>
- Piaget, J. (1970). *Science of education and the psychology of the child*. New York: Viking.

- Pilcher, N., & Richards, K. (2017). Challenging the power invested in the International English Language Testing System (IELTS): Why determining 'English' preparedness needs to be undertaken within the subject context. *Power and Education*, 9,1, 3–17. <https://doi.org/10.1177/1757743817691995>
- Pill, J., & McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing*, 33,2, 217–234. <https://doi.org/10.1177/0265532215607402>
- Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice*, 16,2, 9–13.
- Pressley, M. (2006). *Reading instruction that works* (3rd ed). New York: Guilford Press.
- Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Mahwah, NJ: Erlbaum
- Pritchard, R., & O'Hara, S. (2008). Reading in Spanish and English: A Comparative Study of Processing Strategies. *Journal of Adolescent & Adult Literacy*, 51,8, 630-638.
- Proctor, C. P., August, D., Carlo, M. S., & Snow, C. (2006). The intriguing role of Spanish language vocabulary knowledge in predicting English reading comprehension. *Journal of Educational Psychology*, 98,1, 159–169. <https://doi.org/10.1037/0022-0663.98.1.159>
- Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.
- Qian, H., Woo, A., & Banerjee, B. (2014). Setting an English language proficiency passing standard for entry-level nursing practice Michigan English Language Assessment Battery. *NCLEX Technical Report*. 1-4.
- RAND Reading Study Group. (2002). *Reading for understanding: Toward an R & D program in reading comprehension*. RAND Corporation, Santa Monica, CA.
- Rapley, T. (2007). *The Sage qualitative research kit. Discourse and document analysis*. Sage Publications Ltd. <https://doi.org/10.4135/9781849208901>
- Read, J. (2015). *Assessing English for University Study*. London: Palgrave Macmillan.

- Read, J. (2016) (Ed.). *Post-admission Language Assessment of University Students*. Springer
- Reppen, R. (2001). Register variation in student and adult speech and writing. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 187–199). London: Longman.
- Rasti, I. (2009). Iranian candidates' attitude towards IELTS. *The Asian EFL Journal Quarterly*, 11,3, 110-155.
- Révész, A., Mitchel, M., & Lee, M. (2014). Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory, *IELTS Research Reports*.
- Rayner, K., Pollatsek, A., Ashby, J., Clifton, C (2012). *Psychology of reading*, (2nd Edition). New York: Psychology Press.
- Roothoof, H., & Breeze, R. (2019). Investigating the development of 'grammatical range and accuracy' at different proficiency levels in the IELTS Speaking test. *IELTS Research Reports*.
- Rosenfeld, P., Oltman, P. & Sheppard, K. (2004). Investigating the validity of TOEFL: A feasibility study using content and criterion-related strategies, *TOEFL Research Report*, RR-71.
- Runyan M. K. (1991). The effect of extra time on reading comprehension scores for university students with and without learning disabilities. *Journal of Learn Disability*, 24,2,104-108.
<https://doi.org/10.1177/002221949102400207>
- Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing*, 23,4, 441-474.
- Sadoski, M., & Paivio, A. (2007). Toward a unified theory of reading. *Scientific Studies of Reading*, 11,4, 337–356. <http://doi.org/10.1080/10888430701530714>
- Sadoski, M., & Paivio, A. (1994). A dual coding view of imagery and verbal processes in reading comprehension. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (p. 582–601). International Reading Association.
- Saldaña, J. (2009). *The coding manual for qualitative researchers*. Los Angeles: SAGE Publications.

- Sambell, K., Mcdowell, L., & Brown, S. (1997). But Is It Fair? An Exploratory Study of Student Perceptions of the Consequential Validity of Assessment. *Studies in Educational Evaluation* 23,4, 349–371. Web.
- Saraswati, D. (2015). Reading strategies in the Nepalese School Leaving Exam: Establishing construct validity. *Pro Quest Dissertations Publishing*. Web.
- Sarojani, D. K. (2011). Careful versus expeditious reading: The case of the IELTS reading test *Academic Research International*. 1,3, 1-12.
- Schindler, J, Richter, T., Isberner, M., Naumann, J., & Neeb, Y. (2018): Construct validity of a process-oriented test assessing syntactic skills in German primary school children, *Language Assessment Quarterly*, 3, 183-203. <https://doi.org/10.1080/15434303.2018.1446142>.
- Schmitt, N. (2002). Do reactions to tests produce changes in the construct measured? *Multivariate Behavioral Research*, 37,1, 105–126.
- Scott, D., & Morrison, M. (2005). *Key ideas in educational research*. London: Continuum.
- Sedgwick, C., Garner, M.W., & Vicente-Macia, I. (2016). Investigating the language needs of international nurses: insiders' perspectives. *IELTS Research Reports*.
- Seedhouse, P., & Morales, S. (2017). Candidates questioning examiners in the IELTS Speaking Test: An intervention study. *IELTS Research Report*.
- Sheehan, K.M. & Ginther, A. (2001). What do passage-based MC verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section. Paper presented at the 2000 Annual Meeting of the National Council of Measurement in Education
- Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners*. *Studies in Language Testing*, 32. New York, NY: Cambridge University Press.
- Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24,1, 99-128. <https://doi.org/10.1177/0265532207071513>

- Sheorey, R., & Baboczky, E. S. (2008). Metacognitive awareness of reading strategies among Hungarian college students. In K. Mokhtari, & R. Sheorey (Eds.), *Reading strategies of first- and second-language learners: See how they read* (pp. 161-173). Norwood, MA: Christopher-Gordon Publishers.
- Sheorey, R., & Mokhtari, K. (2001). Differences in the metacognitive awareness of reading. *System*, 29, 431–449. [https://doi.org/10.1016/S0346-251X\(01\)00039-2](https://doi.org/10.1016/S0346-251X(01)00039-2)
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language in language tests*. Harlow, UK: Pearson Education.
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics*, 31,4, 487–512.
- Sinclair, J, Larson, E. J., & Rajendram, S. (2019). Be a machine: International graduate students' narratives around high-stakes English tests, *Language Assessment Quarterly*. 16,2, 236–252. <https://doi.org/10.1080/15434303.2019.1628238>
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36,8, 477–481. <https://doi.org/10.3102/0013189X07311609>
- Solano-Flores, G., & Trumbull, E. (2008). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32,2, 3-13.
- Spielberger, C. D., & Vagg, P. R. (1995). *Test anxiety: Theory, assessment, and treatment*. Washington D.C.: Taylor and Francis.
- Stake, R. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.
- Stanovich, K. E. (2000). *Progress in understanding reading*. The Guilford Press: New York.
- Staub, A. (2015). Reading Sentences: Syntactic parsing and semantic interpretation. In A. Pollatsek and R. Treiman (Eds.). *The Oxford handbook of reading*. 202-216. London: Oxford University Press.
- Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7, 17-29.

- Sternberg, R. J. (1979). *The construct validity of aptitude tests: An information processing assessment*. Paper presented at an ETS Conference on Construct Validation, October, Princeton, NJ.
- Stevens, R.J., Slavin, R.E., & Farnish, A.M. (1991). The effects of cooperative learning and direct instruction in reading comprehension strategies on main idea identification. *Journal of Educational Psychology, 83*, 8-16.
- Stevenson, M., Schoonen, R. and de Glopper, K. (2007), Inhibition or compensation? A multidimensional comparison of reading processes in Dutch and English. *Language Learning, 57*, 115-154.
- Strauss, A., & Corbin, J. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks, CA: Sage Publications, Inc.
- Swales, J.M. (1990). *Genre analysis*. Cambridge: Cambridge University Press.
- Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*, Vol. 3: *Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.
- Tanya R. Shuy, T. R., McCardle, P. & Albro, E. (2006). Introduction to this special issue: Reading comprehension assessment, *Scientific Studies of Reading, 10*,3, 221-224.
- Tavakoli, H. (2014). The effectiveness of metacognitive strategy awareness in reading comprehension: The case of Iranian university EFL students. *The Reading Matrix, 14*,2, 314–336.
- Taylor, L. (2007). The impact of the joint-funded research studies on the IELTS writing test. In IELTS Collected papers: Research in speaking and writing assessment, (Eds.) L. Taylor and P. Falvey, (pp. 479-492). Cambridge University Press, Cambridge.
- Tsagari D., Cheng L. (2017). Washback, impact, and consequences revisited. In Shohamy E., May S. (Eds.) *Language Testing and Assessment: Encyclopedia of Language and Education* (3rd ed.). (pp. 359-372). Springer, Cham
- Tsai, Y., Ernst, C., & Talley, P.C. (2010). L1 and L2 strategy use in reading comprehension of Chinese EFL readers. *Reading Psychology, 31*, 1-29.

- Tweissi, A. I. (1998). The effects of the amount and the type of simplification on foreign language reading comprehension. *Reading in a Foreign Language*, 11, 191–206. *Reading in a Foreign Language*, 28,1, 1–19.
- Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., & Dobbs, C. L. (2015). Beyond vocabulary: Exploring cross-disciplinary academic-language proficiency and its association with reading comprehension. *Reading Research Quarterly*, 50,3, 337-356.
- Unaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27,2, 191-211.
- Urquhart, A. H. (1984). The effect of rhetorical ordering on readability. In C. Alderson, & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 16-28). London & New York: Longman.
- Urquhart, S., & Weir, C. (1998). *Reading in a second language: Process, product and practice*. New York: Addison Wesley Longman.
- Valencia, S. W., Smith, A. T., Reece A. M, Li, M., Wixson, K. K., & and Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, 45,3, 270- 291.
- van den Broek, P. (1994). Comprehension and memory of narrative texts: Inferences and coherence. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 539–588). Academic Press.
- van den Broek, P., Kendeou, P., Kremer, K., Lynch, J. S., Butler, J., White, M.J. *et al* (2005). Assessment of comprehension abilities in young children. In S. Stahl, S. Paris (Eds.), *Children's reading comprehension and assessment* (pp. 107-130) Erlbaum, Mahwah, NJ.
- van den Broek, P., Rapp, D.N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse processes*, 39,2/3, 299-316.
- Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

- Van de Watering, G., Gijbels, D., Dochy, F., & Van der Rijt, J. (2008). Students' assessment preferences, perceptions of assessment and their relationships to study results. *Higher Education*, 56,6, 645–658.
- Vellutino, F. R, Tunmer. W. E., Jaccard, J. J., & Chen, R. (2007) Components of reading ability: Multivariate evidence for a convergent skills model of reading development, *Scientific Studies of Reading*, 11,1, 3-32
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Massachusetts: Harvard University Press.
- Wallace, C. (1992). *Reading*. Oxford: Oxford University Press.
- Weir, C.J., Hawkey, R., Green, A., & Devi, S. (2009). The cognitive processes underlying the academic reading construct as measured by IELTS. *IELTS Research Reports*, 9, 157-189.
- Weir, C. J., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2012). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp. 37–119). Cambridge University Press.
- Weir, C. J., Hawkey, R. Green, T., Devi, S., & Unaldi, A. (2009) The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. *ELTS Research Report*, 9, 97-156.
- Weir, C., Hawkey, R., Green, A., & Devi, S. (2008). The cognitive processes underlying the academic reading construct as measured by IELTS. *IELTS Research Reports*, 9, 157-189.
- Weir, C. J., & Khalifa, H. A. (2008). Cognitive processing approach towards defining reading comprehension, *Research Notes*, 31, 2-10.
- Weir, C. J., & Urquhart, AH. (1998). *Reading in a second language: Process, product and practice*, Longman, New York

