



The impact of computer interface design on Saudi students' performance on a L2 reading test

Serge Korevaar

This is a digitised version of a dissertation submitted to the University of Bedfordshire.

It is available to view only.

This item is subject to copyright.

THE IMPACT OF COMPUTER INTERFACE DESIGN ON
SAUDI STUDENTS' PERFORMANCE ON A L2 READING TEST

S. A. A. Korevaar

Ph.D

2015

UNIVERSITY OF BEDFORDSHIRE

THE IMPACT OF COMPUTER INTERFACE DESIGN ON
SAUDI STUDENTS' PERFORMANCE ON A L2 READING TEST

by

Serge Antonie Aart Korevaar

A thesis submitted to the University of Bedfordshire in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

January 2015

Abstract

This study investigates the effect of testing mode on lower-level Saudi Arabian test-takers' performance and cognitive processes when taking an L2 reading test on computer compared to its paper-based counterpart from an interface design perspective.

An interface was developed and implemented into the computer-based version of the L2 reading test in this study, which was administered to 102 Saudi Arabian University students for quantitative analyses and to an additional eighteen for qualitative analyses. All participants were assessed on the same L2 reading test in two modes on two separate occasions in a within-subject design. Statistical tests such as correlations, group comparisons, and item analyses were employed to investigate test-mode effect on test-takers' performance whereas test-takers' concurrent verbalizations were recorded when taking the reading test to investigate their cognitive processes. Strategies found in both modes were compared through their frequency of occurrence. In addition, a qualitative illustration of test-takers cognitive behavior was given to describe the processes when taking a lower-level L2 reading test. A mixed-method approach was adhered to when collecting data consisting of questionnaires think-aloud protocols, and post-experimental interviews as main data collection instruments.

Results on test-takers' performance showed that there was no significant difference between the two modes of testing on overall reading performance, however, item level analyses discovered significant differences on two of the test's items. Further qualitative investigation into possible interface design related causes for these differences showed no identifiable relationship between test-takers' performance and the computer-based testing mode. Results of the cognitive processes analyses showed significant differences in three out of the total number of cognitive processes employed by test-takers indicating that test-takers had more difficulties in processing text in the paper-based test than in the computer-based test. Both product and process analyses carried out further provided convincing supporting evidence for the cognitive validity, content validity, and context validity contributing to the construct validity of the computer-based test used in this study.

Acknowledgements

I would like to thank my supervisor Professor Stephen Bax for putting up with me for this long stretch, for his invaluable advice throughout the various stages of this study, and for having been tremendously flexible and helpful throughout this study. I would also like to thank my second supervisor Dr. Fumiyo Nakatsuhara for her helpful comments at crucial stages in the quantitative data analysis process.

I would like to thank my parents for everything they did and have done to raise me in a moral and liberal way, always encouraging me to be dedicated and persistent when trying to achieve my goals in life.

I would like to take this opportunity to thank my wife who in addition to being a loving companion and full-time mother had to function as a stand-in father more often than not during these past 4 years of studies. I would like to thank my children for being patient with me and for consistently buying into my excuses when ‘rescheduling’ our quality time for the 1000th time due to having to work late... again. I would like to thank the UOH and, especially, Dr. Eid Al-Haisoni, for facilitating the data collection process at the Preparatory Year Program and for providing me with the resources needed to successfully complete this study’s data collection. I would further like thank my friends and colleagues who helped me at various stages throughout this study; Dr. Saad Al-Amri, Mr. Meteb, Mr. Mustafa George, Professor Motasim Badri, Mr. Paul Condon, Mr. Suleiman Jenkins, Professor Muhammad Khan, Dr. Muhammad Al-Roomy. I would like to show my heartfelt gratitude to this study’s participants for providing me with their ‘thoughts’ which were critical to this study’s outcomes and contributions to the field.

Table of Contents

Abstract	i
Acknowledgements	ii
List of Tables	xi
List of Figures	xiv
List of Abbreviations	xv
Chapter 1: Introduction	1
1.1 Background to the Study	1
1.2 Aims of the Study	4
1.3 Significance of the Topic	6
1.4 Educational System in Target Context	9
1.5 Examination in Saudi Arabia	10
1.6 Research Questions	12
1.7 Thesis Overview	12
Chapter 2: Literature Review	14
2.1 Introduction	14
2.2 L2 Reading Ability	15
2.3 Reading Types	16
2.4 Processes Models in Reading	19
2.5 Process Levels in Reading	21
2.5.1 Lower-Level Processes	21
2.5.1.1 Lexical Access	21
2.5.1.2 Syntactic Parsing	22
2.5.1.3 Establishing Propositional Meaning	22
2.5.1.4 Inferencing at the Local Level	23
2.6 Test Validity	23
2.6.1 Background	24
2.6.2 Concepts of Validity	25
2.7 Theoretical Framework for Conceptualizing Test Validity	28
2.8 Cognitive Validity	31

2.8.1 Establishing Cognitive Validity	32
2.9 Strategies in L2 Reading	38
2.9.1 Distinguishing between Skills & Strategies	38
2.9.2 Defining Test-Taking Strategies.....	40
2.5.2.1 Test-taking Strategies in Reading Assessment	41
2.10 Theoretical Cognitive Processes Model.....	46
2.11 Reading Assessment Format	48
2.11.1 Short Answer Questions (SAQ's)	48
2.12 Overview Reviewed Elements	52
2.13 Computer Interface Design	54
2.13.1 Introduction	54
2.13.2 Earlier Reviews	54
2.14 Interface Evaluation Model.....	57
2.15 Review Interface Design: Presentation (typographical factors).....	59
2.15.1 Font Characteristics	59
2.15.2 Line Length (characters per line)	68
2.15.3 Number of Lines.....	73
2.15.4 Interlinear Spacing.....	75
2.15.5 White Space.....	77
2.15.6 Number of Columns	80
2.15.7 Text/Background	81
2.16 Review User Interface: Presentation (graphical factors).....	84
2.16.1 Screen Size and Resolution	84
2.16.2 Icon & Button Design.....	86
2.17 Review User Interface: Interaction	88
2.17.1 Scrolling	88
2.17.2 Item Review.....	91
2.17.3 Item Presentation	92
2.18 Reviewed Elements and Recommended Settings User Interface.....	95
2.19 Summary of the Chapter	97
Chapter 3: Research Methodology.....	98

3.1 Introduction	98
3.2 Research Questions and Hypotheses.....	99
3.3 Research Design Rationale.....	99
3.4 Frameworks and Design.....	101
3.4 Interface Design	105
3.5 Pilot Study.....	110
3.5.1 Objectives	110
3.5.2 Participants	111
3.5.3 Instruments Pilot Study	111
3.5.4 Interface Design Pilot Study.....	112
3.5.5 Procedure Pilot Study	115
3.6 Results Pilot Study	117
3.6.1 Computer Familiarity Questionnaire Results	117
3.5.4.3 Usability Questionnaire	119
3.6.2 Usability Questionnaire Results	121
3.6.2.1 Buttons	121
3.6.2.2 Reading Passage Scrolling Feature	122
3.6.2.3 Test Timer	122
3.6.2.4 Screen capture software (SCS).....	123
3.6.2.5 Recording Devices	123
3.7 Implications for the Main Study	124
3.7.1 Computer Familiarity Questionnaire.....	124
3.7.2 Interface Design.....	125
3.8 Summary	125
3.9 Main Study	126
3.9.1 Target Population and Participants.....	126
3.9.2 Permissions.....	127
3.9.3 Informed Consent	127
3.9.4 Instruments	128
3.9.4.1 Study Tests	128
3.9.4.2 Types of Reading and CEF Level	129

3.9.4.3 Reliability Reading Tests	130
3.9.4.4 Test Contents.....	131
3.10 Study Test’s Validity checks.....	132
3.10.1 Face Validity	132
3.10.2 Content Validity	133
3.11 Processes	136
3.11.1 Text Analysis.....	138
3.11.2 Passage Order vs. Item Order.....	143
3.12 Test Administration Procedure.....	144
3.13 Interface Design Main Study.....	146
3.13.1 Interface Design: Presentation.....	147
3.13.1.1 Typographical Factors.....	149
3.13.1.2 Graphical Factors	153
3.13.2 Interface Design: Interaction	155
3.14 Test Analysis	158
3.15 Study Questionnaires.....	158
3.15.1 Questionnaire Design	159
3.15.2 Computer Familiarity Questionnaire (CFQ)	160
3.15.2.1 Administering the CFQ	163
3.15.2.2 Data Analysis CFQ	163
3.15.2.3 Computer Familiarity Scale Development.....	163
3.15.3 Post-Test Questionnaire (PTQ)	167
3.15.3.1 Administering the PTQ	169
3.15.3.2 Data Analysis of the PTQ.....	169
3.16 Interviews	170
3.16.1 Instrument Rationale	170
3.16.2 Procedure	171
3.16.3 Interview Data Analysis	172
3.17 Think Aloud Protocol.....	173
3.17.1 Instrument Rationale	173
3.17.2 Sampling of Think-Aloud Participants.....	174

3.17.3 Instruments used in TA-Sessions	175
3.17.3.1 Reading Passage	175
3.17.3.2 Training Materials TA-Sessions.....	176
3.17.3.3 TA Training Session	177
3.17.3.4 Procedure Think Aloud	177
3.18 Think Aloud Data Analysis.....	178
3.18.1 Protocol Transcription	178
3.18.2 Development of the Coding Scheme	180
3.18.2.1 Segmentation and Coding Stage	180
3.18.2.2 Categorization/Classification Stage	185
3.18.2.3 Strategy Counting.....	186
3.18.3 Reliability Checks	187
3.18.3.1 Intrajudge Reliability.....	188
3.18.3.2 Interjudge Reliability.....	189
3.19 Chapter Summary.....	190
Chapter 4: Study Results Part 1	192
4.1 Introduction	192
4.2 Testing Mode Effect on Test-Taker Performance.....	193
4.2.1 Reliability Figures PBT and CBT	193
4.2.2 Descriptive Statistics PBT	196
4.2.3 Descriptive Statistics CBT	197
4.2.4 Test of Normality	198
4.2.5 Score Comparisons PBT and CBT.....	199
4.2.6 Correlational Analyses PBT and CBT	201
4.2.7 Item Performance Comparison PBT and CBT.....	202
4.2.8 Post-Test Questionnaire.....	204
4.3 Discussion of Results Part 1: Test-Takers' Performance in PBT and CBT	209
4.3.1 Comparability of Scores in PBT and CBT.....	209
4.3.1.1 Reliability in PBT and CBT	209
4.3.1.2 Test-Takers' Performance in PBT and CBT	211
4.4 Summary	214

Chapter 5 Results & Discussion, Part 2a: Comparing Processes.....	215
5.1 Introduction	215
5.2 Overview Overall Strategy Types and Tokens in PBT and CBT	218
5.3 Overview Strategies by Category.....	221
5.3.1 Category 1: Overall Test-Level Strategies	221
5.3.1.1 Descriptive Statistics and Paired Samples T-test’s Results	222
5.3.2 Category Two: Strategies related to Initial Reading of the Passage	224
5.3.2.1 Descriptive Statistics and Paired Samples T-test’s Results	225
5.3.3 Category Three: Strategies related to Reading of Questions.....	227
5.3.3.1 Descriptive Statistics and Paired Samples T-test’s Results	228
5.3.4 Category Four: Strategies related to Reading of Passage.....	230
5.3.4.1 Descriptive Statistics and Paired Samples T-test’s Results	231
5.3.5 Category Five: Strategies related to Aiding in Answering Questions.....	232
5.3.5.1 Descriptive Statistics and Paired Samples T-test’s Results	233
5.3.6 Category Six: Strategies related to Items after having answered them	234
5.3.6.1 Descriptive Statistics and Paired Samples T-test’s Results	235
5.3.7 Category Seven: Supporting Strategies	236
5.3.7.1 Descriptive Statistics and Paired Samples T-test’s Results	237
5.3.8 Category Eight: Executive Strategies	238
5.3.8.1 Descriptive Statistics and Paired Samples T-test’s Results	238
5.3.9 Category Nine: Evaluative Strategies.....	239
5.3.9.1 Descriptive Statistics and Paired Samples T-test’s Results	240
5.3.10 Category Ten: Inferencing Strategies	240
5.3.10.1 Descriptive Statistics and Paired Samples T-test’s Results	241
5.3.11 Category Eleven: Affective Strategies	242
5.3.11.1 Descriptive Statistics and Paired Samples T-test’s Results	242
5.4 Discussion of Results Part 2a: Processes in PBT and CBT	243
5.4.1 Differences Test-Takers’ Processes in PBT and CBT	243
5.5 Results & Discussion 2, Part 2b: Describing Cognitive Processes	247
5.5.1 Introduction	247
5.5.2 Students’ Performance on Think-Aloud Study Test in PBT & CBT	248