- Wendt, A., Woo, A., & Kenny, L. (2009). Setting a passing standard for English proficiency on the Internet-based Test of English as a Foreign Language. *JONA's Healthcare Law, Ethics, and Regulation*, 11,3, 85–90.
- Whitford, V., & Titone, D.A. (2017). The effects of word frequency and word predictability during first- and second-language paragraph reading in bilingual older and younger adults. *Psychology and Aging*, 32, 158–177.
- Whitford, V., & Titone, D.A. (2015). Second-language experience modulates eye movements during first- and second-language sentence reading: evidence from a gaze-contingent moving window paradigm. *Journal of experimental psychology. Learning, memory, and cognition*, 41,4, 1118-29.
- Wray, A., & Pegg, C.A. (2009). The effect of memorized learning on the writing scores of Chinese IELTS test takers. *IELTS Research Reports*, 9, 191-216.
- Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing*, 11,4, 324-348.
- Xie, Q. (2008). Students' perception of the CET4 listening and test preparation practices—Implications for washback. *Research Studies in Education*, 6, 32–47.
- Yalin, S., & Wei, T. (2011). The Relative significance of vocabulary breadth and syntactic knowledge in the prediction of reading comprehension test performance, *Chinese Journal of Applied Linguistics*, 34,3, 113-126. <https://doi.org/10.1515/cjal.2011.028>
- Yamashita, J. (1999). *Reading in a first and a foreign Language: A study of reading comprehension in Japanese (the L1) and English (the L2)*. Unpublished Ph.D. thesis. Lancaster University.UK.
- Yang, Y. (2006). Reading strategies or comprehension monitoring strategies. *Reading Psychology*, 27, 313-343.
- Yano, Y., Long, M., & Ross, S. (1994). Effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning*, 44, 189–219.
- Yin, R. K. (2002). *Case study research: Design and methods*. Thousand Oaks, CA: SAGE Publications.
- Yin, R. K. (2012). *Applications of case study research*. Sage Publications.

- Young, J.W. (2008). Ensuring valid content tests for English language learners. *R&D Connections*, 8, *ETS Research and Development*, Princeton, NJ.
- Yu, G., He, L., & Isaacs, T. (2017). The cognitive processes of taking IELTS academic writing task one: An eye-tracking study. *IELTS Research Reports*.
- Yu, G., Rea-Dickins, P., & Kiely, R. C. (2012). The cognitive processes of taking IELTS academic writing task 1. *IELTS Research Reports*.
- Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *The Journal of Educational Research*, 80,6, 352–358. <https://doi.org/10.1080/00220671.1987.10885782>
- Ziesing, M. (2001). *Cultural literacy and language fluency*. A paper presented in Celebrations of the 20th Anniversary of the University of the Thai Chamber of Commerce, Bangkok, Thailand.
- Zhang, L, Goh, C, & Kunnan, A. (2014). Analysis of test takers' metacognitive and cognitive strategy use and EFL reading test performance: A multi-sample SEM approach. *Language Assessment Quarterly*, 11, 76–102, <https://doi.org/10.1080/15434303.2013.853770>
- Zheng, Y., Klinger, D., Cheng, L., Fox, J. & Doe, C. (2011). Test-Takers' background, literacy activities, and their views of the Ontario Secondary School Literacy Test. *Alberta Journal of Educational Research*, 57, 2, 115–136.
- Zwaan, R. A., & Rapp, D. N. (2006). Discourse comprehension. In M. A. Gernsbacher & M. J. Traxler (Eds.), *Handbook of psycholinguistics* (pp. 725-764). San Diego, CA: Elsevier.
[HTTPS://DOI.ORG/:10.1016/B978-012369374-7/50019-5](https://doi.org/10.1016/B978-012369374-7/50019-5)
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123,2, 162–185. <https://doi.org/10.1037/0033-2909.123.2.162>

Appendices



Appendix A: Test takers' invitation email

Project Clearance Number, Carleton University: #110368

Dear Potential Participant

My name is Raof Moeini. I am a PhD candidate in Applied Linguistics and Discourse Studies at Carleton University. I am writing to you to invite you to participate in a research project that I am working on under the supervision of Dr. Janna Fox.

The purpose of this proposed study is to explore the skills, processes, and strategies used in the Reading Comprehension Module of IELTS as a test of general English general language proficiency

If you participate in the study, we will arrange for two or three sessions of data collection, about one hour each. In the first session, you will read a short passage and you do a Cloze test which is a short passage with some blanks which need to be filled in by you. In the second session, you will be introduced to the Reading Comprehension Module of IELTS test and practice the test to better know what the test is like. In the third session which is I the main session of data collection, you will take another sample of the Reading Comprehension Module of IELTS. You are asked do the test. After you finish the test, you explain what you did and how you did each task in the test. Each session should last about 60 minutes in total.

All the information that you provide in the third session will be audio recorded but it will be kept confidential. Only I will be able to access all the raw un-coded data, which includes the audio-recordings, interview transcripts, and letters of consent. All data will be stored on my personal password-protected computer for the duration of my PhD program.

You may withdraw after the first session, in which case your data will be removed immediately. You may withdraw by contacting me or my supervisor by email or in person. Once the study is completed, if you wish, I will provide you with a summary and an electronic copy of the final paper.

The Carleton University Research Ethics Board reviewed this project. If you have any ethical concerns with the study, please contact Dr. Dr. Bernadette Campbell, Chair, Carleton University Research Ethics Board-A (by phone at 613-520-2600 ext. 2517 or via email at ethics@carleton.ca).

If you would like to participate in this research project, or have any questions or concerns, please feel free to contact me by email, raof.moeini@carleton.ca. If you wish you may also contact my supervisor, Dr. Janna Fox, through the School of Linguistics and Language Studies by email at jannafox@cunet.carleton.ca

Sincerely,
Raof Moeini

Appendix B: Testing Experts' invitation email



Project Clearance Number, Carleton University: #110368

Dear Potential Participant

My name is Raof Moeini. I am a PhD candidate in Applied Linguistics and Discourse Studies at Carleton University. I am writing to you to invite you to participate in a research project that I am working on under the supervision of Dr. Janna Fox.

The purpose of this proposed study is to study testing experts' judgements and accounts of the Reading Comprehension Module of IELTS as a test of general English general language proficiency and what skills, processes and strategies it might measure.

If you participate in the study, we will arrange for one session of data collection, about 90 minutes. You will take the Reading Comprehension Module of IELTS test and. After you finish each reading test task, you will take a questionnaire which addresses your judgements of the tasks. Also, you will participate in an interview to explain what specific activities and processes you used in doing each reading task.

All the information that you provide will be audio recorded but it will be kept confidential. Only I will be able to access all the raw un-coded data, which includes the audio-recordings, interview transcripts, and letters of consent. All data will be stored on my personal password-protected computer for the duration of my PhD program.

You may withdraw during the session in which case your data will be removed immediately.

You may withdraw by contacting me or my supervisor by email or in person. Once the study is completed, if you wish, I will provide you with a summary and an electronic copy of the final paper.

The Carleton University Research Ethics Board reviewed this project. If you have any ethical concerns with the study, please contact Dr. Dr. Bernadette Campbell, Chair, Carleton University Research Ethics Board-A (by phone at 613-520-2600 ext. 2517 or via email at ethics@carleton.ca).

If you would like to participate in this research project, or have any questions or concerns, please feel free to contact me by email, raoofmoeini@carleton.ca. If you wish you may also contact my supervisor, Dr. Janna Fox, through the School of Linguistics and Language Studies by email at jannafox@cunet.carleton.ca

Sincerely,
Raof Moeini



Appendix C: Consent form

I have read this form and the research study has been explained to me. I have been given the opportunity to ask questions and my questions have been answered. If I have additional questions, I have been told whom to contact. I agree to participate in the research study described above and will receive a copy of this consent form.

Participant's Name (printed)

Participant's Signature

Date

If you have any questions about your rights as a participant in this research, you can contact the following office at the Carleton University:

The Carleton University Research Ethics Board reviewed this project. If you have any ethical concerns with the study, please contact

Dr. Dr. Bernadette Campbell,
Carleton University Research Ethics Board-A
613-520-2600 ext. 2517
ethics@carleton.ca.

Appendix D: Proficiency Cloze Test

Cloze test: Read the following text and fill in each blank with ONE appropriate word that fits the passage. Write your answers in the space provided.

What does food tell us about culture?

Have you ever wondered what the food you eat everyday can tell you about where you come from? Have you ever wondered why people from different parts of the world eat different types of food? Do you ever ask yourself why 1..... foods or culinary traditions are so 2..... to your culture? There is more 3..... a connection between food and culture 4..... you may think. On an individual 5....., we grow up eating the food 6..... our cultures. It becomes a part 7..... who each of us are. Many 8..... us associate food from our childhood 9..... warm feelings and good memories and 10..... ties us to our families, holding 11..... special and personal value for us. 12..... from our family often becomes the 13..... food we seek as adults in 14..... of frustration and stress. When I 15..... sick as a kid, I couldn't 16..... rice because I was too weak, 17..... my mother would cook soup and 18..... it to bed for me. The 19..... and taste of the soup became 20..... very familiar to me. Now, whenever 21..... feel tired or stressed, I remember 22..... soup my mom used to make 23..... me and I feel hungry for 24..... soup. On a larger scale, food 25..... an important part of culture. Traditional 26..... is passed down from one generation 27..... the next. It also operates as 28..... expression of cultural identity. Immigrants bring 29..... food of their countries with them 30..... they go and cooking traditional food 31..... a way of preserving their culture 32..... they move to new places. Continuing 33..... make food from their culture for 34..... meals is a symbol of pride 35..... their ethnicity and a means of 36..... with homesickness. Many open their own 37..... and serve traditional dishes. However, the 38..... does not remain exactly the same. 39..... example some ingredients needed to make 40..... dishes may not be readily available, so the taste and flavor of the dishes that they would prepare in their home countries. Additionally, when immigrants sell food in another country, they do not sell it to people from the same countries as them, but to people from different countries. Therefore, they have to alter the original dishes to cater to a wider range of customers with distinct tastes and flavor preferences. Alternations to original dishes can create new flavors that still retain the cultural significance of the dish.

Appendix E: Reading Essay

Nonverbal Behavior: Culture, Gender, and the Media

Ten Kwal Gamble and Michael W. Gamble

Much of this chapter addresses how men and women use written and spoken language. But what about our nonverbal communication skills—the way we speak without words? Do men and women use different nonverbal cues? Are smiles and nods a female communication trait, whereas interruptions and touching are a male one? These are just a few of the body-language questions addressed by this essay.

Ten Kwal Gamble is a professor of communication at the College of New Rochelle, and Michael Gamble is a professor of communication at New York Institute of Technology. They have coauthored many books, including *Communication Works* (1990), *Literature Alive!* (1994), and *Public Speaking in the Age of Diversity* (7998). The following article is an excerpt from their latest collaboration, *Contacts: Communicating Interpersonally* (7998).

1 Throughout the world, people use nonverbal cues to facilitate self-expression. To a great extent, however, the culture of a people modifies their use of such cues. For example, individuals who belong to contact cultures, which promote interaction and encourage displays of warmth, closeness, and availability, tend to stand close to each other when conversing, seek maximum sensory experience, and touch each other frequently. In contrast, members of noncontact cultures discourage the use of such behaviors. Saudi Arabia, France, and Italy are countries with contact cultures; their members relish the intimacy of contact when conversing. In contrast, Scandinavia, Germany, England, Japan, and the United States are low- or lower-contact cultures whose members value privacy and maintain more distance from each other when interacting.

2 Individuals who grow up in different cultures may display emotion or express intimacy in different ways. It is normal, for example, for members of Mediterranean cultures to display highly emotional reactions that are uninhibited and greatly exaggerated; it is common for them to express grief or happiness with open facial displays, magnified gestures, and vocal cues that support the feelings. On the other hand, neither the Chinese nor the Japanese readily reveal their feelings in public, preferring to display less emotion, maintain more self-control, and keep their feelings to themselves; for these reasons, they often remain expressionless.

3 Even when different cultures use the same nonverbal cues, their members may not give the cues the same meaning. In the United States, for example, a nod symbolizes agreement or consent, while in Japan it means only that a message was received.

4 If we hope to interact effectively with people from different cultures, it is important that we make the effort to identify and understand the many ways culture shapes nonverbal communication. We need to acknowledge that one communication style is not intrinsically better than any other; it is that awareness that can help contribute to more successful multicultural exchanges.

5 Men and women commonly use nonverbal communication in ways that reflect societal expectations. For example, men are expected to exhibit assertive behaviors that demonstrate their power and authority; women, in contrast, are expected to exhibit more reactive and responsive behaviors. Thus, it should not surprise us that men talk more and interrupt women more frequently than vice versa.

6 Men are also usually more dominant during interactions than women. Visual dominance is measured by comparing the percentage of time spent looking while speaking with the percentage of time spent looking while listening. When compared with women, men display higher levels of looking while speaking than women do, and lower levels than women when they are listening. Thus, the visual dominance ratio of men

is usually higher than that of women, and again reflects the use of nonverbal cues to reinforce perceptions of social power.

7 Men and women also differ in their use of space and touch. Men use space and touch to assert their dominance over women. As a result, men are much more likely to touch women than women are to touch men. Women are thus more apt to be the recipients of touching actions than they are to be the initiators of such actions. Men also claim more personal space than women usually do, and they more frequently walk in front of women rather than behind them. Thus, in general, males are the touchers, not the touchees, and the leaders rather than the followers.

8 There are nonverbal behaviors that women display more than men do. Women tend to smile more than men. They also commonly display their feelings more overtly than men. In general, women are more expressive than men and exhibit higher levels of involvement when engaged in person-to-person interaction than men. Women also use nonverbal signals to draw others into conversation to a greater extent than men. While women demonstrate an interest in affiliation, men are generally more interested in establishing the strength of their own ideas and agendas than they are in sharing the floor with others. Women also are better interpreters of nonverbal messages than men.

9 All too often, the media and technology help legitimize stereotypical nonverbal displays. The contents of various media contain a plethora of open sexual appeals, portrayals of women obsessed with men, and male—female interactions that portray the man as physically dominant and the female as subordinate. They also include numerous repetitions of the message that “thin is in.”

10 After repeated exposure to such media messages, men and women come to believe and ultimately emulate what they see and hear. Thus, females are primed to devote considerable energy to improving their appearance, preserving their youthfulness, and nurturing others, while males learn to display tougher, more aggressive take-charge cues, trying all the while to control their emotions.

11 Nonverbal power cues echo the male dominance/female subservience mediated message. In advertisements, for example, men are typically portrayed superior to women, who are usually shown in various stages of undress. In the media, nonverbal behaviors portray women as vulnerable and men in control.

12 The repetition of such myths can make us feel dissatisfied and inadequate. If we rely on the media as a reference point for what is and is not desirable in our relationships and interactions, we may find it difficult to be ourselves.

13 Even mediated vocal cues suggest that it is the male and not the female who is the authority. In up to 90 percent of all advertisements male voices are used in voiceovers even when the product being sold is aimed at women.

14 Further complicating the situation is the continued growth of the use of computer-generated virtual reality simulations in addition to allow us to feel as if we were really interacting in different, but make-believe environments and even giving us the opportunity to change our gender, such simulations are also being used to enforce violent gender scenarios resulting in women being threatened and controlled. Even when erotic rather than violent, the media offerings all too often reinforce the notion that men have physical control over women⁷

THINKING CRITICALLY

1. How does culture influence nonverbal communication? What are contact cultures? How can our understanding of contact cultures likewise improve our understanding of nonverbal communication between genders?

2. In paragraph 5, Gamble and Gamble say “it should not surprise us that men talk more and interrupt women more frequently than vice versa.” Why do they feel this statement to be true? Based on your own experience does it seem like a reasonable assessment? Explain.

3. How can looking and listening behaviors of men and women reinforce perceptions of social power?

4. In the final sentence of paragraph 7, the authors connect the act of touching to the act of leading. “Thus, in general, males are the touchers not the touched, and the leaders rather than the followers.” Evaluate the accuracy of this connection. Do you agree with their conclusion? Explain.

- 5) What are some of the difference between men and women in their non-verbal behavior? Which difference is more striking to you?

Appendix F: Writing Summary Task

Write a brief summary of the essay you just read. You may also include your comments and ideas about the ideas discussed. (200 words)

Appendix G: Sample RCM IELTS

Raising the Mary Rose

How a sixteenth-century warship was recovered from the seabed

On 19 July 1545, English and French fleets were engaged in a sea battle off the coast of southern England in the area of water called the Solent, between Portsmouth and the Isle of Wight. Among the English vessels was a warship by the name of Mary Rose. Built in Portsmouth some 35 years earlier, she had had a long and successful fighting career, and was a favourite of King Henry VIII. Accounts of what happened to the ship vary: while witnesses agree that she was not hit by the French, some maintain that she was outdated, overladen and sailing too low in the water, others that she was mishandled by undisciplined crew. What is undisputed, however, is that the Mary Rose sank into the Solent that day, taking at least 500 men with her. After the battle, attempts were made to recover the ship, but these failed.

The Mary Rose came to rest on the seabed, lying on her starboard (right) side at an angle of approximately 60 degrees. The hull (the body of the ship) acted as a trap for the sand and mud carried by Solent currents. As a result, the starboard side filled rapidly, leaving the exposed port (left) side to be eroded by marine organisms and mechanical degradation. Because of the way the ship sank, nearly all of the starboard half survived intact. During the seventeenth and eighteenth centuries, the entire site became covered with a layer of hard grey clay, which minimised further erosion.

Then, on 16 June 1836, some fishermen in the Solent found that their equipment was caught on an underwater obstruction, which turned out to be the Mary Rose. Diver John Deane happened to be exploring another sunken ship nearby, and the fishermen approached him, asking him to free their gear. Deane dived down, and found the equipment caught on a timber protruding slightly from the seabed. Exploring further, he uncovered several other timbers and a bronze gun. Deane continued diving on the site intermittently until 1840, recovering several more guns, two bows, various timbers, part of a pump and various other small finds.

The Mary Rose then faded into obscurity for another hundred years. But in 1965, military historian and amateur diver Alexander McKee, in conjunction with the British Sub-Aqua Club, initiated a project called 'Solent Ships'. While on paper this was a plan to examine a number of known wrecks in the Solent, what McKee really hoped for was to find the Mary Rose. Ordinary search techniques proved unsatisfactory, so McKee entered into collaboration with Harold E. Edgerton, professor of electrical engineering at the Massachusetts Institute of Technology. In 1967, Edgerton's side-scan sonar systems revealed a large, unusually shaped object, which McKee believed was the Mary Rose.

Further excavations revealed stray pieces of timber and an iron gun. But the climax to the operation came when, on 5 May 1971, part of the ship's frame was uncovered. McKee and his team now knew for certain that they had found the wreck, but were as yet unaware that it also housed a treasure trove of beautifully preserved artefacts. Interest in the project grew, and in 1979, The Mary Rose Trust was formed, with Prince Charles as its President and Dr Margaret Rule its Archaeological Director. The decision whether or not to salvage the wreck was not an easy one, although an excavation in 1978 had shown that it might be possible to raise the hull. While the original aim was to raise the hull if at all feasible, the operation was not given the go-ahead until January 1982, when all the necessary information was available.

An important factor in trying to salvage the Mary Rose was that the remaining hull was an open shell. This led to an important decision being taken: namely to carry out the lifting operation in three very distinct stages. The hull was attached to a lifting frame via a network of bolts and lifting wires. The problem of the hull being sucked back downwards into the mud was overcome by using 12 hydraulic jacks. These raised it a few centimetres over a period of several days, as the lifting frame rose slowly up its four legs. It was only when the hull was hanging freely from the lifting frame, clear of the seabed and the suction effect of the surrounding mud, that the salvage operation progressed to the second stage.

In this stage, the lifting frame was fixed to a hook attached to a crane, and the hull was lifted completely clear of the seabed and transferred underwater into the lifting cradle. This required precise positioning to locate the legs into the stabbing guides of the lifting cradle. The lifting cradle was designed to fit the hull using archaeological survey drawings and was fitted with air bags to provide additional cushioning for the hull's delicate timber framework. The third and final stage was to lift the entire structure into the air, by which time the hull was also supported from below. Finally, on 11 October 1982, millions of people around the world held their breath as the timber skeleton of the Mary Rose was lifted clear of the water, ready to be returned home to Portsmouth.

Questions 1-4 (the True, False, Not given Task)

Do the following statements agree with the information given in Reading Passage? In boxes 1-4, write

TRUE if the statement agrees with the information

FALSE if the statement contradicts the information

NOT GIVEN if there is no information on this

- 1 There is some doubt about what caused the Mary Rose to sink.
- 2 The Mary Rose was the only ship to sink in the battle of 19 July 1545.
- 3 Most of one side of the Mary Rose lay undamaged under the sea.
- 4 Alexander McKee knew that the wreck would contain many valuable historical objects.

Your answer:

1	2	3	4
.....

Questions 5-8 (the Matching Features Task)

Look at the following statements (Questions 5-8) and the list of dates below.

Match each statement with the correct date, **A-G**.

Write the correct letter, **A-G**, in boxes 5-8 on your answer sheet.

List of Dates

A-1836 **B- 1840.** **C- 1965** **D-1967**
E-1971 **F-1979** **G- 1982**

- 5** A search for the Mary Rose was launched.
- 6** One person's exploration of the Mary Rose site stopped.
- 7** It was agreed that the hull of the Mary Rose should be raised.
- 8** The site of the Mary Rose was found by chance.