5.6 Overview Expeditious Reading Operations	249
5.7 Description Cognitive Processes/Strategies Utilized as per Test-Item	251
5.7.1 Item 11	251
5.7.1.1 Operations/Strategies Item 11	251
5.7.1.2 Descriptive Account Processes/Strategies Item 11	251
5.7.1.3 Levels of Processing Item 11	253
5.7.2 Item 12.....	254
5.7.2.1 Operations/Strategies Item 12.....	254
5.7.2.2 Descriptive Account Processes/Strategies Item 12.....	255
5.7.2.3 Levels of Processing Item 12.....	256
5.7.3 Item 13.....	257
5.7.3.1 Operations/Strategies Item 13.....	257
5.7.3.2 Descriptive Account Processes/Strategies Item 13.....	258
5.7.3.3 Levels of Processing Item 13.....	259
5.7.4 Item 14.....	260
5.7.4.1 Operations/Strategies Item 14.....	260
5.7.4.2 Descriptive Account Processes/Strategies Item 14.....	261
5.7.4.3 Levels of Processing Item 14.....	262
5.7.5 Item 15.....	263
5.7.5.1 Operations/Strategies Item 15.....	263
5.7.5.2 Descriptive Account Processes/Strategies Item 15.....	263
5.7.5.3 Levels of Processing Item 15.....	265
5.7.6 Item 16.....	266
5.7.6.1 Operations/Strategies Item 16.....	266
5.7.6.2 Descriptive Account Processes/Strategies Item 16.....	266
5.7.6.3 Levels of Processing Item 16.....	268
5.7.7 Item 17.....	268
5.7.7.1 Operations/Strategies Item 17.....	269
5.7.7.2 Descriptive Account Processes/Strategies Item 17.....	269
5.7.7.3 Levels of Processing Item 17.....	270
5.7.8 Item 18.....	270

5.7.8.1 Operations/Strategies Item 18.....	271
5.7.8.2 Descriptive Account Processes/Strategies Item 18.....	271
5.7.8.3 Levels of Processing Item 18.....	272
5.7.9 Item 19.....	273
5.7.9.1 Operations/Strategies Item 19.....	273
5.7.9.2 Descriptive Account Processes/Strategies Item 19.....	273
5.7.10 Item 20.....	274
5.7.10.1 Operations/Strategies Item 20.....	275
5.7.10.2 Descriptive Account Processes/Strategies Item 20.....	275
5.7.10.3 Levels of Processing Item 19&20.....	275
5.8 Summary	276
5.9 Establishing Cognitive Validity	277
5.9.1 Expeditious Reading Operations	278
5.9.2 Levels of Processing.....	279
Chapter 6: Overview, Conclusions, Implications, and Recommendations.....	281
6.1 Introduction	281
6.2 Overview of Research Findings	282
6.2.1 Overview and Conclusions Performance in PBT and CBT	282
6.2.1.1 RQ1	282
6.2.1.2 Conclusions RQ1.....	283
6.2.2 Processes in PBT and CBT.....	284
6.2.2.1 RQ2	284
6.2.2.2 Conclusions RQ2.....	285
6.2.3 Interface Design.....	285
6.2.3.1 Scrutinizing Item 14 in PBT and CBT	285
6.2.3.2 Suitability of Computer Interface.....	287
6.3 Conclusions on Cognitive Validity CBT	288
6.4 Overall Contributions of this Study	289
6.5 Study Limitations and Future Research	292
6.6 Concluding Remarks	295
Bibliography	297

Appendix A: Study’s Reading Test	336
Appendix B: Computer Familiarity Questionnaire English Version	344
Appendix C: Computer Familiarity Questionnaire Arabic Version	347
Appendix D: Post-Test Questionnaire English Version	350
Appendix E: Post-Test Questionnaire Arabic Version	353
Appendix F: University’s Placement Test	356
Appendix G: University’s Permission Letter.....	362
Appendix H: Informed Consent English & Arabic.....	364
Appendix I: Samples of Textbooks Used in Target Context	367
Appendix J: Kobrin’s (2000) List of Strategies	370
Appendix K: Al-Amri’s (2008) Taxonomy	372
Appendix L: Strategy Counting Template	377
Appendix M: Think Aloud Protocol PBT.....	381
Appendix N: Think-Aloud Protocol CBT.....	385
Appendix O: Cohen & Upton’s (2007) Strategies	388
Appendix P: This Study’s Identified Strategies (Template).....	393

List of tables

Table 1. Grading System Universities Saudi Arabia	11
Table 2. Urquhart & Weir’s (1998) four-level Matrix of Reading Types	17
Table 3. Cognitive Processing in Reading Tests (Bax, 2013).	35
Table 4. Included Font Types in Bernard et al.’s Study	62
Table 5. Summary Reviewed Studies Addressing Typeface and Type Size	66
Table 6. Suggested On-Screen Typeface/Type Size Settings	68
Table 7. Summary of Studies Reviewed on Line Length	71
Table 8. Suggested Line Length Settings	73

Table 9. Suggested Text/Background Colour Settings	84
Table 10. Suggested Screen Size and Resolution	86
Table 11. Design Pilot Study	116
Table 12 Reading Types in relation to Cambridge ESOL Levels KET and PET.....	129
Table 13. Reliability Statistics Main Study Test	130
Table 14. Lexical Profile Reading Passage.....	140
Table 15. Readability Statistics Reading Passage 1, 2, & 3	141
Table 16. Lexical Profiles Course Book Sample Texts	142
Table 17. Readability Figures Course Book Sample Texts	142
Table 18. Question Order vs. Passage Order	143
Table 19. Timeline Data Collection Main Study	144
Table 20 Test Administration Procedure Quantitative Study	145
Table 21. Overview Division CFQ Questions	161
Table 22 Loadings of Computer Familiarity Questions on Familiarity Scale	165
Table 23. Loadings of CFQ Questions on Familiarity Scale after item omissions.....	166
Table 24. Overview Division Questions Questionnaire Two	168
Table 25. Intrajudge Coding Reliability Figures	188
Table 26. Interjudge Coding Reliability Figures	189
Table 27. Reliability Coefficient PBT Main Study	193
Table 28. Reliability Coefficient CBT Main Study	194
Table 29. Standard Error of Measurement for PBT & CBT.....	195
Table 30. Descriptive Statistics for the PBT Reading Test	196
Table 31. Descriptive Statistics for the CBT Reading Test	197

Table 32. Shapiro-Wilk’s Normality Test PBT & CBT	198
Table 33. Wilcoxon Test Results of Difference Total PBT & CBT Scores	200
Table 34. Correlations CBT and PBT Main Study.....	201
Table 35. Comparison of Item Level Performance between PBT and CBT	203
Table 36. Descriptive Summary of Participants’ Ease of Use Questionnaire Results Part 1	206
Table 37. Summary of Participants’ Ease of Use Questionnaire Results Part 2.....	208
Table 38. Overview Strategy Types and Strategy Type Tokens PBT and CBT.....	219
Table 39. Total Strategy Tokens each Participant in PBT and CBT	220
Table 40. Descriptive Statistics Strategy Category 1 PBT & CBT	222
Table 41. Descriptive Statistics Strategy Category 2 PBT & CBT	225
Table 42. Descriptive Statistics Strategy Category 3 PBT & CBT	228
Table 43. Descriptive Statistics Strategy Category 4 PBT & CBT	231
Table 44. Descriptive Statistics for Strategy Category 5 PBT & CBT.....	233
Table 45. Descriptive Statistics Strategy Category 6 PBT & CBT	235
Table 46. Descriptive Statistics for Strategy Category 7 PBT & CBT.....	237
Table 47. Descriptive Statistics for Strategy Category 8 PBT & CBT.....	238
Table 48. Descriptive Statistics for Strategy Category 9 PBT & CBT.....	240
Table 49. Descriptive Statistics for Strategy Category 10 PBT & CBT.....	241
Table 50. Descriptive Statistics for Strategy Category 11 PBT & CBT.....	242
Table 51. Descriptive Statistics of Significant Differences PBT & CBT.....	244
Table 52 Test-Takers’ Scores Think-Aloud Test PBT and CBT.....	248

List of Figures

Figure 1. Messick’s (1989) Validity Framework.....	26
Figure 2 Overview of Weir’s (2005) Socio-Cognitive Framework for Test Validity	29
Figure 3 Khalifa & Weir’s (2009) Model of Reading	33
Figure 4. Model for Expected Manifestation of Cognitive Processes in this Study.....	46
Figure 5. Reviewed Elements in Literature	52
Figure 6 Identified Human Computer related Variables in CBT by Leeson (2006)	56
Figure 7. User Interface Evaluation Model for a Computer Based Language Test.....	58
Figure 8. Worked out Interface Design Evaluation Model of Reviewed Elements	95
Figure 9. Overview Recommended Settings based on reviewed Literature through InterfaceDesign Evaluation Model	94
Figure 10. Research design.	102
Figure 11. Devised Data Collection model.....	103
Figure 12. Essential components of a CBT interface design process (Fulcher, 2003).	107
Figure 13. This study’s interface design process adapted from Fulcher (2003).	109
Figure 14. Screenshot Computer Interface Pilot Study	113
Figure 15. Screenshot Lextutor Input	138
Figure 16. Screenshot Lextutor Output.....	139
Figure 17. Screenshot Basic Configuration Hot Potatoes.....	148
Figure 18. Screenshot Other Optional Amendments	149
Figure 19. Screenshot Font Amendments.....	150
Figure 20. Screenshot Text/Background Colour options in Hot Potatoes	153
Figure 21. Screenshot Button Amendment Hot Potatoes	154

Figure 22. Screenshot Instructions Scrolling Amendments.....	155
Figure 23. Screenshot Interface Used in this Study	157
Figure 24. Score Distribution in PBT	Figure 25. Score Distribution in CBT
.....	198
Figure 26. Boxplot Median CBT and PBT Scores	199

List of Abbreviations

CALT: Computer-Assisted language testing

CAT: Computer-Adaptive testing

CBT: Computer-based testing

CFQ: Computer Familiarity Questionnaire

CPL: Characters per Line

CRT: Cathode Ray Tube

DPI: Dots per Inch

EFL: English as a foreign language

ESL: English as a second language

HCI: Human Computer Interaction

L1: Native language or mother tongue

L2: Second or foreign language

MCQ: Multiple-choice question

PBT: Paper-based testing

PTQ: Post-Test Questionnaire

PYP: Preparatory Year Program

SA: Saudi Arabia

SAQ: Short Answer Question

SCS: Screen Capture Software

TA: Think-Aloud

TTS: Test-taking strategies

Chapter 1: Introduction

1.1 Background to the Study

This study aims to contribute to the field of language testing by investigating the effect of computer interface design on performance and cognitive processing of Saudi Arabian male first-year preparatory students when taking an English L2 reading test. Literature from educational psychology (e.g. Pollock and Sullivan, 1990; Ployhart, Weekley, Holtz and Kemp, 2003), ergonomics (e.g. Noyes and Garland, 2003; 2008), computers and human factors (e.g. Dillon 1992; Haas, 1992; Lee and Tedder, 2003), education (e.g. Azevedo and Bernard, 1995) and language testing (e.g. Pomplun et al., 2002; Choi et al., 2003; Pommerich, 2004, 2007; Higgins et al., 2005; Horkay et al., 2006; Puhan et al., 2007) demonstrates that the comparability of both paper-based tests (henceforth PBT) and computer-based tests (henceforth CBT) of reading has been under investigation for decades. Several comparability studies have been carried out of which some concluded that the newly introduced CBT's in their studies were equivalent to their paper-based counterparts whereas others found significant performance differences between the two modes. The conclusions drawn in a number of these comparability studies showed that either specific elements of the computer interface or the testing mode itself as a whole were thought to have contributed to observed differences between PBT and CBT performance. For example, Pomplun et al (2002) introduced a CBT version of the Nelson-Denny Reading Test to a sample population of 215 high school and college students in a between group study design to investigate its equivalence to its paper-based counterpart. They found a significant difference in the vocabulary section of the test, which they attributed to the difference in presentation of stimuli between the two modes.