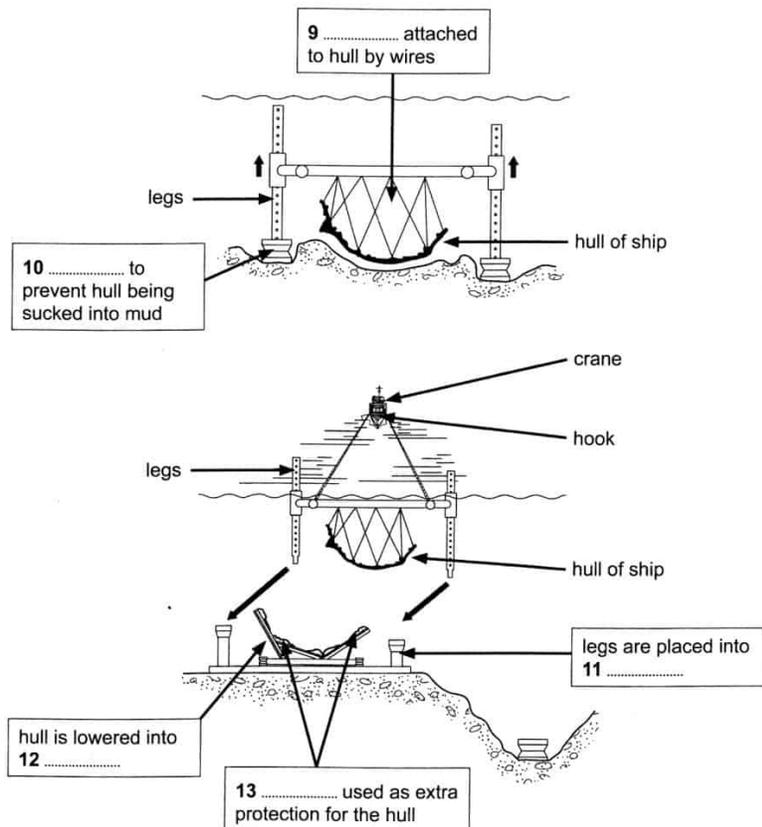
Your answers:

5 **6** **7** **8**
.....

Questions 9-13 (the Diagram Completion Task)

Label the diagram below. Choose **NO MORE THAN TWO WORDS** from the passage for each answer.

Raising the hull of the *Mary Rose*: Stages one and two



What destroyed the civilisation of Easter Island?

A

Easter Island, or Rapa Nui as it is known locally, is home to several hundred ancient human statues - the moai. After this remote Pacific island was settled by the Polynesians, it remained isolated for centuries. All the energy and resources that went into the moai - some of which are ten metres tall and weigh over 7,000 kilos - came from the island itself. Yet when Dutch explorers landed in 1722, they met a Stone Age culture. The moai were carved with stone tools, then transported for many kilometres, without the use of animals or wheels, to massive stone platforms. The identity of the moai builders was in doubt until well into the twentieth century. Thor Heyerdahl, the Norwegian ethnographer and adventurer, thought the statues had been created by pre-Inca peoples from Peru. Bestselling Swiss author Erich von Daniken believed they were built by stranded extraterrestrials. Modern science - linguistic, archaeological and genetic evidence - has definitively proved the moai builders were Polynesians, but not how they moved their creations. Local folklore maintains that the statues walked, while researchers have tended to assume the ancestors dragged the statues somehow, using ropes and logs.

B

When the Europeans arrived, Rapa Nui was grassland, with only a few scrawny trees. In the 1970s and 1980s, though, researchers found pollen preserved in lake sediments, which proved the island had been covered in lush palm forests for thousands of years. Only after the Polynesians arrived did those forests disappear. US scientist Jared Diamond believes that the Rapanui people - descendants of Polynesian settlers - wrecked their own environment. They had unfortunately settled on an extremely fragile island - dry, cool, and too remote to be properly fertilised by windblown volcanic ash. When the islanders cleared the forests for firewood and farming, the forests didn't grow back. As trees became scarce and they could no longer construct wooden canoes for fishing, they ate birds. Soil erosion decreased their crop yields. Before Europeans arrived, the Rapanui had descended into civil war and cannibalism, he maintains. The collapse of their isolated civilisation, Diamond writes, is a 'worst-case scenario for what may lie ahead of us in our own future'.

C

The moai, he thinks, accelerated the self-destruction. Diamond interprets them as power displays by rival Chieftans who, trapped on a remote little island, lacked other ways of asserting their dominance. They competed by building ever bigger figures. Diamond thinks they laid the moai on wooden sledges, hauled over log rails, but that required both a lot of wood and a lot of people. To feed the people, even more land had to be cleared. When the wood was gone and civil war began, the islanders began toppling the moai. By the nineteenth century none were standing.

D

Archaeologists Terry Hunt of the University of Hawaii and Carl Lipo of California State University agree that Easter Island lost its lush forests and that it was an 'ecological catastrophe' - but they believe the islanders themselves weren't to blame. And the moai certainly weren't. Archaeological excavations indicate that the Rapanui went to heroic efforts to protect the resources of their wind-lashed, infertile fields. They built thousands of circular stone windbreaks and gardened inside them, and used broken volcanic rocks to keep the soil moist. In short, Hunt and Lipo argue, the prehistoric Rapanui were pioneers of sustainable farming.

E

Hunt and Lipo contend that moai-building was an activity that helped keep the peace between islanders. They also believe that moving the moai required few people and no wood, because they were walked upright. On that issue, Hunt and Lipo say, archaeological evidence backs up Rapanui folklore. Recent experiments indicate that as few as 18 people could, with three strong ropes and a bit of practice, easily manoeuvre a 1,000 kg moai replica a few hundred metres. The figures' fat bellies tilted them forward, and a D-shaped base allowed handlers to roll and rock them side to side.

F

Moreover, Hunt and Lipo are convinced that the settlers were not wholly responsible for the loss of the island's trees. Archaeological finds of nuts from the extinct Easter Island palm show tiny grooves, made by the teeth of Polynesian rats. The rats arrived along with the settlers, and in just a few years, Hunt and Lipo calculate, they would have overrun the island. They would have prevented the reseedling of the slow-growing palm trees and thereby doomed Rapa Nui's forest, even without the settlers' campaign of deforestation. No doubt the rats ate birds' eggs too. Hunt and Lipo also see no evidence that Rapanui civilisation collapsed when the palm forest did. They think its population grew rapidly and then remained more or less stable until the arrival of the Europeans, who introduced deadly diseases to which islanders had no immunity. Then in the nineteenth century slave traders decimated the population, which shrivelled to 111 people by 1877.

G

Hunt and Lipo's vision, therefore, is one of an island populated by peaceful and ingenious moai builders and careful stewards of the land, rather than by reckless destroyers ruining their own environment and society. 'Rather than a case of abject failure, Rapa Nui is an unlikely story of success', they claim. Whichever is the case, there are surely some valuable lessons which the world at large can learn from the story of Rapa Nui.

Questions 14-20 (the Matching Headings Task)

The Reading Passage has seven paragraphs, A-G.

Choose the correct heading for each paragraph from the list of headings below.

Write the correct number, i-ix, in boxes 1-7 on your answer sheet.

List of Headings:

- i** Evidence of innovative environment management practices
- ii** An undisputed answer to a question about the moai
- iii** The future of the moai statues
- iv** A theory which supports a local belief
- v** The future of Easter Island
- vi** Two opposing views about the Rapanui people
- vii** Destruction outside the inhabitants' control
- viii** How the statues made a situation worse
- ix** Diminishing food resources

Your answers:

- 14 Paragraph A -----iv----
- 15 Paragraph B -----ix----
- 16 Paragraph C ----viii-----
- 17 Paragraph D ----i-----
- 18 Paragraph E -----ii----
- 19 Paragraph F ----vii-----
- 20 Paragraph G ---v-----

Questions 21-24 (the Summary Completion Task)

Complete the summary below.

Choose **ONE WORD ONLY** from the passage for each answer.

Write your answers in boxes **8-11** on your answer sheet.

Jared Diamond's View

Diamond believes that the Polynesian settlers on Rapa Nui destroyed its forests, cutting down its trees for fuel and clearing land for **21**------. Twentieth-century discoveries of pollen prove that Rapa Nui had once been covered in palm forests, which had turned into grassland by the time the Europeans arrived on the island. When the islanders were no longer able to build the **22** -----they needed to go fishing, they began using the island's **23** ----- as a food source, according to Diamond. Diamond also claims that the moai were built to show the power of the island's chieftains, and that the methods of transporting the statues needed not only a great number of people, but also a great deal of **24** -----.

Questions 25-26 (the Multiple-Choice (two answers) Task)

Choose **TWO** letters, **A-E**.

Write the correct letters in boxes **25** on your answer sheet.

On what points do Hunt and Lipo disagree with Diamond?

- A the period when the moai were created
- B how the moai were transported
- C the impact of the moai on Rapanui society
- D how the moai were carved
- E the origins of the people who made the moai

Neuroaesthetics

An emerging discipline called neuroaesthetics is seeking to bring scientific objectivity to the study of art, and has already given us a better understanding of many masterpieces. The blurred imagery of Impressionist paintings seems to stimulate the brain's amygdala, for instance. Since the amygdala plays a crucial role in our feelings, that finding might explain why many people find these pieces so moving. The discipline of neuroaesthetics aims to bring scientific objectivity to the study of art. Neurological studies of the brain, for example, demonstrate the impact which Impressionist paintings have on our emotions.

Could the same approach also shed light on abstract twentieth-century pieces, from Mondrian's geometrical blocks of colour, to Pollock's seemingly haphazard arrangements of splashed paint on canvas? Sceptics believe that people claim to like such works simply because they are famous. We certainly do have an inclination to follow the crowd. When asked to make simple perceptual decisions such as matching a shape to its rotated image, for example, people often choose a definitely wrong answer if they see others doing the same. It is easy to imagine that this mentality would have even more impact on a fuzzy concept like art appreciation, where there is no right or wrong answer.

Angelina Hawley-Dolan, of Boston College, Massachusetts, responded to this debate by asking volunteers to view pairs of paintings - either the creations of famous abstract artists or the doodles of infants, chimps and elephants. They then had to judge which they preferred. A third of the paintings were given no captions, while many were labelled incorrectly -volunteers might think they were viewing a chimp's messy brushstrokes when they were actually seeing an acclaimed masterpiece. In each set of trials, volunteers generally preferred the work of renowned artists, even when they believed it was by an animal or a child. It seems that the viewer can sense the artist's vision in paintings, even if they can't explain why.

Robert Pepperell, an artist based at Cardiff University, creates ambiguous works that are neither entirely abstract nor clearly representational. In one study, Pepperell and his collaborators asked volunteers to decide how powerful they considered an artwork to be, and whether they saw anything familiar in the piece. The longer they took to answer these questions, the more highly they rated the piece under scrutiny, and the greater their neural activity. It would seem that the brain sees these images as puzzles, and the harder it is to decipher the meaning, the more rewarding is the moment of recognition.

And what about artists such as Mondrian, whose paintings consist exclusively of horizontal and vertical lines encasing blocks of colour? Mondrian's works are deceptively simple, but eye-tracking studies confirm that they are meticulously composed, and that simply rotating a piece radically changes the way we view it. With the originals, volunteers' eyes tended to stay longer on certain places in the image, but with the altered versions they would flit across a piece more rapidly. As a result, the volunteers considered the altered versions less pleasurable when they later rated the work.

In a similar study, Oshin Vartanian of Toronto University asked volunteers to compare original paintings with ones which he had altered by moving objects around within the frame. He found that almost everyone preferred the original, whether it was a Van Gogh still life or an abstract by Miro. Vartanian also found that changing the composition of the paintings reduced activation in those brain areas linked with meaning and interpretation.

In another experiment, Alex Forsythe of the University of Liverpool analysed the visual intricacy of different pieces of art, and her results suggest that many artists use a key level of detail to please the brain. Alex Forsythe of the University of Liverpool believes many artists give their works the precise degree of complexity which most appeals to the viewer's brain. Too little and the work is boring, but too much results in a kind of 'perceptual overload', according to Forsythe. What's more, appealing pieces both abstract and representational, show signs of "fractals" - repeated motifs recurring in different scales, fractals are common throughout nature, for example in the shapes of mountain peaks or the branches of trees. She also observes that pleasing works of art often contain certain repeated images which occur

frequently in the natural world. It is possible that our visual system, which evolved in the great outdoors, finds it easier to process such patterns.

It is also intriguing that the brain appears to process movement when we see a handwritten letter, as if we are replaying the writer's moment of creation. This has led some to wonder whether Pollock's works feel so dynamic because the brain reconstructs the energetic actions the artist used as he painted. This may be down to our brain's 'mirror neurons', which are known to mimic others' actions. The hypothesis will need to be thoroughly tested, however it might even be the case that we could use neuroaesthetic studies to understand the longevity of some pieces of artwork. While the fashions of the time might shape what is currently popular, works that are best adapted to our visual system may be the most likely to linger once the trends of previous generations have been forgotten.

It's still early days for the field of neuroaesthetics - and these studies are probably only a taste of what is to come. It would, however, be foolish to reduce art appreciation to a set of scientific laws. We shouldn't underestimate the importance of the style of a particular artist, their place in history and the artistic environment of their time. Abstract art offers both a challenge and the freedom to play with different interpretations. In some ways, it's not so different to science, where we are constantly looking for systems and decoding meaning so that we can view and appreciate the world in a new way.

Questions 27-31 (the *Multiple-Choice Task*)

Choose the correct letter, A, B, C or D.

- 27.** In the second paragraph, the writer refers to a shape-matching test in order to illustrate
- A the subjective nature of art appreciation.
 - B the reliance of modern art on abstract forms.
 - C our tendency to be influenced by the opinions of others.
 - D a common problem encountered when processing visual data.
- 28.** Angelina Hawley-Dolan's findings indicate that people
- A mostly favour works of art which they know well.
 - B hold fixed ideas about what makes a good work of art.
 - C are often misled by their initial expectations of a work of art.
 - D have the ability to perceive the intention behind works of art.
- 9.** Results of studies involving Robert Pepperell's pieces suggest that people
- A can appreciate a painting without fully understanding it.
 - B find it satisfying to work out what a painting represents.
 - C vary widely in the time they spend looking at paintings.
 - D generally prefer representational art to abstract art.
- 30.** What do the experiments described in the fifth paragraph suggest about the paintings of Mondrian?
- A They are more carefully put together than they appear.
 - B They can be interpreted in a number of different ways.
 - C They challenge our assumptions about shape and colour.
 - D They are easier to appreciate than many other abstract works.

Questions 32-34 (the Summary Completion Task)

Complete the summary using the list of words, **A-H**, below.

Write the correct letters, **A-H**, in boxes **31-33** on your answer sheet.

Art and the Brain

The discipline of neuroaesthetics aims to bring scientific objectivity to the study of art. Neurological studies of the brain, for example, demonstrate the impact which Impressionist paintings have on **31**-----
 ---c----- Alex Forsythe of the University of Liverpool believes many artists give their works the precise degree of **32** ----- which most appeals to the viewer's brain. She also observes that pleasing works of art often contain certain repeated **33** ----- which occur frequently in the natural world.

- | | | | |
|-------------------|---------------|------------|--------------|
| A- interpretation | B- complexity | C- emotion | D. movements |
| E- Skill | F- layout | G. Concern | H. images |

Questions 34-39 (the Yes, No, Not given Task)

Do the following statements agree with the views of the writer in Reading Passage?

In boxes **34-40**, Write

YES if the statement agrees with the views of the writer

NO if the statement contradicts the views of the writer

NOT GIVEN if there is no information on this

NO 34 Forsythe's findings contradicted previous beliefs on the function of 'fractals' in art.

YES 35 Certain ideas regarding the link between 'mirror neurons' and art appreciation require further verification.

NG 36 People's taste in paintings depends entirely on the current artistic trends of the period.

YES 37 Scientists should seek to define the precise rules which govern people's reactions to works of art.

NG 38 Art appreciation should always involve taking into consideration the cultural context in which an artist worked.

NG 39 It is easier to find meaning in the field of science than in that of art.

40. What would be the most appropriate subtitle for the article?

- A Some scientific insights into how the brain responds to abstract art
- B Recent studies focusing on the neural activity of abstract artists
- C A comparison of the neurological bases of abstract and representational art
- D How brain research has altered public opinion about abstract art

Appendix H: Test Performance Observation Scheme

Name:		Gender:		Proficiency level:
Major:		L1:		
	Time spent reading the text	Time spent on the task	Total Time	# back and forth movements between the text and the task
True, False, Not given Task				
Matching Features Task				
Diagram Completion Task				
Matching Headings Task				
Summary completion 1 Task				
Multiple-Choice (Two answers) Task				
Multiple-Choice Task				
Summary Completion 2 Task				
Yes, No, Not Given Task				

Appendix I: Construct of the RCM IELTS Questionnaire

Level of reading	Word	Phrase	Sentence	Inter-sentence	Paragraph	Text
Knowledge needed	Lexical	Grammatical (sentential/ inter-sentential)		Textual (cohesion, coherence)	World knowledge	Others
Type of comprehension	Literal	Inferential		Evaluative		
Reading speed	Careful ...1.....10.....				Expeditious	
Difficulty of the text	Easy1.....10.....				Difficult	
Difficulty of the task	Easy1.....10...				Difficult	
Level of task processing	Low level ...1.....10...				High level	
Time needed	<input type="checkbox"/> -1m.	<input type="checkbox"/> 1m	<input type="checkbox"/> 90s	<input type="checkbox"/> 2ms	<input type="checkbox"/> more	
Reading skills measured	<input type="checkbox"/> Understanding the main idea <input type="checkbox"/> Paragraph structure <input type="checkbox"/> Vocabulary knowledge <input type="checkbox"/> Sentence comprehension <input type="checkbox"/> Lexical Inferencing <input type="checkbox"/> World knowledge inferencing <input type="checkbox"/> Scanning <input type="checkbox"/> Text type knowledge Other skills:		<input type="checkbox"/> Understanding details <input type="checkbox"/> Text structure <input type="checkbox"/> Grammar knowledge <input type="checkbox"/> Inter-sentential comprehension <input type="checkbox"/> Text based inferencing <input type="checkbox"/> Skimming <input type="checkbox"/> Writer's Attitude <input type="checkbox"/> Reading speed			

Appendix J: Rating scale for the summary tasks

	Advanced 5	Upper intermediate 4	Intermediate 3	Lower intermediate 2	Beginner 1
Task achievement					
Coherence and cohesion					
Lexical resourcefulness					
Grammatical range and accuracy					

Appendix K: Definitions, examples, and distribution of the processes and strategies used in Searching Theme

1. Definition of the processes and strategies used in Searching Theme along with some actual samples from the transcripts.

Table 1

Definition and examples of search processes used by different participants

Search processes	Definition	Examples
1. Paragraph level search	The TT reads through a paragraph searching for the relevant information.	<i>"I started reading the first paragraph. I found the relevant info in the fourth line of the first paragraph"</i> (Abed)
2. Scanning specific info in the text	The TT looks for some words or phrases from the task in the text.	<i>"I went back and re-read it and I tried to see what the text is about and then I focused about the "doubt" mentioned to see what the answer is."</i> Zade
3. Searching several paragraphs for the relevant information	The reader skims 2-3 paragraphs to locate the relevant information or check if there is some contradictory information.	<i>"I read the text to see if I am right. I read as far I could find the answer to the questions. I did not read more than the first few paragraphs."</i> (Abed)
4. Reading all the paragraph carefully*	The TT reads all the paragraph carefully.	<i>I read the whole paragraph carefully just once. For paragraph A and B, I read them carefully."</i> (Nabi0)
5. Using previous reading	The TT makes use of his/her comprehension of the text from the first reading.	<i>"I could locate the information very quickly because I knew it from previous reading. I read the relevant sentences very quickly just needed to remind myself what they said. It was mostly skimming"</i> (Abed)
6. Skimming the text to locate relevant paragraph(s)	The TT skims the text or the paragraph to locate the relevant info.	<i>"I skimmed the paragraph, no careful reading, to get the main point."</i> (Mosa)
7. Sentence level search	The TT reads a sentence from a paragraph to find the relevant information.	<i>"then I continued reading until I came to the last blank (not only a great number of people and a great deal of blank-and in here in this sentence in the passage it says a lot of people and a lot of wood so I choose wood as an answer."</i> (Cour)
8. Reading more text to make sure nothing is missed	The TT reads more sentences or paragraphs to make sure s/he is not missing anything, and s/he has got the answer right.	<i>"I should also mention that I looked at some other paragraphs to make sure the name is not repeated there."</i> (Alip)
9. Careful reading	The TT reads the text carefully word by word to develop an overall understanding of the text.	<i>"Then I read through the text word by word from the very beginning to the end as I normally read a text."</i> (Broo)
10. Focusing on doing the task not comprehension of the text	The TT focuses on reading and doing the test task instead of reading for comprehension of the text.	<i>"I just tried to fill in the blanks. It made no sense to me. I looked for this word in the text and match it with the words in the textbox."</i> Angl
11. Struggling with comprehension of the paragraph(s)	The TT struggles with some words/phrases and cannot relate them to make sense of what is said.	<i>"I had to read this over and over trying to understand it but with a lot of vocabulary I could not understand."</i> Angl
12. Text level search	The TT goes through the whole text to locate the relevant info.	<i>"I went through the paragraphs to choose the best heading for the</i>

passage, but I relied on my overall understanding of the text to choose an answer” (Jaha)

759065

Table 6. 24.