Chapter 1: Introduction

Choi et al. (2003) investigated the comparability of the PBT and CBT versions of the TEPS, a proficiency test (covering the four skills) developed by Seoul National University, which involved 258 Korean EFL university students. Their aim was to evaluate the construct validity and content validity of the proficiency test used through comparing performance in both modes. The results of their reading subsection, which is of particular relevance to this study, showed a significant mode effect indicating a relative difficulty on the computer-based version in particular. One of the possible explanations Choi et al. (2003) gave for this result was the fact that their test-takers might not have been familiar with the new testing mode, i.e. testing mode (un) familiarity, which has been indicated as a possible construct irrelevant measure in several earlier studies (e.g. de Beer & Visser, 1998; Sawaki, 2001; e.g. Horkay et al., 2006). Choi et al. (2003) further mentioned that a large number of the students assessed in their study reported some form of ‘eye-fatigue’, which could have been a possible cause for the difficulty in their CBT, which Choi (2000) and others’ also indicated in earlier studies where students had to read longer passages on screen (e.g. Larson, 1999; Sawaki, 2001; Blackhurst, 2005).

Pommerich (2004) investigated the effect of text and item presentation, and computer interface features on test performance of 1893 grade 11 & grade 12 students to whom she administered a self-designed scientific reasoning test that included a mathematics component and a passage based reading component. The reading component consisted of four reading passages each accompanied by 15 test items. The question format used was MCQ and the test events were divided into two sessions. After the first testing session, amendments were made in response to differences that were found at the item level due to presentation elements of the interface. The second data collection session was conducted and showed improvement of the CBT compared to

Chapter 1: Introduction

the first session. Pommerich (2004) suggested to make further amendments in follow up studies in order to further improve CBT item performance but did not specify how to accomplish this.

Higgins et al (2005) investigated the effect of scrolling on test-taker reading performance. They examined a total of 219 fourth grade students on a paper-based test, a computer-based version of the PBT requiring scrolling as a navigation tool, and a CBT version where page turning was required to navigate through the test (page-turning requires the test-taker to click on a button that will then turn the page for him). Higgins et al. (2005) did not find any significant performance differences between the three modes; however, they did mention that their sample population consisted of unusually high computer familiar students, which could have been the reason for this.

Puhan et al. (2007) investigated 1122 reading examinees on a large-scale certification test from the Praxis program, which involved mathematics, a writing, and a reading component. The reading test consisted of 40 MCQ test items administered to test-takers in both PBT and CBT mode. Results showed neither significant difference between the two modes at the test-level nor at the item level for the reading component but DIF analyses showed that three writing items functioned different between the two modes.

Although the comparability studies above yielded inconclusive results as to the effect elements of the computer interface had on test performance, it is evident that it holds a significant position in PBT and CBT equivalence studies as a potential source of construct irrelevant variance. Choi et al (2003) mentioned about interface design in relation to test performance that it is one of ‘the major factors which significantly affect test performance that have been investigated in previous comparability studies’ (p.297). Other language testing researchers such as Fulcher (2003) agreed, as he stressed that an inadequately developed

interface ‘may easily become a source of construct irrelevant variance, thus threatening the score users’ ability to make meaningful inferences from test scores’ (p.385).

As for the effect of interface on cognitive behaviour, Pommerich (2004) added based on her investigation of the influence of interface design on test, passage, an item presentation that, ‘differences in how the test is presented could influence examinee behaviour while testing’ (p. 40), which could in turn affect examinee performance. This study’s overall aim and associated objectives were based on these indications above, i.e. the potential effects of computer interface elements on test-takers’ behaviour and performance, which is further detailed in the following section.

1.2 Aims of the Study

This study aims to contribute to the field of language testing by investigating the effect of computer interface design on performance and cognitive processing of Saudi Arabian male first-year preparatory students when taking an English L2 reading test. An interface was developed aiming to minimize interference with the test’s constructs, which is to be validated through evidencing comparability between the PBT mode and the CBT mode of this study’s reading test. Conditions that determine the comparability/equivalence of a computer-based test and a paper-based test mentioned by the American Psychological Association in their guidelines are as follows: ‘Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode’ (APA, 1986, p. 18).

Many of the comparability studies that were carried out, particularly from the 80s up to the beginning of the 21st century reflected these guidelines by focusing mainly on post-hoc, quantitative analyses involving mean comparisons and correlational analyses (among others) to support mode equivalence (e.g. Boo, 1997; Russell and Haney, 1997; Russell, 1999; Pomplun et al., 2002; Choi et al., 2003). One of the limitations of the abovementioned guidelines pertaining mode equivalence, however, is that this focus on post-hoc, quantitative analyses, fails to address the process itself, which is equally an essential element of equivalence establishment (Messick, 1989; Weir, 2005; Khalifa & Weir, 2009; Field, 2012). The Standards for Educational and Psychological Testing later acknowledged its importance as they mentioned in standard 1.8: ‘If the rationale for test use or score interpretation depends upon the premises about psychological processes or cognitive operations used by examinees, then theoretical or empirical evidence in support of those premises should be provided’ (p.19).

This study involves both aspects as it investigates test-takers’ cognitive processes in addition to the conventional post-hoc score comparisons between a PBT and CBT of reading resulting in the following two levels at which to establish equivalence:

1. The performance level, which entails equivalence of test scores, score distributions, significant correlations, and item-level equivalence in PBT and CBT.
2. The cognitive level, which entails the equivalence of test-takers’ cognitive processes/strategies between the two modes.

In addition to investigating cognitive equivalence, the data obtained from the cognitive processes would further provide a qualitative insight into the processes test-takers employ when answering test items. Providing both performance (i.e. product) and cognitive (i.e. process) related evidence in support of mode equivalence (McDonald, 2002) is expected to not only

confirm the suitability of the CBT itself, but likewise the appropriateness of the interface that was developed for this study's purpose. The following section further elaborates on this by presenting this study's anticipated contributions to the field of L2 reading and language testing.

1.3 Significance of the Topic

This study aims to contribute to a number of areas that have not been explored greatly in the existing language testing literature, which are described below.

1. Computer Interface of an L2 Reading Test (CBT). Many comparability studies have indicated possible negative effects of elements of a computer interface on language test performance (e.g. Lee et al., 1986; Pomplun et al., 2002; Choi et al., 2003). Although each of these studies highlighted a particular element of the interface as a possible source of discrepancies between CBT and PBT (e.g. item presentation, scrolling, screen resolution), none have approached the interface design from a combined elements perspective, i.e. reviewed all involved elements grouped together and developed an interface based on that to (a-priori) try to minimize possible construct irrelevant variance emerging from the computer interface. The contribution of this study to the field of language testing is by addressing this gap through the development of an interface based on a thorough review of the published literature aiming to limit possible construct irrelevant variance in the CBT-version of the reading test by making this interface 'invisible' as it were to the test-taker (Nielsen, 1990) leading up to process and product equivalence. This 'invisible' interface is formed through a by the researcher established interface design evaluation model incorporating the various interface design elements applicable to human computer interaction and language testing (see section 2.7).

2. Comparability Studies in L2 Reading (Performance). Chalhoub-Deville & Deville (1999) and later Sawaki (2001) pointed out that studies investigating comparability of PBT and CBT language test scores were still limited and therefore encouraged further contributions to the field in due course. However, only a small number of studies have done this since then and, the studies that did, likewise suggested that more comparability research investigating the effect of CBT on test-taker performance is needed (e.g. Choi et al., 2003; Pommerich, 2004; Higgins et al., 2005). Because part of this study involved score (i.e. product) comparisons (i.e. RQ1), it aims to contribute to this particular aspect of the L2 language testing literature.

3. Comparability Studies in L2 Reading (Processes). Particularly in L2 language testing research, only a few studies investigated processes equivalence between PBT and CBT involving an L2 reading test. In fact, only Kobrin (2000) and later Al-Amri (2008) compared cognitive processes in L2 reading tests in both modes. Studies such as Cohen & Upton's (2007) explored reading and test-taking processes used by test-takers when performing the reading tasks of the new TOEFL test; however, they did not compare these between two testing modes (i.e. PBT and CBT). This study aims to further contribute to the relatively underrepresented comparability research in this area by comparing test-takers' cognitive processes in the two modes.

4. L2 Expeditious Reading Processes in PBT and CBT. With regards to L2 reading, a gap in the current L2 reading literature exists in relation to text processing in L2; particularly, the way in which test-takers perform expeditious reading operations according to the set purpose (e.g. answering a test item) has been left underrepresented. Urquhart & Weir (1998) mentioned the following about this, 'We have theories of careful reading but very little on how readers process texts quickly and selectively, i.e. expeditiously, to extract important information in line with intended purpose(s)' (Urquhart & Weir, 1998, p.101). This entails processes such as skimming,

scanning, and search reading in L1 as well as in L2 that aid in locating required information in a given text. As Urquhart & Weir (1998) further mentioned, ‘We have somewhat ignored expeditious reading behaviors such as skimming, search reading and scanning in both L1 and L2 teaching of reading’ (ibid, p.101).

This study’s aim is to contribute to the field of L2 reading and language-testing by exploring test-taker’s cognitive behaviour when taking an L2 reading test in PBT and CBT mode on items requiring expeditious reading operations to locate relevant information in addition to careful reading operations by qualitatively describing these processes (see 2.2.1 for further discussion of these reading types).

5. Cognitive Validity of a Computer-Based L2 Reading Test. A test is considered to be cognitively valid when the processes elicited by that particular test emulate (as much as possible) cognitive processing used in that particular situation in real life (e.g. Glaser, 1991; Khalifa & Weir, 2009; Field, 2012). Through comparing the cognitive processes in PBT and CBT in this study’s reading test, it is anticipated that evidence is generated for cognitive processes equivalence between the two modes, which is the first step towards investigating this study’s test’s cognitive validity. Qualitatively describing the processes and identifying whether the appropriate cognitive processes are employed by the test-takers is expected to provide further supporting evidence for the cognitive validity of the L2 reading test used in this study (further discussed in section 2.4), which, in turn contributes towards providing evidence for the test’s construct validity.

6. Contribution to Target Context. As far as the target context is concerned, the majority of studies that investigated the effect of CBT on test-taker performance were largely conducted in countries where either the first language was English or the participants were ESL students. Choi

et al (2003) and Al-Amri (2008) are among the few that have targeted EFL students in a comparability study, and, for this reason, this study is expected to contribute in this aspect. Furthermore it would be the second study to the researcher's knowledge that included Saudi Arabian EFL students for this purpose, which further indicates its novelty. A summarized preliminary overview of the educational system in Saudi Arabia and its examination system is therefore given in section 1.4 for further contextualization purposes.

1.4 Educational System in Target Context

Overall, the educational system in the target context (i.e. Saudi Arabia) is somewhat similar to other educational systems worldwide. Pre-primary education is the initial stage for children aged 3-5, which is kindergarten in western terminology. The pre-primary stage is not compulsory in Saudi Arabia and therefore not a requirement for beginning primary education. Primary education starts at the age of six and is compulsory for all nationals. At the end of grade 6, , students have an exam that they must pass in order to be admitted to the next level of education, which is named intermediate school. The total duration of intermediate school is three years (from ages 11-14), which then would be equivalent to grade 9, as grade 6 is the final grade in primary school when students are about 11 years of age. From there secondary school commences, which lasts another three years (from 14-18 years of age) totaling 12 years (pre-primary school included). The duration of tertiary education depends on the field of study. Humanities and social sciences take four years to complete whereas medicine and engineering both take five years.

The Saudi ministry of education and ministry of higher education dictate the public school curriculum for primary, intermediate, and secondary education. This is also the case at

tertiary level where it is the ministry of higher education that sets the curriculum for public universities. The private schools and universities that exist, however, generally closely approximate the curricula set out by both ministries and in some cases private establishments follow the exact same curriculum. For this reason, it can be reasonably confidently argued that the test-takers involved in this study have gone through the same educational system and have been assessed in the same way throughout the four levels of education. The following section discusses the examination system in more detail.

1.5 Examination in Saudi Arabia

In tertiary education, under which this study's sample comes, the ministry of higher education sets out the academic policies, administrative structures, as well as assessment implementations. The examination policies are identical throughout, where examination and marking is carried out by course instructors and coordinators. Midterm and final exams are double-marked to assure overall consistency within and between institutions. Generally, the exams are achievement tests, i.e. students are assessed on what they've been taught during the course throughout the semester. For the English language programs in the preparatory year these involve reading, writing, listening, and speaking, which together make up an overall score of 100% for English. There may be slight divergence between institutions on the assigned weight to each skill; however, these differences are not greatly significant as the same overall marking scheme is used by most of the universities in Saudi Arabia. The marks for each course are made up of quizzes, midterm(s), and continuous assessment, which add to around 40% of the total mark whereas around 60% of the total mark is from the final examination. Table 1 below summarizes the grade letters, percentages, and associated points on the GPA-scale.