Definition and examples of answer processes used by different participants

Answering Processes	Definition	Examples
1. phrase level answer	The TT gets the answer from a single phrase in the text.	<i>“I read the first question and searched for exact words “attached to hull by wires”. I located it and re-read it several times to see what the exact word missing is. Lifting wires was there. “By wires” was there and I though lifting frame should be the answer.” Zade</i>
2. sentence level answer	The TT finds the answer in one single sentence in the paragraph.	<i>“For the last blank, it was in the same sentence and I could easily find it. This was the easiest.” Mosa</i>
3. paragraph level answer	The TT finds the answer in the sum of ideas expressed in the paragraph.	<i>“Paragraph F, here they talk about rats and so the rats destroying everything and eating everything, bird, they were destroying. It is destruction and it was not controllable by them, so I chose vii, (destruction outside their control”. This is the context of the paragraph that helps see all this.” Angl</i>
4. several paragraphs level	The TT refers to the whole text or different parts of the text in answering the question.	<i>“I just read the question and the options. I did not go back to the text. I had enough ideas in mind to decide on the correct answers. So, I chose the options based on my understanding of the text. I should say I am not sure if the answers are right.” Pari</i>
5. attention to literal meaning	The TT makes use of the exact meaning expressed in the text to answer the question.	<i>“The next sentence 33 says works of art contain certain repeated blank and the third last paragraph says contained repeated images then I put images for the blank” Nabil</i>
Using vocabulary knowledge		
6. answer question without re-reading the text	The TT answers the question based on the initial reading without going back to the text.	<i>“I remembered they said they did not need so many people. I chose B first.” Angl “I just read through the summary then I just filled in the words that came to my mind. They were very straight forward and I did not have any doubt with any of them. As soon as I read each sentence, I knew from context what each blank needs. I even did not look at the text at all.” Jack</i>
7. careful reading of the relevant info	After locating the relevant info through skimming/ scanning, etc., the test taker reads the relevant info carefully	<i>“The dates in the table are all in order so I started with 1836. I found the</i>

		<i>relevant date in the text and I read carefully” Abed</i>
8. inferencing	As part of getting the answer to a question, the test taker infers meaning from a phrase, sentence, or paragraph in the relevant info. For example, “I saw it as an intentional search”.	<i>“Paragraph D, this was easy too, I put it as an option, but I then crossed it out. They found out, the archeologists, that they were developing instruments to make their life easier and the environment and solve the problems they were facing. So they created this stones and wind break inside them and many other innovations. These are innovative, they were doing something to solve the problems they were facing.” Angl</i>
9. paraphrasing	The TT paraphrases the question or the relevant info in the text to get the answer.	<i>“and then in the second blank it says the islanders were no longer able to build blank so I skimmed again through and I it was actually in the same sentence where I had stopped which says they could no longer construct wooden canoes for fishing and so here after it mentions they needed to go fishing so I chose canoe for 22.”Nabi</i>
10. recall info from the text	The TT answers the question based on what he has in mind from the text.	<i>“For 38, as I said I got the general idea and memorized it. So, I remembered no idea in the text that talks about cultural element into account. So, I knew I will give it a Not Given.” Mart</i> <i>“Then I scanned the third paragraph. I remembered that there is one person mentioned in the text who found Mary Rose. I read more carefully, and I could get the answer.” Zade</i>
11. re-reading the paragraph or the relevant info	The TT re-read some parts of the paragraph or some sentences several time to get to the answer.	<i>“I had to read every sentence several times but there were some words in each paragraph that directed me.” Zade</i> <i>“To answer the questions in the p I read these paragraphs like 5 times.” Nabi</i>
12. lexical inferencing	The reader connects words from the text to words in the questions and makes inferences how they are connected. For instance, the word “vary” indicates “doubt”.	<i>“Item 36 talks about peoples taste and artistic trends of the period. I continued reading the text. I read, “While the fashions of the time might shape what is currently popular, works that are best adapted to our visual system” it says might shape. It does not say it depends. It might be related not dependent, therefore I said No.” Hona</i>
13. highlighting the answer in the text	The test taker highlights, underlines the key phrase or sentence that he thinks contains the answer.	<i>As I was reading, I highlighted the parts pertained to the answers and I got to the end I got anything that is relevant to this item then I answered it.” Nabi</i>

Answer the question while reading	The TT answers the questions as he read. He has the questions in mind and can answer them as he reads and encounters information relevant to the question.	<i>"I answered questions 2 and 3 while I was reading bk the names and the dates were just there and I said, yeah this is the answer. Right off the hook."</i> Broo
-----------------------------------	--	--

2. Distribution of the process and strategies used in Searching Theme across test takers and test tasks

Table 2

Results of frequencies of the Searching Theme across participants and test tasks

	L1 Test takers	L2: More successful	L2: Less successful	Total
1. Search at paragraph level				
<i>True, False, Not given</i>	14	12	7	33
<i>Matching Features</i>	11	10	7	28
<i>Diagram Completion</i>	13	6	11	30
<i>Matching Headings</i>	41	20	3	64
<i>Summary Completion 1</i>	5	4	4	13
<i>MC- two answers</i>	1	1	2	4
<i>Multiple-Choice</i>	27	15	11	53
<i>Summary Completion 2</i>	7	4	5	16
<i>Yes, No, Not given</i>	11	12	6	29
Total	130	84	56	270
Mean frequency per person	13	16.8	4.3	
2. Search several paragraphs				
<i>True, False, Not given</i>	3	1	5	9
<i>Matching Features</i>	5	3	3	11
<i>Diagram Completion</i>	4	2	1	7
<i>Matching Headings</i>	0	2	0	2
<i>Summary Completion 1</i>	2	0	0	2
<i>MC- two answers</i>	5	4	4	13
<i>Multiple-Choice</i>	3	2	1	5
<i>Summary Completion 2</i>	2	0	0	2
<i>Yes, No, Not given</i>	2	0	1	3
Total	26	14	15	54
Mean frequency per person	2.6	2.8	2.5	
3. Sentence level search				
<i>True, False, Not given</i>	0	1	0	1
<i>Matching Features</i>	2	0	0	2
<i>Diagram Completion</i>	0	0	2	2

<i>Summary Completion 1</i>	6	2	0	8
<i>MC- two answers</i>	1	0	0	1
<i>Multiple-Choice</i>	1	1	1	3
<i>Summary Completion 2</i>	2	2	1	5
<i>Yes, No, Not given</i>	0	1	1	2
Total	12	7	5	24
Mean frequency per person	1.2	1.4	0.83	
4. Scanning specific words, phrases, or ideas				
<i>True, False, Not given</i>	5	2	6	13
<i>Matching Features</i>	17	5	6	28
<i>Diagram Completion</i>	14	2	6	22
<i>Matching Headings</i>	0	1	0	1
<i>Summary Completion 1</i>	4	2	1	7
<i>MC- two answers</i>	3	3	2	8
<i>Multiple-Choice</i>	21	6	8	35
<i>Summary Completion 2</i>	7	5	4	16
<i>Yes, No, Not given</i>	13	3	7	23
Total	84	29	40	153
Mean frequency per person	8.4	5.8	.83	
5. Read the whole paragraph carefully				
<i>True, False, Not given</i>	0	0	0	0
<i>Matching Features</i>	0	3	0	3
<i>Diagram Completion</i>	1	0	0	1
<i>Matching Headings</i>	6	11	22	39
<i>Summary Completion 1</i>	2	1	3	6
<i>MC- two answers</i>	0	0	3	3
<i>Multiple-Choice</i>	3	1	2	6
<i>Summary Completion 2</i>	2	1	5	8
<i>Yes, No, Not given</i>	3	1	1	5
Total	17	18	36	71
Mean frequency per person	1.7	3.6	6	
6. Careful reading				
<i>True, False, Not given</i>	1	1	3	5
<i>Matching Features</i>	2	1	3	6
<i>Diagram Completion</i>	2	2	2	6
<i>Matching Headings</i>	1	1	4	6
<i>Summary Completion 1</i>	1	2	3	6
<i>MC- two answers</i>	0	2	0	2

<i>Multiple-Choice</i>	0	1	3	4
<i>Summary Completion 2</i>	1	0	1	2
<i>Yes, No, Not given</i>	0	0	3	3
Total	8	10	22	40
Mean frequency per person	0.8	2	3.6	
7. Use previous reading				
<i>True, False, Not given</i>	1	0	0	1
<i>Matching Features</i>	2	1	0	3
<i>Diagram Completion</i>	1	0	0	1
<i>Matching Headings</i>	2	0	0	2
<i>Summary Completion 1</i>	3	2	2	7
<i>MC- two answers</i>	6	4	2	12
<i>Multiple-Choice</i>	6	3	2	11
<i>Summary Completion 2</i>	6	2	0	8
<i>Yes, No, Not given</i>	1	0	0	1
Total	28	8	7	46
Mean frequency per person	2.8	1.6	1.1	
8. Skim the text to locate the relevant information				
<i>True, False, Not given</i>	5	3	3	11
<i>Matching Features</i>	5	3	0	8
<i>Diagram Completion</i>	2	0	2	4
<i>Matching Headings</i>	0	0	0	0
<i>Summary Completion 1</i>	4	2	1	7
<i>MC- two answers</i>	2	0	2	4
<i>Multiple-Choice</i>	4	2	0	6
<i>Summary Completion 2</i>	4	3	0	7
<i>Yes, No, Not given</i>	4	2	1	7
Total	30	15	9	54
Mean frequency per person	3	3	1.5	
9. Struggle with comprehension of the paragraph				
<i>True, False, Not given</i>	0	0	2	2
<i>Matching Features</i>	0	0	3	3
<i>Diagram Completion</i>	3	3	8	14
<i>Matching Headings</i>	0	2	4	4
<i>Summary Completion 1</i>	0	0	1	1
<i>MC- two answers</i>	0	0	2	2
<i>Multiple-Choice</i>	0	0	2	2

<i>Summary Completion 2</i>	0	0	0	0
<i>Yes, No, Not given</i>	0	0	6	6
Total	3	5	28	34
Mean frequency per person	0.3	1	4.6	
10. Focus on doing the task				
<i>True, False, Not given</i>	1	1	0	2
<i>Matching Features</i>	1	0	3	4
<i>Diagram Completion</i>	0	2	5	7
<i>Matching Headings</i>	0	1	0	1
<i>Summary Completion 1</i>	0	0	1	1
<i>Multiple-Choice</i>	1	2	1	4
<i>Yes, No, Not given</i>	0	0	2	2
Total	3	6	12	21
Mean frequency per person	0.3	1	2.4	
11. Read more to make sure nothing is missing				
<i>True, False, Not given</i>	4	1	2	7
<i>Matching Features</i>	0	1	0	1
<i>Diagram Completion</i>	1	0	3	4
<i>Matching Headings</i>	2	2	0	4
<i>Multiple-Choice</i>	2	0	2	4
<i>Yes, No, Not given</i>	1	0	1	2
Total	10	4	8	22
Mean frequency per person	1	0.8	1.3	
Total	351	200	238	789
Mean frequency per person	35	40	40	

Appendix L: Definitions, examples, and distribution of the processes and strategies used in Answering Theme

1. Definitions and examples of the processes and strategies used in Answering Theme

Answering Processes	Definition	Examples
1. phrase level answer	The TT gets the answer from a single phrase in the text.	<i>"I read the first question and searched for exact words "attached to hull by wires". I located it and re-read it several times to see what the exact word missing is. Lifting wires was there. "By wires" was there and I though lifting frame should be the answer." Zade</i>
2. sentence level answer	The TT finds the answer in one single sentence in the paragraph.	<i>"For the last blank, it was in the same sentence and I could easily find it. This was the easiest." Mosa</i>
3. paragraph level answer	The TT finds the answer in the sum of ideas expressed in the paragraph.	<i>"Paragraph F, here they talk about rats and so the rats destroying everything and eating everything, bird, they were destroying. It is destruction and it was not controllable by them, so I chose vii, (destruction outside their control". This is the context of the paragraph that helps see all this." Angl</i>
4. several paragraphs level	The TT refers to the whole text or different parts of the text in answering the question.	<i>"I just read the question and the options. I did not go back to the text. I had enough ideas in mind to decide on the correct answers. So, I chose the options based on my understanding of the text. I should say I am not sure if the answers are right." Pari</i>
5. attention to literal meaning	The TT makes use of the exact meaning expressed in the text to answer the question.	<i>"The next sentence 33 says works of art contain certain repeated blank and the third last paragraph says contained repeated images then I put images for the blank" Nabil</i>
Using vocabulary knowledge		
6. answer question without re-reading the text	The TT answers the question based on the initial reading without going back to the text.	<i>"I remembered they said they did not need so many people. I chose B first." Angl "I just read through the summary then I just filled in the words that came to my mind. They were very straight forward and I did not have any doubt with any of them. As soon as I read each sentence, I knew from context what each blank needs. I even did not look at the text at all." Jack</i>
7. careful reading of the relevant info	After locating the relevant info through skimming/ scanning, etc., the test taker reads the relevant info carefully	<i>"The dates in the table are all in order so I started with 1836. I found the relevant date in the text and I read carefully" Abed</i>