Table 1. Grading System Universities Saudi Arabia

Percentage (%)	Grade Letter	GPA (5)	GPA (4)
95-100	A+	5	4
90-94	A	4.75	3.75
85-89	B+	4.5	3.5
80-84	B	4	3
75-79	C+	3.5	2.5
70-74	C	3	2
65-69	D+	2.5	1.5
60-64	D	2	1
0-59	F (Fail)	1	0

As table 1 above shows, the pass/fail cut-off point in tertiary education in Saudi Arabia is 60%, which corresponds with a D-grade. Anything below the cut-off point of 60% corresponds with F (i.e. fail). The assessment is based on achievement, as is the case for primary, intermediate, and secondary education in Saudi Arabia. Likewise, the reading test used in this study is an achievement test where mainly local level text processing is assessed.

1.6 Research Questions

The formulated research questions and hypotheses based on previous indications on the possible effect of interface design on test-takers' performance and test-taking behaviour are as follows:

Related to test-takers' performance

RQ1. *What is the effect of administration mode on students' performance when taking a lower-level L2 reading test?*

H₀: *There is no effect of administration mode on students' performance when taking a lower-level L2 reading test (PBT=CBT).*

Related to test-takers' processes

RQ2. *What is the effect of administration mode on test-taking strategies students employ when completing a lower-level L2 reading test?*

H₀: *There is no effect of administration mode on test-taking strategies students employ when completing a lower-level L2 reading test (PBT=CBT).*

1.7 Thesis Overview

This thesis contains a total of six chapters. This chapter briefly outlined the background to this study followed by mentioning its aims, its importance/contribution(s) to the field of language testing, and gave a brief insight into the target context with regards to education and examination.

Chapter two of this thesis reviews the relevant literature briefly introducing the reading concept, followed by reviewing the reading types and cognitive models assumed in this study. It further discusses the contemporary view of validity and how this study works within the socio-

Chapter 1: Introduction

cognitive framework of language test validity, which is a product of the contemporary view of validity. The final part, comprising more than half of the literature review in word count, is a comprehensive review of the interface elements relevant to computer-based language testing leading up to a worked out model encompassing the optimal settings for a computer-interface that can be used as a template in the field of reading and language testing.

Chapter three discusses the methodology employed in this study in terms of research design and the methods/instrumentation chosen to collect quantitative and qualitative data required to investigate research questions one and two. Two pilot studies that were carried out are described in terms of sample description, instruments used, procedures followed, post-hoc analyses and implications leading up to the main study, which is subsequently described in the same manner.

Chapter four presents the analyses and results for research question one (RQ1) and subsequently discusses these findings.

Chapter five is subdivided into two parts (i.e. part A and part B). In part A, the results i.e. comparison of cognitive processes between PBT and CB to answer RQ2 are presented and subsequently discussed. Part 2B involves a qualitative description of test-taker behaviour when answering the test's items illustrated through excerpts from the think-aloud in order to provide evidence in support of this study's test's cognitive validity.

Chapter six concludes the thesis by summarizing the main points and concluding the findings in relation RQ1, RQ2, and the conclusions regarding the test's cognitive validity. It further outlines this study's limitations, and recommendations are made for further research based on this study's outcomes

Chapter 2: Literature Review

2.1 Introduction

This chapter reviews and discusses the relevant literature followed by the gaps identified to which this study contributes in terms of L2 comparability studies in L2 reading, its interface design, expeditious reading processes involved when reading in L2, the test's cognitive validity, and its overall construct validity. The review begins with a brief overview of reading ability, reading types, and reading processes in L2 reading in order to provide a background and to substantiate the assumptions about reading made in this study. After that, test validity is introduced leading up to Weir's (2005) socio-cognitive framework of test validity highlighting cognitive validity in particular due to its relevance to this study. Following this, establishing cognitive validity of a language test from a cognitive processes perspective is reviewed followed by a brief review of strategies found in related studies leading up to a two-stage process/strategy model through which test-takers' cognitive processes are expected to occur when taking this study's reading test in PBT and CBT mode. After that, the short-answer question assessment format chosen for this study's test is reviewed (i.e. SAQ).

The final part of this literature review comprehensively discusses computer interface design and its elements in relation to computerized reading assessment and its effect on test-takers' behaviour and performance leading up to an interface design model illustrating the optimal settings for each element involved based on this review.

2.2 L2 Reading Ability

Earlier definitions of reading were mostly limited to describing its concept in the narrowest sense, which was related to the view that the reading process consisted mainly of decoding (e.g. Perfetti, 1985). Later research opposed this view based on empirical evidence that decoding itself did not necessitate understanding word meaning, i.e., comprehension (e.g. Urquhart & Weir, 1998).

Defining reading ability in a single sentence has been argued to be extremely difficult, if not, impossible, as there are several processes involved which act in combination with skills, strategies, and knowledge bases to achieve this (Grabe & Stoller, 2011). L2 research on reading, which evolved mainly from L1 research, resonates this view, as Kim (2009) remarks: ‘The nature of second language (L2) reading ability is extremely complex and its components are yet to be agreed upon’ (Kim, 2009: p.1). Despite its complexity, the generally accepted view of the reading process in itself is that it is a cognitive activity as it largely takes place in the mind (e.g. Just & Carpenter, 1987; Bernhardt, 1991; Urquhart & Weir, 1998; Kim & Huynh, 2008; Cohen & Upton, 2007; Shiotsu, 2010). Earlier studies by cognitive psychologists and reading theorists indicated this through their involvement in examining cognitive processing throughout the 20th century since its inception in the early 1900s (i.e. Huey, 1908) to the 1960s (e.g. Goodman, 1967), 1970s (Gough, 1972), and 1980s (e.g. Stanovich, 1980).

Khalifa & Weir (2009) argue that, as far as the testing of reading ability is concerned; the cognitive view of reading, reflected through a cognitive processes perspective, provides the most adequate and workable theoretical foundation for this purpose, for which supporting evidence will accumulate as this review develops. This study assumes a cognitive processes perspective when investigating test-takers’ cognitive behaviour while taking an L2 reading test and will

therefore be reflected in the discussions that follow in relation to cognitive processing in L2 reading assessment and further throughout this study.

2.3 Reading Types

In the past, language researchers and cognitive psychologists alike described reading as a slow, incremental process mainly reflecting careful reading, which meant understanding each and every text element (Rayner and Pollatsek, 1989). Urquhart & Weir (1998) indicated that, because of this focus on mainly careful reading (particularly at the local level), other aspects of the reading process were left neglected. They mentioned that, for example, expeditious reading operations like skimming, scanning, and search reading, which aid in quick and selective text processing, had been largely ignored in L1 and L2 reading research. Urquhart & Weir (1998) view reading as a multicomponential construct, which is reflected through their proposed four-level componential matrix from a reading components perspective based on Weir's (1993) and Pugh's (1978) earlier conceptualization of the reading process. The matrix included four reading types: global expeditious reading, local expeditious reading, global careful reading, and local careful reading, with each reading type contingent to the purpose/goal of the reader. This multicomponential view guided by its purpose has found support by other researchers as integral parts of the reading process (e.g. Perfetti, 1997; Enright et al., 2000; Grabe, 2002; Grabe and Stoller, 2011). Table 2 below further illustrates the four reading types proposed by Urquhart & Weir (1998) and each reading type is subsequently explained.

Table 2. Urquhart & Weir's (1998) four-level Matrix of Reading Types

	Global	Local
Expeditious	A. Skimming quickly to establish discourse topic and main ideas. Search reading to locate quickly and understand information relevant to predetermined needs.	B. Scanning to locate specific information; symbol or group of symbols; names, dates, figures or words (<i>also includes search reading at local level</i>).
Careful	C. Reading carefully to establish accurate comprehension of the explicitly stated main ideas the author wishes to convey; propositional inferencing.	D. Understanding syntactic structure of sentence and clause. Understanding lexical and/or grammatical cohesion. Understanding lexis/deducing meaning of lexical items from morphology and context.

As table 2 shows, global expeditious reading (level A) in the top left corner is a way of quickly reading a text by being selective, which results in optimal efficiency. In order to achieve this, strategies such as skimming, and search reading are employed to aid in understanding of a text at the global level. Urquhart & Weir (1998) refer to this global processing as macro-structure whereas scanning (and also local level search reading), as seen in expeditious reading level B in the top right corner, aids to processing text at the local level (i.e. micro-structure). As shown in reading level C, careful reading at the global (macro) level aims to develop an understanding of the entire text and helps to make inferences whereas careful reading at the local (micro) level (i.e. reading level D) aims to understand word meaning and clause or sentence structure understanding.

What can be implied from this framework based on the four reading types operating at the micro and macro levels is that:

1. Reading involves multidivisible skills that a reader applies depending on the (reading) goal he sets.
2. Reading processes appear to be at least bi-divisible into lower-level processes and higher-level processes operating at the local and global level (e.g. Just & Carpenter, 1980; Grabe and Stoller, 2002; Cohen and Upton, 2006).

This means that the view of reading as a construct would indicate multi-componentiality in terms of reading itself as well as in the assessment of it due to this division into reading types contingent to its purpose, which is supported by other researchers such as Grabe (2002). This framework, which distinguishes between reading types and levels of processing, has become a common framework of reference in the reading literature (e.g. Field, 2000; Ridgway, 2003; Walter, 2003; Meng, 2009; Akyel and Ozek, 2010).

Urquhart & Weir's (1998) reading components perspective reflected through the four reading types mentioned above is assumed in this study in its investigation of test-takers cognitive processes when taking an L2 reading test. The test items included in this study's reading test are anticipated to elicit local level expeditious reading operations in order to locate relevant information in the test's passage followed by more careful reading operations to achieve accurate comprehension to ensure correctly answering the test items (i.e. reading type B and D in the framework). The matrix forms an integral part of the later developed multidivisible model of reading by Khalifa & Weir (2009), which further outlines the processing levels in reading according to their set purpose/goal and is discussed in section 2.4 of this chapter. This inclusion of the expeditious reading dimension in relation to careful reading, investigating and illustrating

these reading operations combined with their associated levels of processing, is expected to significantly contribute to the scarcely available literature in this area.

2.4 Processes Models in Reading

The processes involved when reading a text have been described in different ways each through its own model placing emphasis on the active part the reader has in this process. For example, one of the earlier process models was the Reader Response Theory proposed by Rosenblatt (1938) who argued that readers interact with the text bringing in their own perceptions and background knowledge.

Another often mentioned theory is the Metacognitive Theory proposed by Baker & Brown (1984), which emphasized a metacognitive approach through on-going comprehension monitoring during reading and using strategies to compensate for lack of comprehension when encountered. This theory contributed to understanding processes involved when monitoring one's understanding of the text read and applied compensatory processes related to comprehension shortcomings.

The Schema Theory of Anderson & Pearson (1984) described the reading process as a reader creating his own schemata based on concepts he discovers within the text. This theory contributed, for example, to understanding the processes involved when integrating prior knowledge into newly acquired information from the text. The schemata referred to is in reality prior knowledge activation, which allows for making inferences about text content based on the created schemata.

Other models of the reading process have provided insights into main idea extraction (e.g. Afflerbach, 1990), global processes such as integrating different parts of a passage/text to

enrich understanding (e.g. Kintsch & van Dijk, 1978; Pressley & Afflerbach, 1995), inferencing (e.g. Trabasso & Magliano, 1996), and text processing at the local level, i.e. word recognition (e.g. Gough, 1972; LaBerge and Samuels, 1974).

Just & Carpenter's (1980) critique of a number of these previous models at the time was that they only described one certain level of processing by illustrating the mechanisms for only one aspect of the reading process (i.e. either word recognition *or* inferencing *or* main idea extraction etc.) but not together as a whole thereby neglecting various other stages in the reading process. In response to these identified shortcomings in describing the various levels of processing in reading, Just & Carpenter (1980) proposed one of the earlier models that described overall reading comprehension attempting to include various levels of processing from encoding at the word level up to higher-level processes such as text integration. This 'all-inclusive' model describing these processes gained prominence in the late 20th century. Examples of these processing stages were 'encoding, lexical access, assigning semantic role,' referring to lower-level processes, and 'relating the information in a given sentence to previous sentences and previous knowledge' referring to higher-level processes (Just & Carpenter, 1980, p.331). Urquhart & Weir's (1998) reading matrix, as mentioned earlier, too emphasized a distinction between lower-level and higher-level text processes inaugurating at the local level and transiting into more global processes depending on the task at hand.