8. inferencing	As part of getting the answer to a question, the test taker infers meaning from a phrase, sentence, or paragraph in the relevant info. For example, "I saw it as an intentional search".	<i>"Paragraph D, this was easy too, I put it as an option, but I then crossed it out. They found out, the archeologists, that they were developing instruments to make their life easier and the environment and solve the problems they were facing. So they created this stones and wind break inside them and many other innovations. These are innovative, they were doing something to solve the problems they were facing."</i> Angl
9. paraphrasing	The TT paraphrases the question or the relevant info in the text to get the answer.	<i>"and then in the second blank it says the islanders were no longer able to build blank so I skimmed again through and I it was actually in the same sentence where I had stopped which says they could no longer construct wooden canoes for fishing and so here after it mentions they needed to go fishing so I chose canoe for 22."</i> Nabi
10. recall info from the text	The TT answers the question based on what he has in mind from the text.	<i>"For 38, as I said I got the general idea and memorized it. So, I remembered no idea in the text that talks about cultural element into account. So, I knew I will give it a Not Given."</i> Mart <i>"Then I scanned the third paragraph. I remembered that there is one person mentioned in the text who found Mary Rose. I read more carefully, and I could get the answer."</i> Zade
11. re-reading the paragraph or the relevant info	The TT re-read some parts of the paragraph or some sentences several time to get to the answer.	<i>"I had to read every sentence several times but there were some words in each paragraph that directed me."</i> Zade <i>"To answer the questions in the p I read these paragraphs like 5 times."</i> Nabi
12. lexical inferencing	The reader connects words from the text to words in the questions and makes inferences how they are connected. For instance, the word "vary" indicates "doubt".	<i>"Item 36 talks about peoples taste and artistic trends of the period. I continued reading the text. I read, "While the fashions of the time might shape what is currently popular, works that are best adapted to our visual system" it says might shape. It does not say it depends. It might be related not dependent, therefore I said No."</i> Hona
13. highlighting the answer in the text	The test taker highlights, underlines the key phrase or sentence that he thinks contains the answer.	<i>As I was reading, I highlighted the parts pertained to the answers and I got to the end I got anything that is relevant to this item then I answered it."</i> Nabi
Answer the question while reading	The TT answers the questions as he read. He has the questions in mind and can answer	<i>"I answered questions 2 and 3 while I was reading bk the names and the dates</i>

them as he reads and encounters information relevant to the question.	<i>were just there and I said, yeah this is the answer. Right off the hook.</i> ” Broo
---	--

2. Distribution of “answering Theme” across different test tasks

Table 2

Results of processes and strategies used in Answering Theme across participants and test tasks

	L1 TT	L2: More successful	L2: Less successful	Total
1. Sentence level answer				
<i>True, False, Not given</i>	11	4	6	21
<i>Matching Features</i>	9	3	3	15
<i>Diagram Completion</i>	18	8	3	29
<i>Matching Headings</i>	10	4	0	14
<i>Summary Completion 1</i>	9	5	5	19
<i>MC- two answers</i>	2	1	0	3
<i>Multiple-Choice</i>	22	12	6	40
<i>Summary Completion 2</i>	7	2	2	11
<i>Yes, No, Not given</i>	15	8	2	25
Total	103	47	27	177
Mean frequency per person	10.3	9.4	4.5	
2. Paragraph level answer				
<i>True, False, Not given</i>	5	5	0	10
<i>Matching Features</i>	10	7	3	20
<i>Diagram Completion</i>	0	0	1	1
<i>Matching Headings</i>	35	16	6	57
<i>Yes, No, Not given</i>	9	3	3	15
Total	59	31	13	103
Mean frequency per person	5.9	6.2	2.1	
3. Several paragraph level				
<i>Matching Headings</i>	1	1	0	2
<i>MC- two answers</i>	6	2	4	12
<i>Multiple-Choice</i>	4	2	5	11
Total	11	5	9	25
Mean frequency per person	1.1	1	1.5	
4. Phrase level answer				
<i>Diagram Completion</i>	1	1	2	4
<i>Summary Completion 1</i>	2	0	1	3
<i>Multiple-Choice</i>	2	0	0	2

<i>Summary Completion 2</i>	1	1	0	2
<i>Yes, No, Not given</i>	3	1	0	2
Total	9	3	3	15
Mean frequency per person	0.9	0.6	0.5	
5. Use of vocabulary knowledge				
<i>True, False, Not given</i>	0	2	0	2
<i>Matching Features</i>	4	5	4	13
<i>Diagram Completion</i>	21	8	21	50
<i>Matching Headings</i>	8	11	8	27
<i>Summary Completion 1</i>	9	6	9	24
<i>Multiple-Choice</i>	3	7	3	21
<i>Summary Completion 2</i>	5	1	5	11
<i>Yes, No, Not given</i>	2	4	0	6
Total	52	44	50	146
Mean frequency per person	5.2	8.8	8.3	
6. Literal meaning				
<i>True, False, Not given</i>	1	1	0	2
<i>Matching Features</i>	0	2	2	4
<i>Diagram Completion</i>	4	7	4	15
<i>Matching Headings</i>	1	1	0	2
<i>Summary Completion 1</i>	10	3	1	14
<i>Multiple-Choice</i>	6	5	0	11
<i>Summary Completion 2</i>	6	5	2	13
<i>Yes, No, Not given</i>	2	2	2	6
Total	30	26	11	67
Mean frequency per person	3.0	5.2	1.8	
7. Careful reading of the relevant information				
<i>True, False, Not given</i>	2	2	4	8
<i>Matching Features</i>	6	2	3	11
<i>Diagram Completion</i>	1	0	0	1
<i>Matching Headings</i>	0	1	0	1
<i>Summary Completion 1</i>	3	0	1	4
<i>MC- two answers</i>	2	3	2	7
<i>Multiple-Choice</i>	11	3	4	18
<i>Summary Completion 2</i>	2	0	3	5
<i>Yes, No, Not given</i>	9	3	5	17
Total	36	14	22	72

Mean frequency per person	3.6	2.8	3.6	
8. Inferencing				
<i>True, False, Not given</i>	1	2	0	3
<i>Matching Features</i>	1	0	2	3
<i>Diagram Completion</i>	1	0	0	1
<i>Matching Headings</i>	15	5	0	20
<i>MC- two answers</i>	8	2	2	12
<i>Multiple-Choice</i>	9	1	0	10
<i>Yes, No, Not given</i>	19	10	2	31
Total	54	20	6	80
Mean frequency per person	5.4	4.0	1	
9. Paraphrasing				
<i>Matching Features</i>	0	0	2	2
<i>Diagram Completion</i>	7	1	2	10
<i>Matching Headings</i>	4	1	0	5
<i>Summary Completion 1</i>	6	5	1	12
<i>MC- two answers</i>	1	1	0	2
<i>Multiple-Choice</i>	3	5	0	8
<i>Summary Completion 2</i>	3	1	2	8
<i>Yes, No, Not given</i>	5	1	0	6
Total	29	15	7	49
Mean frequency per person	2.9	3.0	1.1	
10. Recall info from the text				
<i>True, False, Not given</i>	8	1	0	7
<i>Matching Features</i>	2	4	2	8
<i>Diagram Completion</i>	1	0	0	1
<i>Matching Headings</i>	4	0	0	4
<i>Summary Completion 1</i>	4	1	1	6
<i>MC- two answers</i>	4	3	0	7
<i>Multiple-Choice</i>	1	0	0	1
<i>Summary Completion 2</i>	8	1	1	10
<i>Yes, No, Not given</i>	5	2	0	7
Total	37	12	4	53
Mean frequency per person	3.7	2.4	.66	
11. Re-reading the paragraph or the relevant info				
<i>True, False, Not given</i>	3	1	7	8
<i>Matching Features</i>	1	2	8	9

<i>Diagram Completion</i>	5	3	6	12
<i>Matching Headings</i>	6	2	6	11
<i>Summary Completion 1</i>	1	0	1	2
<i>MC- two answers</i>	1	1	2	4
<i>Multiple-Choice</i>	0	0	4	2
<i>Yes, No, Not given</i>	1	2	6	2
Total	18	11	40	69
Mean frequency per person	1.8	2.2	6.66	
12. Lexical inferencing				
<i>True, False, Not given</i>	4	6	1	11
<i>Matching Features</i>	3	0	1	4
<i>Diagram Completion</i>	3	3	4	10
<i>Matching Headings</i>	2	0	0	2
<i>Multiple-Choice</i>	1	0	0	1
<i>Yes, No, Not given</i>	1	0	0	1
Total	14	9	6	29
Mean frequency per person	1.4	1.8	1.0	
13. Highlighting the answer in the text				
<i>True, False, Not given</i>	2	0	0	2
<i>Matching Features</i>	1	0	0	1
<i>Diagram Completion</i>	0	1	1	2
<i>Summary Completion 1</i>	0	0	1	1
<i>Multiple-Choice</i>	2	0	0	2
<i>Summary Completion 2</i>	0	1	0	1
<i>Yes, No, Not given</i>	2	1	0	3
Total	7	3	2	12
Mean frequency per person	0.7	0.2	0.3	
14. Answering the question while reading				
<i>True, False, Not given</i>	2	0	0	2
Total	2	0	0	2
Mean frequency per person	0.2	0.0	0.0	

Appendix M: Definitions and examples of the metacognitive strategies used

1. Definitions and examples of the metacognitive strategies used in test takers' performance

Table 1. Definitions and *actual examples of metacognitive strategies used*

Strategies	Definition	Sample examples from participants' transcripts
1. Using paragraph/ text structure knowledge	The TT makes use of text type knowledge in searching for an answer. For instance, he is aware of the chronology of events in the Mary Rose text and chooses to read the end of the passage not the beginning for answering certain questions.	<p><i>"this time I did not read the whole text because I remembered that there are two paragraphs that talk about all these situations."</i> (Broo)</p> <p><i>"I know to look just in the first paragraph because it talked about the battle and I knew I did not need to read the whole paragraph."</i> (Hele)</p> <p><i>"And I also noticed that questions 26, 27, and 28 that all looked at the results. I checked the end of each paragraph because it is where the results are put."</i> (Rich)</p>
2. Eliminating/ narrowing down the options	The TT eliminates some of the options/distractors after careful analysis and gets to the answer.	<p><i>"The options also helped to choose the answers. Option A for example, was not discussed in the text at all. Option D had an unfamiliar word for me, the word "carved" I do not know what it means."</i> (Abed)</p> <p><i>"In some items I felt I need to read the whole paragraph or the relevant info twice. I tried to eliminate some of the options for each question because they were very irrelevant, or I could easily reject and mostly I ended up choosing between two options. For the final answer, I would go back to the text and re-read the relevant info to make the final choice."</i> (Beig)</p>
3. Developing a gist of the text.	The TT focuses on the main point of the text not details. He keeps the chronology of the main events in mind but not the details.	<p><i>"I looked at the task first and then read the whole text carefully every sentence and paragraph to get an overall understanding."</i> (Angl)</p> <p><i>"I read through the paragraph after I read the question and looked for the response options and I chose A. That was pretty easy to respond to. So, I read the paragraph and kind of summarized it for myself and went back to the question."</i> (Isab)</p> <p><i>"I read these paragraphs one by one thoroughly. I tried to generate a general concept and use that general concept to provide my heading answers."</i> (Rich)</p> <p><i>"I usually read all the paragraph carefully, especially the first and last two sentences to get a good grasp of the paragraph. The point is to see what the paragraph is saying."</i> (Abed)</p> <p><i>I had the main idea of the paragraph in mind which I considered in my answers. The idea was kind of providing different pieces of evidence and support."</i> (Hona)</p>
4. Delay answering	The TT is not sure if the answer is right, so he delays it. Or he does not know the answer at all and delays answering it. Hoping that he will find the answer as he does other items.	<p><i>"I read it and delayed it because I was not sure and I did not remember and I was short of time and thought I read the other tasks first."</i> (Angl)</p> <p><i>"Then as I could not do all the headings and paragraphs, so I decided to go to the next task- the summary task and go back to them later."</i> (Khei)</p>
5. applying test wiseness strategy	The TT uses his test taking experience and knowledge while doing the test task. For instance, he considers	<p><i>"I could not get a clear idea of the text so I could not rely on my guesses. I went back to the previous blank, I though it should be related to paragraph 1 because there was no question related to this paragraph in other tasks."</i> (Hash)</p>

	the order of the questions and the sequence of ideas discussed in the text. He knows that the questions are not in order.	<p><i>"Knowing that the task is basically a copy of the original text, I focused on "for example" I searched for "for example" in the paragraph and then the other words an I could find the word that fits the blank." (Hash)</i></p> <p><i>"I knew the task would be dealing with dates, so I paid attention to the dates mentioned." (Mitic)</i></p>
6. Using background knowledge	The TT makes reference to his background knowledge in answering the question.	<p><i>"I noticed protection and airbags are related to one another. We know airbags are sued for protection, so I chose it." (Beig)</i></p> <p><i>"I am not sure of the answer, but I had an idea what a lifting frame might look like." (Jack)</i></p>
7. Attention to details	The TT pays attention to the details of the paragraph.	<i>"For paragraph B, it talks about huge trees and grassland and then talking about forests disappearing like how the fish was used for so long and a lot of environmental and food problems and what not so I ended up selecting the last heading "diminishing food resource.." because it talked about food resources. I was confident about this answer." (Cour)</i>
8. Double checking the answer	The TT knows the answer, yet he double checks the answer. Most of the answers prove right.	<p><i>"for the summary task I was looking for words and I already knew the answer. I basically checked the answers." (Mitic)</i></p> <p><i>"I kind of knew the answer is False after reading the first two paragraphs but I just double checked and postpone the answer after finished reading the whole text." (Nabi)</i></p>
9. Guessing the answer	The TT guesses what the answer might be.	<p><i>"I get an understanding of the whole sentence. For both 22 and 23, I could recall information from the text and guess what the answer is. Then again, I went back to the passage and double checked and they were correct." (Abed)</i></p> <p><i>"I did the last question based on my understanding of the passage. I knew I had not read the whole passage yet, but I just guessed and answered it based on my overall comprehension and I chose a subtitle." (Alp)</i></p>
10. Re-phrasing or rereading the question	The TT reads the question and analyses and paraphrases it in more tangible and sensible terms and ideas for himself.	<p><i>"I was looking for some objects that match the description." (Aabed)</i></p> <p><i>"First I read the question and then highlighted the key words in the question which was "shape-matching" and I turned it to a question to myself to make things clear for myself what I should be specifically answering." (Beig)</i></p>
11. Planning a strategy	The TT plans how to do the item/ test task. He decides to use a certain strategy to do the task. For instance, "I decided to skim for dates in the text."	<p><i>"because of the task and how it is divided up by each task I decided to approach it differently. I was trying to answer some specific info in Mary rose but here I was trying to develop a general understanding of each paragraph. I have to say I still tried to develop an overall understanding of what each paragraph is about." (Rich)</i></p> <p><i>"I decided to skim through the text to find the dates." (Mart)</i></p>
12. Developing and using text representation of the text	The TT reads the text and develops a representation of the paragraph/paragraphs or the whole text and knows what has happened, where, and how ideas in the paragraph(s) are related. He recalls the main points and some details from the paragraph in his retrospection. This	<p><i>"I had the information in mind. I mostly memorized what information is in each paragraph and I went back to them to answer the question more carefully" (Rich)</i></p> <p><i>"I read the whole text and I understood what it is all about. It read carefully. The most difficult part of the text for me was the last paragraph because it was too technical, I know the main points the chronological order of events and that they could finally take the</i></p>

	representation is used in searching for relevant information and answering some of the questions.	<i>ship out of the sea. It was a successful task. I could understand the text.” (Angl)</i>
13. Highlighting the answer in the text	As he reads the text, the TT underlines some words, phrases. He thinks they are important and might be used in answering some questions.	<p><i>“So if I have circled or underlined the names and the dates I can then know where in the paragraph I can find the information.” (Pari)</i></p> <p><i>“I could read this attached to the hull by wires so I went back to my underlined things in the passage to see which one fits. And I did this for each question.” (Broo)</i></p>
14. Careful analysis of the options	The TT reads the options carefully and explains why they are right or wrong.	<p><i>“I also eliminated D and E because they did not talk much about the ideas in D and E in the passage. The ideas in B and C were strongly talked about but not D and E.” (Cour)</i></p> <p><i>“So I read the options before I go to the text. I noticed moai is repeated in all the option which indicates the focus is on moai or the islanders. Role of the moai in the destruction of the island was something I had in mind and I read parts of paragraph D which discusses moai differently. So, I knew C is one of the answers, the role of moai on Rapanui society. B was another answer which I could find in paragraph E which says, “they also believed that the moai were moved with few people and no wood”. Diamond, on the other hand, believed it required a lot of people and a lot of wood.” (Jah)</i></p>
15. Attention to certain features of the text	The TT pays more attention to certain features of the text. For instance, dates, numbers, and proper names are underlined.	<i>“So, when I read the text, I took note of the dates.” (Rich)</i>
16. Recalling information from previous reading	The TT reads the question and makes reference to ideas from previous reading.	<p><i>“For item 30, I did not read anything again because after having read it the first time and a couple of times for the other tasks. I read through all the options and answered the task” (Mitic)</i></p> <p><i>“For question 8, I could answer it quickly because when I was reading for question 5, I had read the paragraph that could help me answer question 8.” (Hash)</i></p>
17. Noticing grammar	The TT notices an element of grammar at word or sentence level. For example, he says, we need a verb here.	<i>“In this case blank 31, I filled it in with the word emotions. (reading the sentence) Neurological studies of the brain, for example, demonstrate the impact which Impressionist paintings have on our emotions- have on emotions” then reading all through the options, it was one of the only words that would have really made sense in the space grammatically and based on the understanding that I got from the main passage” (Kyle)</i>
18. Read the text with the questions in mind	The test takers read the item and goes to the text and reads it.	<i>I read the questions, so I have an idea what each question was asking. So that when I began to read, I know what I want to find as an answer. (Isab)</i>
19. Attention to the repetition of inter-related words (lexical cohesion)	The TT pays attention to repetition of certain words that are semantically related to one another. For example, evidence, question, answer	<p><i>“Paragraph D, I could find a match i “Evidence of innovative environment management practices”. The par mentions several examples for these practices; “they believe the islanders were not to blame or “heroic efforts to protect resources for their wild .. in fertile fields. There is also a mention of “they built thousands of ...” or “they used unbroken volcanic rocks to keep the soil...” all these are supporting innovative environment management. All these ideas are inter-connected, and you need to relate them and put them all together to choose the answer. You need to piece these together.” (Hona)</i></p> <p><i>“I needed to read the whole text carefully to het the answer. This is why I highlighted some words in each paragraph and kind of connected them to each other to get a certain theme that integrates</i></p>

		<i>these words together. For instance, in paragraph B there are a number of words about food, fishing, cannibalism. Therefore, I thought It should be related to heading ix- diminishing food resources.” (Jaha)</i>
20. More attention to the topic or concluding sentences	The TT reads the paragraph but pays more attention to the first and last sentences in the paragraph(s) to get an overall meaning of the paragraph.	<i>“I often knew that the beginning and the last sentence are kind of really important in knowing what this paragraph is about and the concluding sentence at the end.” (Cour)</i> <i>“I was more careful with the first few sentences in each paragraph because this is where you may find the main idea.” (Hona)</i>
21. Recalling the questions while reading	The TT recalls the question while reading the whole text. Before reading the text, she has looked at the questions.	<i>I went back to the text to answer question 5 but I started by answering question 8 because I had read the relevant info for it and I answered it first. (Jaha)</i> <i>“Originally, I was looking at the first question but then I realized because question 8 was in the back of my mind it was question 8 that first popped up in my mind.” (Isab)</i>
22. Answering without going back to the text	The TTs immediately answers the question without going back to find the answer or double check the answer.	<i>when I went to the items, 1,3 and 4, I answered them pretty much immediately. I knew the answer. I remembered the answer exactly being discussed in the text. I knew each one of those in the specific parts in the text that discussed them. (Jack)</i>
23. Noticing the order of the questions in the text	The TT notices the order of the questions and how they are related to the sequence of ideas presented in the paragraph(s).	<i>“Then I moved down to item 7, which is -it was agreed that the hull of the Mary Rose should be raised- so I knew that they are timeline wise a chronologically ordered so I went towards the end of the passage ...” (Cour)</i> <i>“I knew the questions were not in order which was confusing.” (Isab)</i> <i>“I knew from doing the previous test that the questions are in order in the text.” (Isab)</i>
24. Re-reading the last few sentences in each paragraph	The TT re-read the last few sentences in the paragraph because it might summarize the main idea of the paragraph.	<i>“For item 30, since I had not read all the paragraphs, I read and re-read the last sentence of each paragraph and answered the question.” (Hash)</i>
25. Tagging paragraphs	The TT tags each paragraph with words and phrases that summarize main idea of the paragraph.	<i>“as I read each paragraph at the end, I tagged it with a term that covered what it was talked about.” (Broo)</i>
26. Using the heading to understand the paragraph	After failing to comprehend the gist of the paragraph, the TT reads the headings for clues that can help him understand and explain the main points of the paragraph.	<i>“I used the headings to make sense of the ideas in the paragraph. For instance, when I read paragraph A, I went to the heading to see which one fits it best.” (Mosa)</i>