Khalifa & Weir (2009) expanded on Just & Carpenter's (1980) earlier work by devising a cognitive model of reading describing cognitive processes likely to occur when reading a text further exemplifying the role which the previously discussed reading types by Urquhart & Weir (1998) have in relation to these processes. Their model formed this study's theoretical framework for investigating test-takers' cognitive behaviour when taking an L2 reading test and further

functioned as a framework of reference for establishing this study's test cognitive validity, which will be discussed in more detail in section 2.4.

2.5 Process Levels in Reading

Based on the indications in section 2.4 above, the processes in reading can be divided into two levels: lower-level processes and higher-level processes. Lower-level processes involve text processes at the local level such as lexical access, syntactic parsing, and establishing propositional meaning (at clause and/or sentence level). Higher-level processes refer to more global text processing such as global level inferencing, building a mental model, text level and intertextual representation (Khalifa & Weir, 2009). Descriptions of the lower-level processes are given in section 2.5.1 due to their relevance to this study's reading test, as this study's test items are thought to mainly tap into these.

2.5.1 Lower-Level Processes

2.5.1.1 Lexical Access

In this process the reader develops an orthographic (i.e. word recognition) and/or phonological representation of the lexical item (i.e. word/phrase) he encounters. In simpler terms this would mean that when the reader recognizes a word, it is matched with stored information of that word's form and meaning through retrieving it from the lexicon (Field, 2004). The lexicon is accessed through two routes:

1. Visual input into word meaning excluding sound (orthographic route/direct route)
2. Visual input into sound into word meaning

Regardless of the route taken, this process is largely automatic (i.e. beyond conscious

control) for the L1 reader and the skilled L2 reader. However, it could pose significant difficulties for the lower-level L2 reader through, for example, wrongly assigning stored information to the (visually) retrieved word. Examples of this are shown and further discussed in section 5.5 of the second results and discussion chapter (chapter five).

2.5.1.2 Syntactic Parsing

Khalifa & Weir (2009) view syntactic parsing the same as grammatical knowledge, which therefore, in addition to word order, includes word form, and structural elements in a clause/sentence such as prepositions, helping verbs/auxiliary verbs etc. Urquhart & Weir (1998) mention about the process of syntactic parsing that it aims to establish relationships between the word that is recognized and its grammatical connotation in order to store it in working memory. As for the successful readers or good readers, Grabe and Stoller (2002) mention that by having a clear understanding of words and their position and function within a clause/sentence, it enables them to disambiguate different words easily like, for example, words that have context related meanings.

2.5.1.3 Establishing Propositional Meaning

Grabe and Stoller (2002) describe the process of establishing propositional meaning as the reader constructing meaning at the clause-level from words that contain structural information. In other words, it is an abstract understanding of a unit of meaning without external factors brought in by the reader such as background knowledge or contextual factors that might contribute to enriching the propositional meaning. In short, it is like a literal interpretation of the unit of information that is attended to.

2.5.1.4 Inferencing at the Local Level

Local inferencing is effectively inferencing that takes place at the word level. Local inferencing mainly comprises the following two types:

1. As a response to a word being ambiguous as to what its meaning could be in the target context (i.e. the sentence). This type of inferencing involves guessing word meaning of a word in the context that is unknown.
2. Anaphoric inferencing is another type of local inferencing where the reader has to identify to which entity, for example, a pronoun refers (i.e. to which preceding entity). This type of inferencing is required in a number of items in this study's test of which illustrative examples are given in chapter 5 (i.e. section 5.7.9.2).

The lower-level processes described are anticipated to be elicited by the test's items in this study, for which the answers were explicitly stated in the text and were locatable through employing expeditious reading operations followed by careful reading operations. Providing evidence for appropriate reading operations and process levels for this study's test items in PBT and CBT would contribute to the test's cognitive validity and subsequently, construct validity, which are core elements in Weir's (2005) socio-cognitive framework for test validity. Before reviewing Weir's (2005) validity framework, the following section firstly introduces the contemporary view of validity, its concepts, and types, leading up to this framework and its cognitive validity element, for which this study aims to provide evidence through its reading test.

2.6 Test Validity

Before discussing validity and its concepts, it has to be mentioned that the aim of this study was neither to validate a language test in its entirety, nor was it to provide a comprehensive

discussion on validity, its concepts, and its variations, as this would require considerable additional space, which would be beyond the present study's scope due to its focus. Rather, an account of the main views and generally accepted interpretations of validity is given in order to provide a context that justifies using the contemporary view of validity as a theoretical assumption when examining the effect of the interface design of a computer-based test on test takers' cognitive processes and score outcomes in this study.

2.6.1 Background

The traditional interpretation of test validity suggests that it is established essentially by finding evidence that shows that the test used is measuring what it was intended to measure from the outset (e.g. Cattell, 1946; Hughes, 1989; Brown, 1996). Messick (1995) described validity as, 'an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment' (p. 741).

Messick further stressed that validity is a matter of degree, and validation is a process, which is ongoing (e.g. Messick, 1995; Chapelle, 2001; 2003). Bachman (2004) mentioned that the degree of validity depends on the strength of the evidence one provides *in support* of the validity argument, which means that it is not possible for test developers to directly *prove* that interpretations in themselves are valid. Rather, at best they can, 'provide evidence that the intended interpretations and uses are *more plausible* than other interpretations that might be offered (italics added by author)' (ibid: p. 260).

Even when validity evidence for a test is substantial, it is never automatically generalizable to other tests, but rather 'specific to a particular use or interpretation' (Linn &

Gronlund, 2000, cited in Bachman, 2004, p.259) suggesting that validity is never absolute as it is dependent upon the evidence supporting its argument in the context in which it was established, and its underlying purpose. With this in mind, the validity concept is reviewed in this section from its initial interpretations to the more current views held by contemporary language testing researchers in the field.

2.6.2 Concepts of Validity

The initial concept of validity consisted of three main types, namely *criterion oriented validity* (which consists of predictive validity and concurrent validity), *content validity*, and *construct validity*, which was reflected in several earlier works (e.g. Lado, 1961; Davies, 1977) of which Cronbach & Meehl's (1955) (who introduced it) is arguably the most well-known. Criterion oriented validity is concerned with the relationship between a criterion (e.g. test-takers' ability measures) and test results by effectively predicting the criterion of a construct. When this relationship involves a future criterion, it is called *predictive validity* whereas simultaneously observed measures of a criterion with test scores are called *concurrent validity* (e.g. Fulcher & Davidson; 2007; Al-Amri, 2008). *Content Validity* is evidence for the fact that the content of a test is reflecting the underlying skill(s) accurately (e.g. Hughes, 2003). *Construct Validity* involves the evidence that the operationalized test is measuring the construct that had been theorized it to be measuring beforehand. At the time, content validity and criterion-oriented validity were considered the two main pillars that were heavily relied upon in language testing until the late 70s whereas construct validity was considered an alternative when the former two proved to be insufficient (Kane, 2001). It was only after that; due to the growing dissatisfaction with the limitations of existing validity types that construct validity became more prevalent.

Messick (1989) later expanded the conventional view of validity by unifying the aforementioned validity types making *construct validity* the unifying concept and added the aspect of social consequences of measurement outcomes as shown in figure 1 below.

	<i>Test Interpretation</i>	<i>Test Use</i>
<i>Evidential Basis</i>	Construct Validity	Construct Validity + Relevance and Utility
<i>Consequential Basis</i>	Value Implications	Social Consequences

Figure 1. Messick's (1989) Validity Framework

As shown in figure 1, Messick's (1989) four aspects of the validation process are integrated in a matrix-type model, which is based on its function (i.e. how the test is interpreted and used), and the justification for its validity (i.e. based on its evidence(s) and consequence). The construct validity of the test here is the evidential basis for test (score) interpretation. In order to use these scores and make decisions about them, the need for evidence of their relevance and utility becomes apparent (i.e. the justification for using the construct validity evidence in a particular context from a particular sample to support the inferences made). Making judgments about the value implications (on a social level) in relation to the construct, its underlying theory, and its measurement, is part of the consequential basis in relation to test interpretation. Social consequences are the consequences society could experience based on using a particular measure. This wider framework of test validity in its unified form proved to be of great importance to the field of language testing as well as to the field of psychology and was later included in the Standards for Educational and Psychological Testing (AERA, APA, NCME,

1999) in its chapter on validity.

However, several theorists have criticized the framework, particularly on its social consequential aspect (e.g. Wiley, 1991; Shepard 1993; 1997; Maguire et al., 1994) questioning the necessity to examine social consequences as part of validity potentially leading to confusion by muddling the validity concept (Popham, 1997). Others argued that Messick (1989) did not fully exploit its social dimension despite the importance it portrayed to have in test validation (e.g. Roeber & McNamara, 2006). In case of Popham's (1997) opposition to including social consequences, Hubley & Zumbo (2011) argued that this was largely based on the misconception that it involved test use, and test misuse in particular. However, this was not what was meant by Messick (1989), as his focus was clearly on consequences alone, and, although deemed an important segment in language assessment, test use was not to be regarded as part of the validation process, which he clarified later in order to address this misunderstanding (cf. Messick, 1998).

Despite this criticism, Messick's (1989) expansion of the validity concept has found support among a large body of language testing researchers, who shared the unitary view of validity with construct validity being the all-embracing unifying form over other validity forms (e.g. Anastasi, 1988; Weir, 1988; 2005; Bachman 1990; Bachman & Palmer, 1996; Khalifa & Weir, 2009). Bachman & Palmer (1996) devised a model to pragmatically examine weaknesses and strengths of a language test, which they called the *test usefulness model*. The model comprised of six segments, including construct validity as interpreted by Messick (1989) where social consequences were referred to as *test impact* and were separately addressed as an independent variable. Weir (2005) and later Khalifa & Weir (2009) integrated scoring validity (which includes reliability), context validity, and cognitive validity to make up 'what is

frequently referred to as *construct validity*' (italics in original) (Khalifa & Weir, 2009 p. 143) building on Messick's (1989) initiated framework.

This study assumed the all-inclusive concept of validity referring to Weir's (2005) theoretical validity framework when examining the effect of computer interface design on test-takers' performance and the cognitive processes they utilize when taking a L2 reading test in PBT and CBT. The outcomes were theorized to demonstrate the appropriateness of the interface for this study's tests and context through establishing performance (i.e. test-scores) and processes (i.e. cognitive processes) equivalence, which would subsequently provide a platform for investigating its cognitive validity. The following section further reviews Weir's (2005) framework and this study's relation to the aspect of the cognitive validity element of this framework.

2.7 Theoretical Framework for Conceptualizing Test Validity

Weir's (2005) framework conceptualizes the process of validating a language test. This framework applies (in adapted form) to each of the four skills (i.e. as listening, speaking, reading, and writing). The reason why it is called a *socio-cognitive* framework is because task performance is treated here as a social experience. The framework is *cognitive* because it involves mental processing of the test-taker when testing the ability in question (i.e. in this case reading). A summarized version of the framework including its main characteristics is shown in figure 2 below briefly describing its consisting elements and its relevance to this study. The framework is in agreement with and expands from the contemporary unified view of validity as proposed by Messick (1998). It starts by involving test-taker characteristics and from there includes cognitive validity, context validity, scoring validity, consequential validity, and criterion

related validity.

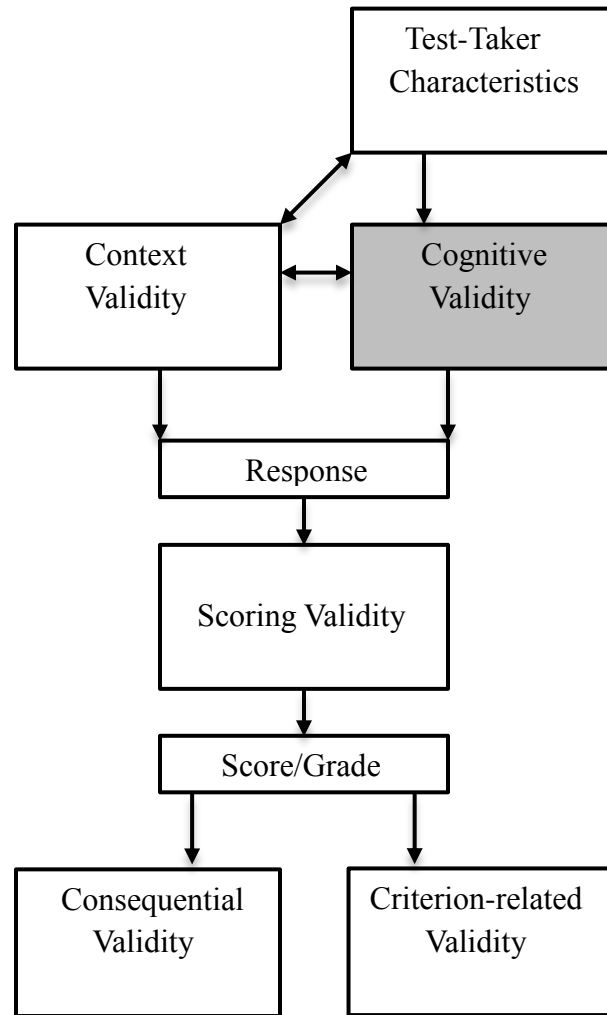


Figure 2. Overview of Weir's (2005) Socio-Cognitive Framework for Test Validity

The framework commences in the top right corner by exploring test-takers' characteristics, which is essential, as the test should ideally cater for test-takers' physical/physiological, psychological, and experiential characteristics. Effectively, these could have a direct influence on both the cognitive validity and the context validity of the test. For example, when a test taker is familiar with taking certain exams as opposed to someone who does not have this experience, it could affect the way he approaches the exam. A test-taker that

prefers to take an MCQ without a time limit may be negatively affected by an exam with open-ended questions that has time constraints (i.e. context validity). This could then affect test performance by generating imbalanced responses and could be a cause of construct-irrelevant variance. This, in turn, could result in a skewed interpretation of the test's scoring validity despite having satisfactorily measured it through required validity parameters such as item difficulty, marker reliability, and internal consistency among others. One can infer from the way the framework is set out that it is important for these three particular segments of the framework (i.e. the test-taker characteristics, cognitive and context validity) to be addressed at the design stage of the test and, therefore, minimize the effect on later (subsequent) segments of the validity framework (i.e. scoring validity, criterion-related and consequential validity). The following segment, i.e. scoring validity, justifies the extent to which the scores obtained from the test could be considered to be reliable. It is, as Khalifa & Weir (2009) say, *symbiotically* related to context and cognitive validity, as the examples of both previous mentioned can impact on, for example, the reliability of the scores on that particular language test. The three together constitute the current interpretation of the inclusive form of *construct validity*. Irrelevancies in construct validity could have consequences, for example, for the test-takers to the extent of affecting their futures. For instance, in high-stakes situations (to give a simplified illustration) this could signify the difference for a university student between entering the College of Medicine, and the College of Dentistry, which would then eventually be the difference between becoming a surgeon and a dentist. These consequences enter into the realm of *consequential validity*, which is, as is *criterion related validity*, beyond the scope of this study to address.

This study investigates the effect of a computer interface (utilized in CBT) on lower-level Saudi Arabian test-takers' cognitive processes and performance when taking a L2 reading test. In

order to investigate the effect of this newly introduced testing mode (i.e. CBT) in this particular context, it was compared to the traditional mode (i.e. PBT) in a repeated measures design where the same test-takers were assessed twice taking the same L2 reading test on separate occasions with ample time in-between the two sessions to control for memory effect (see chapter 3 for a further methodological discussion of this). Equivalence of the two testing modes in terms of utilized cognitive processes in addition to statistical evidence for score equivalence of both tests would be the initial step in investigating the cognitive validity of the newly introduced testing mode in this particular context (i.e. Saudi Arabia) for these particular students. The section that follows elaborates more on cognitive validity and how it is investigated/established in this study.

2.8 Cognitive Validity

Cognitive validity (also known as theory-based validity) in language testing is the extent to which the task/test at hand elicits the (appropriate) cognitive processes relevant to a particular test, and the extent to which it elicits cognitive processes beyond it (Khalifa & Weir, 2009).

Two ways to investigate whether a test can be considered cognitively valid are by either modeling the skill to an expert's behaviour (i.e. target behaviour), or to study candidate behaviour when involved in the test task through verbal reporting to examine to what extent the candidate's processes resemble the skill tested in a non-testing context (Field, 2012). The approach taken to establish whether this study's test is cognitively valid is as follows:

1. By investigating whether the processes anticipated by the test (items) are the processes employed by test-takers when answering these test items through a think-aloud study.
2. Through applying Khalifa & Weir's (2009) cognitive model of reading as the wider *anchor framework* against which the identified processes can be measured.

The following section further illustrates and explains Khalifa & Weir's (2009) cognitive model of reading and how the cognitive processes proposed in this model translate into a testing context.

2.8.1 Establishing Cognitive Validity

Establishing cognitive validity is accomplished through evidencing that the cognitive processes elicited by the test task are representative of the construct it is supposed to measure, which means that evidence from cognitive processing substantiates that the processes employed by test-takers are in agreement with the ones elicited by the test-tasks. For example, the test items in this study are expected to induce local expeditious reading operations to locate relevant information in the text followed by careful reading ensuring to correctly answer the test items. Identifying these processes through test-takers' verbalizations would substantiate the appropriateness of the test items measuring these particular elements of the reading construct. In addition, these processes should reflect processing beyond the test task itself (i.e. in real life). Khalifa & Weir (2009) developed a multi-divisible model illustrating this, which assumes the view of reading as a multidivisible construct where Urquhart & Weir's (1998) reading types initiated by the reading *goal* interact with cognitive processes and knowledge stored in long-term memory outlining the likely processes involved when reading a text in real life. The model is shown in figure 3 below and is subsequently discussed showing the interactions between the three elements (i.e. the reading goal embodied by the reading types, the processing levels, and long-term memory).

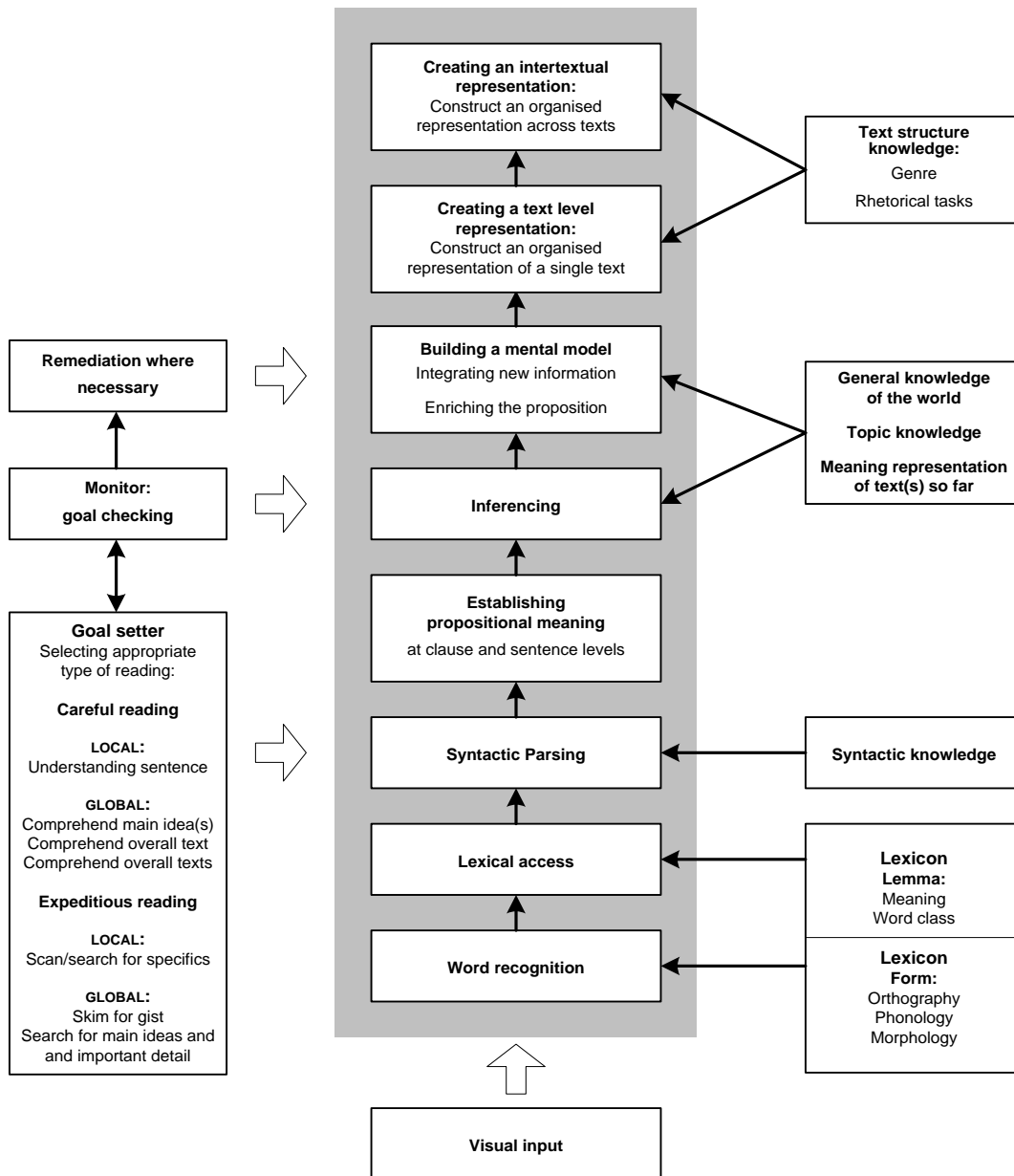


Figure 3. Khalifa & Weir's (2009) Model of Reading

As seen in figure 3 above, the central core shows the various levels of processing involved when reading a text, which is activated by the goal setter to the left of the central core.

The goal setter to the left of the central core serves as a metacognitive element where it activates

the reading type appropriate to the specific reading task. This process is reiterative and could therefore change at any time depending on its goal as shown by the monitor through which the reader (re)evaluates on an on-going basis. The model further shows the connection to the processing elements shown in the central core. In addition, a transition from processing at the local level into more global processing can be identified in the central core of the model, which suggests that the process is hierarchical in complexity. Word recognition, lexical access, syntactic parsing, and establishing propositional meaning are processes that occur at the local level (i.e. word, phrase, clause/sentence level). Inferencing has both global and local applications and is therefore interlinked with building a mental model at either the local level (i.e. word, clause, sentence) or global level (i.e. between sentences, text/passage-level, inter-text level) depending on the set goal i.e. word/clause/sentence comprehension or paragraph/text/inter-text comprehension. The model also illustrates (in the columns to the right of the central core) the role memory has in this process at the various process-levels whether it be at the global or local level through integration of knowledge about, for example, syntactic structures to aid in parsing, or background knowledge aiding in integration of new information in order to build a mental model at either the local or global level depending on the task requirement. How these processes translate into a language test of reading has been clarified by Bax (2013), who illustrated the typical cognitive operations Khalifa & Weir's (2009) identified process types would yield in a language-testing context, as shown in table 3 below.

Table 3. Cognitive Processing in Reading Tests (Bax, 2013).

Process Level	Level of activity (simple to complex processing)	Readers' typical cognitive operations in language tests	Size of typical unit
1	Lexis: word matching	Reader identifies same word in question and text	Word
2	Lexis: synonym and word-class matching	Reader uses knowledge of word meaning or word class to identify synonym, antonym or other related word	Word
3	Grammar/syntax	Reader uses grammatical knowledge to disambiguate and identify answer	Clause/sentence
4	Propositional meaning	Reader uses knowledge of lexis and grammar to establish meaning of a sentence	Sentence
5	Inference	Reader goes beyond literal meaning to infer a further significance	Sentence/paragraph/text
6	Building a mental model	Reader uses several features of the text to build a larger mental model	Text
7	Understanding text function	Reader uses genre knowledge to identify text structure and purpose	Text

The manifestations of cognitive processes in reading tests as illustrated by Bax (2013) are not likely to surpass inferencing at the local level for the test-takers in this study commensurate with Khalifa & Weir's (2009) cognitive model of reading. This is due to the nature of the test task, test items, and test-takers' English proficiency level, which is roughly between A2 and B1 in Common European Framework terms. The test items mainly assess text processing at the local level, which is overall fairly in line with the aforementioned proficiency levels found in Cambridge ESOL exams (see table 13 section 3.7.4.2, for an illustration of the reading types and associated reading operations assessed in Cambridge ESOL exams for both A2 and B1 CEF

levels and how this study's test items relate to them). To locate required information in the text, the test items aimed to elicit local expeditious reading processes such as scanning and search reading in order to contribute to the underrepresented literature in this area as indicated by Urquhart & Weir (1998). In light of Khalifa & Weir's (2009) model this would mean that, as far as reading types are concerned, interactions between mainly *expeditious reading* (to locate required information) and *careful reading* (expected to follow when information has been located) at the local level are expected to be identified from the test-takers in this study corresponding to lower text-level processing. Higher text-level processing would require a combination of reading operations at the local as well as at the global level (see Weir et al., 2000 and Khalifa and Weir, 2009 for examples of this). This study assumes Khalifa & Weir's (2009) cognitive model as a theoretical framework of reference for investigating the processes activated by the test-takers when answering test items interwoven with the types of reading outlined by Urquhart & Weir (1998).

Because this study involved an L2 reading *test*, it was expected to elicit manifestations of behaviour that are different from common reading activities in addition to the discussed reading types due to the goals being distinctly different, i.e. reading to find the answer to a test item as opposed to, for example, reading for entertainment purposes (Cohen, 1986; Farr et al., 1990). As Bax (2013) illustrates, 'The very nature of language tests means that readers frequently jump between the text and test item, and repeatedly regress and jump forward in various ways in their search for answers, in ways quite different from default reading patterns' (p.8). This regressing and jumping is most likely contingent to the particular goal of the test-taker (i.e. successfully answering the items in a language test) and is expected to initiate additional processes specific to this goal. These additional processes are in the literature also referred to as test-taking strategies

(e.g. Kobrin, 2000; Cohen & Upton, 2007; Al-Amri, 2008). Data obtained from the earlier mentioned reading types (Urquhart & Weir, 1998), processes (Khalifa & Weir, 2009), and test-taking strategies employed to answer test items correctly in both testing modes (Kobrin, 2000; Cohen & Upton, 2007; Al-Amri, 2008) through think-aloud reporting is expected to illustrate how lower-level English L2 students approach an English reading test. The first four processing levels shown in the central core of the reading model (i.e. word recognition, lexical access, parsing, and establishing propositional meaning) are thought to be largely automated processes for L1 readers and advanced L2 readers and would, because this, be regarded as *skills* (e.g. Williams & Moran, 1989; Urquhart & Weir, 1998; Afflerbach, 2011). When these processes are automated, verbalization of them might not happen often, which would require different methods to identify how or whether test-takers engage in these microlinguistic processes (e.g. through inferencing), as think-aloud verbalizations are traditionally thought to be a product of merely conscious cognitive operations (Ericsson & Simon, 1993). However, the test-takers in this study are basic L2 readers meaning that automatism for these processes/strategies cannot be assumed, as they might not have reached this level (i.e. it having become a skill) yet in their L2. The implications this would have for the think-aloud reports is that they are expected to reveal a combination of reading operations identifiable through the test-takers reading aloud and other conscious cognitive operations/processes they employ when answering test items such as, for example, responses to problems encountered when executing them (e.g. problem solving processes). Equivalent reading operations and strategies employed by test-takers between PBT and CBT (i.e. RQ2) in addition to score equivalence between the two modes (i.e. RQ1) is a first step towards establishing whether the test used in this study is cognitively valid, which is one of the elements in Weir's (2005) socio-cognitive framework for conceptualizing test validity and is

discussed in section 2.4 of this chapter. Score equivalence, strategy equivalence, and convincing evidence in support of the cognitive validity of this study's test would then contribute to the evidence towards the construct validity of this study's test and simultaneously validate the appropriateness of the interface design developed for the purpose of this study that was devised based on a review of relevant literature from the field of language testing, interface design, and human computer interaction (see 2.7 of this chapter). The following section reviews cognitive processes/strategies that are commonly associated with taking an L2 reading test and closes with a model illustrating the expected cognitive processes to be elicited from the test-takers in this study based on both Urquhart & Weir's (1998) reading matrix and Khalifa & Weir's (2009) cognitive model of reading.

2.9 Strategies in L2 Reading

2.9.1 Distinguishing between Skills & Strategies

Some cognitive processes involved in reading a text are largely automated, subconscious processes (i.e. beyond conscious control), which several reading researchers have defined as (reading) skills (e.g. Williams & Moran, 1989; Afflerbach, 2011). Strategies are thought to act upon these automated (reading) processes (e.g. Cohen, 2005) and have been defined ample times in the reading literature (e.g. Cook & Mayer, 1983; Vann & Abraham, 1990; Brown, 1994). However, arriving at a universally accepted definition of a strategy might not be that straightforward (McDonough, 1995), or even problematic (e.g. Cohen & Pinilla-Herrera, 2009). This is reflected in L2 language learning studies where various researchers had provided their own definitions of the term strategy. Providing these individual definitions was one of Grenfell & Macaro's (2007) main criticisms, as it would add to the difficulty distinguishing between

them. Some researchers, for example, referred to strategies as *moves* (Sarig, 1987), while others named them *choices* (Cohen, 1998), or *actions, steps or techniques* (Phakiti, 2003b). Cohen's (1998) definition implies deliberateness, as a choice is something you make based on a preceding thought (i.e. choose to do something), or a *selection* (Cohen & Upton, 2007). McDonough's (1995, 1999) *articulated plans* wording of a strategy seems to be in agreement with this, as a plan is thought to be something you make (organize) in advance and then execute. This argument is exactly what Paris et al. (1991) and later Afflerbach et al. (2011) used in order to draw a distinction between a reading strategy and a reading skill. Other works in the L2 language learning literature corroborated this view of strategies as being conscious processes (e.g. Williams and Moran, 1989; Feng & Mokhtari, 1998; Urquhart & Weir, 1998; Cohen, 2005), which is in line with McDonough's (1995, 1999) definitions and also Cohen's (1998) earlier mentioned definition where he described strategies as 'mental operations or processes that learners *consciously* select when accomplishing language tasks' (italics added by researcher, p. 92).

Afflerbach et al. (2011), in their literature review, described the defining difference between the two as follows: 'Reading strategies are deliberate, goal-directed attempts to control and modify the reader's efforts to decode text, understand words, and construct meanings of text. Reading skills are automatic actions that result in decoding and comprehension with speed, efficiency, and fluency and usually occur without awareness of the components or control involved' (p. 368). They further mentioned that automated strategies *become* skills, i.e. once a learner becomes proficient in using a certain strategy, it is employed at such high speed that the reader more often than not unconsciously applies it. The implications this would have for identifying these processes through think-aloud reporting is that automated strategies (i.e. skills)

will most likely rarely be verbalized by the reader and would therefore have to be inferred, if obvious. Only when a problem arises with employing strategies, or when the reader has not acquired sufficient speed using a certain strategy to have reached the level of automaticity, processes would likely be revealed through verbalization, which is thought to be the case in this study due to the lower L2 proficiency level of the test-takers. The following section further discusses how these strategies are viewed in a test-taking context.

2.9.2 Defining Test-Taking Strategies

Cohen (1998) mentioned in line with the above, that, when strategies are employed in testing situations, the distinction through the element of consciousness is maintained. Cohen and Upton (2007) clarified this further by saying that in case of test-taking strategies they are, ‘test-taking processes which the respondents have selected and which they are conscious of, at least to some degree’ (p. 211). This would indicate that a strategy that is used to contribute to completing the task at hand during a test will come under the umbrella *test-taking strategies*, which would then also include reading related strategies specific to the test task in that context due to the *goal* of the test-taker, i.e. answering an item correctly. This would then further signify a distinction by definition between general reading strategies employed when reading in daily life activities as opposed to testing situations, which is thought to be the case according to several reading and language testing researchers (e.g. Sternberg, 1991; Kobayashi, 1995; Bax, 2013). However, it is beyond the scope of this study to review the extent to which reading strategies and test-taking strategies overlap or differ and/or try to categorize them in a definite sense, as even in the published literature this still appears to be a point of debate (e.g. Alderson, 2000). Furthermore, the aim of this research was not to investigate the extent to which these strategies do overlap or

differ but rather the effect of the newly introduced mode of testing (i.e. CBT) on the cognitive behaviour of test-takers when completing a reading test reflected through reading operations and (test-taking) strategies was explored (see RQ2). Therefore, in this study, Cohen's (1998) interpretation of test-taking strategies i.e. all strategies employed when completing test tasks (including reading strategies related to answering test-items) will be adhered to for the purpose of simplicity, clarity, and uniformity in terminology used in further discussions. The following section reviews a number of studies that have examined test-taking strategies in PBT and CBT in an L2 reading test. The strategies reviewed in this section will be used as a point of reference to identify strategic behaviour of this study's participants when taking the L2 reading test in PBT and CBT.

2.5.2.1 Test-taking Strategies in Reading Assessment

Researchers have investigated test-taking strategies in the 1970s (e.g. Rowley & Traub, 1977), 1980s (e.g. Anderson, 1989), 1990s (e.g. Kobayashi, 1991; Storey, 1997; Cohen, 1998; Beidel, Turner, & Taylor-Ferreira, 1999) into the 21st Century (e.g. Kobrin, 2000; Abanomey, 2002, Kesselman-Turkel & Peterson, 2004). Collectively these studies have identified a significant number of strategies employed by test-takers when taking a reading test. The majority of these studies, however, were mainly restricted to reading comprehension on paper-based tests. Only a small number of studies involved test-taking strategies employed in either a computer-based reading comprehension test (i.e. Cohen & Upton, 2007) or in both a paper-based and computer-based mode of a reading comprehension test (i.e. Kobrin, 2000; Al-Amri, 2008). Therefore, for relevance purposes, only the studies that investigated test-taking strategies either on computer or both on paper and computer will be succinctly reviewed. The strategies found

will serve as a preliminary guide that most likely will entail many of the common strategies that can be expected to be identified in this study's sample, which will then aid in devising a model describing the expected manifestation of strategies in relation to the reading operations warranted by the test items.

Cohen & Upton (2007) investigated the strategies employed by test-takers when completing the reading subtest of the TOEFL exam on computer. Their sample consisted of 32 students with varying language backgrounds (i.e. Korea, Japan, China, and 'other') each assigned to two of the six subtests of the Language Courseware (2002) materials. Each subtest was between 600 and 700 words long and had 12-13 test items accompanied with it. Verbal reports were used as a means of collecting data about the strategies used by the test-takers. The ten item types included in their study were divided over three main categories; Basic Comprehension tasks, Inferencing Tasks, and Reading to Learn Tasks. Cohen & Upton classified the test-taking strategies found in their study into the following three categories:

Category 1: Reading strategies (i.e. strategies related to the reading of the passage)

These were further divided into four subcategories; the first subcategory was named *Approaches to reading the passage*, which involved strategies such as goal planning, quickly or carefully reading of the passage, reading the whole passage or only a part of it etc. The second subcategory was called *Uses of the passage and the main ideas to help in understanding*, which involved strategies such as rereading to clarify the idea, asking oneself about the overall meaning of the passage or portion of it etc. The third subcategory was named *Identification of important information and the discourse structure of the passage*, which involved strategies such as looking for sentences that convey main ideas, identifying and learning keywords in the passage etc. The fourth subcategory was called *Inferences*, and involved strategies such as pronoun

referencing and inferring meaning of new words using work attack skills.

Category 2: Test-management strategies

The test-management strategies mainly involved strategies that were related to the test items such as going back to the question for clarification, rereading a question item, translating the question or part of it for clarification, predicting answer after having read the item etc.

Category 3: Test-wiseness strategies

Test-wiseness strategies are strategies that are used to arrive at an answer when the test-taker was not able to produce it through the conventional strategies. For example, using the process of elimination with an MCQ item when none of the options stand out as possible answer to the test-taker. The complete list of strategies found in Cohen & Upton's (2007) study can be found in Appendix O.

The remaining two studies that carry relevance to this study as they both involved establishing of cognitive equivalence between a PBT and CBT are Kobrin's (2000) and Al-Amri's (2008). Only the test-taking strategies in Al-Amri's study are discussed here as a guide for determining strategies in this study's sample for the following two reasons:

1. His study was already based on a comprehensive review of the literature which included strategies found in both Cohen & Upton's (2007) and Kobrin's (2000) research.
2. The fact that Al-Amri's (2008) study's context was the same as this study's (i.e. Preparatory students at a Saudi Arabian University), it was expected to reveal more context specific records, which would more likely be useful in informing the segmentation and coding process in this study. However, Al-Amri's (2008), and both Cohen & Upton's (2007) and Kobrin's (2000) studies involved MCQ's (i.e. multiple-choice questions) unlike this study, which consisted of SAQ's (i.e. short-answer questions). Because of this, strategies such as eliminating options are

Chapter 2: Literature Review

not expected to be found in this study due to this innate difference between the two question formats. Al-Amri (2008) placed the strategies identified in his study in the following categories:

Category 1: Affective category

This category included strategies such as self-motivation among others. The strategies in this category were not found (or reported) for example in Cohen & Upton's (2007) study, in particular the saying 'in the name of God' when starting the exam for example.

Category 2: Management category

This category involved strategies related to managing the test itself and was further divided into three subcategories each focusing on a different type of management i.e. overall test management of reading and answering questions, time management, and task management with the latter involving before/during and after task management strategies.

Category 3: (Re) Reading strategies related to individual questions

This category included test related reading strategies and was divided into four subcategories, namely reading of instructions, (Re) reading of the text, (Re) reading of the questions, and (Re) reading of the MCQ options (i.e. a, b, c, etc.).

Category 4: Selecting or attempting to select an answer

This category comprised of strategies related to choosing an answer such as keyword matching, selecting an option through background knowledge, returning to text to confirm selected answer etc.

Category 5: Rejecting or attempting to reject an option

This category involved strategies related to rejection of a possible answer.

Category 6: Reducing options

This category included one strategy only, which was the discarding of options to reduce the

Chapter 2: Literature Review

number of options left. This strategy is in essence very similar to the eliminating of options in Cohen & Upton's (2007) test-wiseness category, as that too is a form of option reducing with a slightly different approach to it.

Category 7: Reviewing / checking a decision

This category included strategies such as reconsidering or double-checking a response and the double-checking of answers during the test (e.g. checking a number of given answers as a group) or after test completion.

Category 8: Changing or attempting to change a decision.

This category involved changing answers after having given them or an attempt to change them

Category 9: Postponing decision

The final category in Al-Amri's study involved postponing a decision by either skipping the question and returning to it later or skipping an option within the item and returning to it later.

The overview of the strategy categories Al-Amri identified is evidently relevant to this study due to both his study's participants and context being very similar. Cohen & Upton's found strategies and categories are relevant as well due to the basic comprehension items included in their study and the processes underlying them. For this reason, the overviews of both Cohen & Upton's (2007) and Al-Amri's studies will serve as a guide to identify strategies in this study. The strategies identified in the PBT and the CBT version of the reading test in this study will be evaluated and if cognitive equivalence is established between the two through processes comparisons, this would be an initial step in support of its cognitive validity as previously discussed in relation to Weir's (2005) framework.

2.10 Theoretical Cognitive Processes Model

To summarize the previous discussions on reading types, process-levels, and test-taking strategies, the devised model below illustrates the manner in which the cognitive processes are expected to occur from this study's test-takers, which was based on an integration of Urquhart & Weir's (1998) identified reading types and Khalifa & Weir's (2009) cognitive model of reading further including Cohen & Upton's (2007) reading and test-taking strategies, and Al-Amri's (2008) test-taking strategies, geared to this study's test purpose.

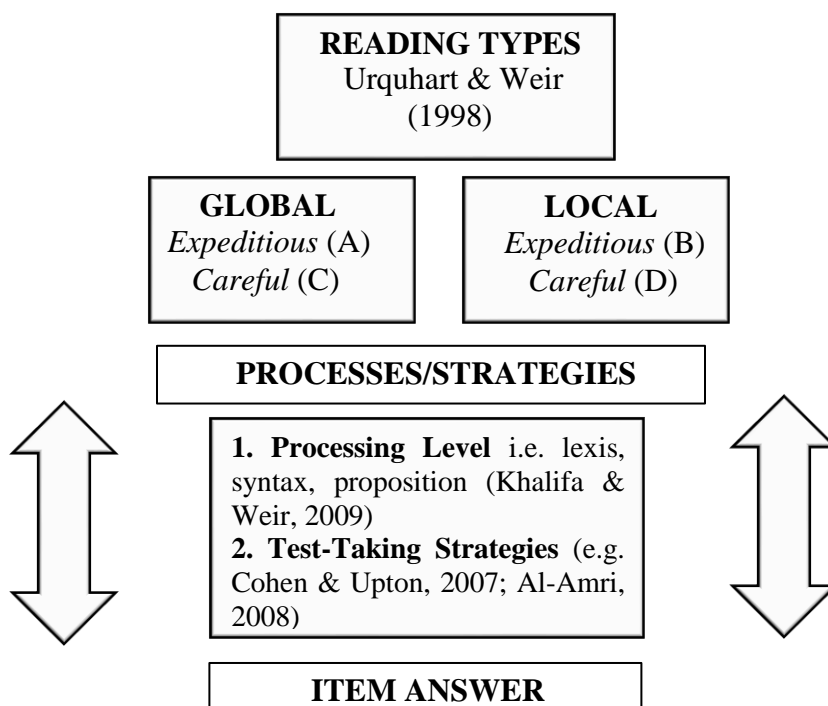


Figure 4. Model for Expected Manifestation of Cognitive Processes in this Study

Figure 4 above illustrates the expected manifestation of test-takers' cognitive processes in this study reflected through reading types, processing levels, and test-taking strategies. The processes manifestation is categorized into two stages. Depending on the nature of the test item, the goal setter activates one of the four reading types (A, B, C, and D). In this study's case, this

would be local expeditious reading (B) employing operations such as scanning and search reading in order to locate the relevant information to the question item (stage 1). Once found, it is expected that the test-takers resort to more local careful reading behaviour (D) in order to accurately understand the clause/sentence where the located keyword/ answer is in (Urquhart & Weir, 1998). This is when further levels of processing are activated depending on what processing level is required to answer that particular test item, e.g. lexical, syntactical, propositional, etc. (Khalifa & Weir, 2009; Bax, 2013). This careful reading behaviour is expected to be further exacerbated by the fact that open-ended questions are used in this study's test, which forces the student even more so to ensure that he completely understands the sentence in which the phrase/clause containing the answer is located, as failing to do so might lead to provide incomplete, and therefore, incorrect answers to the test items. As the model further shows, the reading type(s) activated by the goal setter and subsequent reading operations/strategies could change continuously as it is a potentially reiterative process, which is contingent to the goal of the test-taker. This could in the case of this study's items be, for example, within a reading type, when a test-taker is unable to locate the information related to a test item and adjusts his approach through employing a different reading operation in order to succeed in locating the answer to that particular test item (e.g. from scanning to search reading). It could also be intertype related. For example, employing reading type B to locate information, and, when found, switching to reading type D to sufficiently comprehend the sentence containing that information to answer the test item correctly. The model further implies that, depending on the nature of the test item, different strategy patterns could emerge through combinations of operations and strategies test-takers use when answering the test items.

As for the exact order and what exact strategies are expected to emerge from this, apart

from the anticipated reading types, is not specified because this part is largely exploratory due to the nature of the test items and assessment format in the target context. The results from investigating these processes are therefore expected to contribute substantially to the field of L2 reading and language testing research as very few studies have actually addressed this, and, in this particular context using an L2 reading test with short-answer question items, it is the first study of its kind to the researcher's knowledge. The short-answer question format used in this study is discussed in more detail in the following section.

2.11 Reading Assessment Format

There are a variety of assessment formats in language testing such as open-ended questions, matching, gap filling, summarization, true or false questions, multiple-choice questions, short answer questions, each of them having its advantages and disadvantages. It is beyond the scope of this study to review all and therefore only the SAQ assessment format will be discussed here with specific references to some of the other question formats merely for comparison purposes.

2.11.1 Short Answer Questions (SAQ's)

Short answer questions (henceforth, SAQ's) are questions that require the examinee to write a response consisting from a word or a couple of words up to a complete sentence or sometimes two. SAQ's have been part of reading assessment since the beginning and have been argued to be a plausible way of assessing reading comprehension and have qualities that supersede other question formats.

Alderson (2000) mentioned that in some respects SAQ's better reflect whether test-takers

have understood the question compared to, for example, Multiple-Choice Questions (MCQ), as an answer to an MCQ could be a result of elimination, which is not possible with SAQ's as they require test-takers to generate the answer from the question stem. Magliano et al. (2007) clarify this distinction as follows: 'Short answer questions require readers to generate the answer themselves on the basis of the question stem, which makes a short-answer question distinct from multiple-choice questions, which can be answered partly on recognition memory, information search in the target passage, and reasoning' (p.116). Alderson et al. (1995) further argue that using this type of question format (i.e. SAQ) forces the test-taker to think up the answer for himself. Due to this, Alderson et al. (1995) stress that SAQ's used for reading tests could be 'revealing' (p.59) by potentially showing 'textual misunderstandings' (ibid. p.59) that would otherwise not have been detected by the writer of the test. Other researchers such as Hedgcock & Ferris (2009) and Cunningham (1998) share this view stressing this advantage when using SAQ's particularly over controlled responses such as MCQ's, seeing it as a 'practical alternative' (p. 354) to using MCQ's. Weir (1990) mentions that carefully formulated SAQ's can be potentially useful, provided the answer required is a *brief answer*, i.e. limited to a word or phrase as opposed to a maximum of two sentences that likewise fall into the category of short-answer questions. Bachman & Palmer (1996) refer to this type of SAQ as a limited production response or 'short completion items' (ibid: p.54). SAQ's can be used when assessing a number of strategies such as scanning, skimming, search reading (Weir, 2005), and prediction (Hedgcock & Ferris, 2009), in addition to careful reading. Hedgcock & Ferris (2009) highlighted a number of reading skills such as comprehension and interpretation that could effectively be assessed by SAQ's. Interpretation and comprehension mentioned by Hedgcock & Ferris (2009) can occur at the explicit level as well as implicitly (e.g. Ehara, 2008), and, therefore, at the surface, would

seem to contradict the previous mentioned claim. However, assessing implicit information requires longer answers and would therefore most likely fall into the category *extended production response* instead of the *limited constructed response* (Bachman & Palmer, 1996), which, in turn, would require a more holistic approach when assessing (Van Blerkom, 2009). Furthermore, SAQ's are potentially more efficient in assessing certain reading strategies (e.g. scanning, skimming, and search reading) as opposed to using indirect question formats such as, for example, gap filling. Weir (2005) illustrated this through an example where finding specific information or information pieces can be addressed more effectively by employing scanning strategies as opposed to carefully reading the whole text. The test-takers in this study are lower-level EFL-students and, due to the nature of the test's items, processing is expected to occur largely at the local level where typically explicit information is elicited through basic comprehension questions. This means that, for the most part, reading operations and strategies at the local level are expected to be involved, which makes SAQ's, a valid question format to assess these.

However, Alderson (2000) cautioned regarding developing SAQ's stressing that constructing them is not easy. For example, one of the main concerns expressed about using SAQ's when assessing reading is that it potentially affects the stability of scoring between different raters, which is essential for test fairness and reliability (e.g. Davies et al., 1999; Alderson, 2000; Hedgcock & Harris, 2009). In order to minimize this potential problem, several suggestions surfaced from the reading and language testing literature. One of the suggestions was to include a comprehensive answer key anticipating various responses to a single question (e.g. Alderson et al., 1995; Alderson, 2000) and by including possible alternative answers (Davies et al., 1999). Davies et al. (1999) further added that clear instructions included for the test scorers

explaining how to judge errors made in grammar and spelling could prove to be helpful in this. Another way to help accomplish scoring stability could be by reducing the selection of possible responses in a systematic way through content analysis leaving only a handful of key issues for scoring (Dörnyei, 2003). This could then provide the opportunity to score the items automatically against a template (Fulcher, 1999), which, in turn, would prevent possible inconsistencies between different raters from occurring. The SAQ's in this study elicit mainly explicit information derived *directly* from the text and therefore further contribute to minimizing this possible problem as the answers are mostly literally taken from the text instead of having to use more global processing to formulate an answer.

Another possible problem with SAQ's is the writing activity itself in relation to extended answers. Weir (2005) noted that some research (although largely anecdotal) showed that, because test-takers are involved in writing when answering test items, it could possibly affect the construct measured. Results from a number of comparability studies indicate that this could also be related to the level of familiarity a test-taker has using a computer than the actual writing process itself when it comes to CBT. For example, Russell & Haney (1997) found that the middle school students in their study that were accustomed to writing on computer scored significantly higher on their CBT writing test compared to PBT than those who were not. They concluded that, based on this, students' PBT results on a writing task could be a significant underestimation of their writing ability on computer. However, Yu (2010) who investigated the effect of computer familiarity on summarizing ability involving 157 undergraduate students found that computer familiarity did not affect this ability at all, which makes these results inconclusive. Nevertheless, because of the computer familiarity of the test-takers in this study, it is not expected to affect this ability, if anything, it would then most likely work in their

advantage when answering the test items in CBT. The section that follows is an overview of the reviewed literature so far, and its relation to the research questions.

2.12 Overview Reviewed Elements

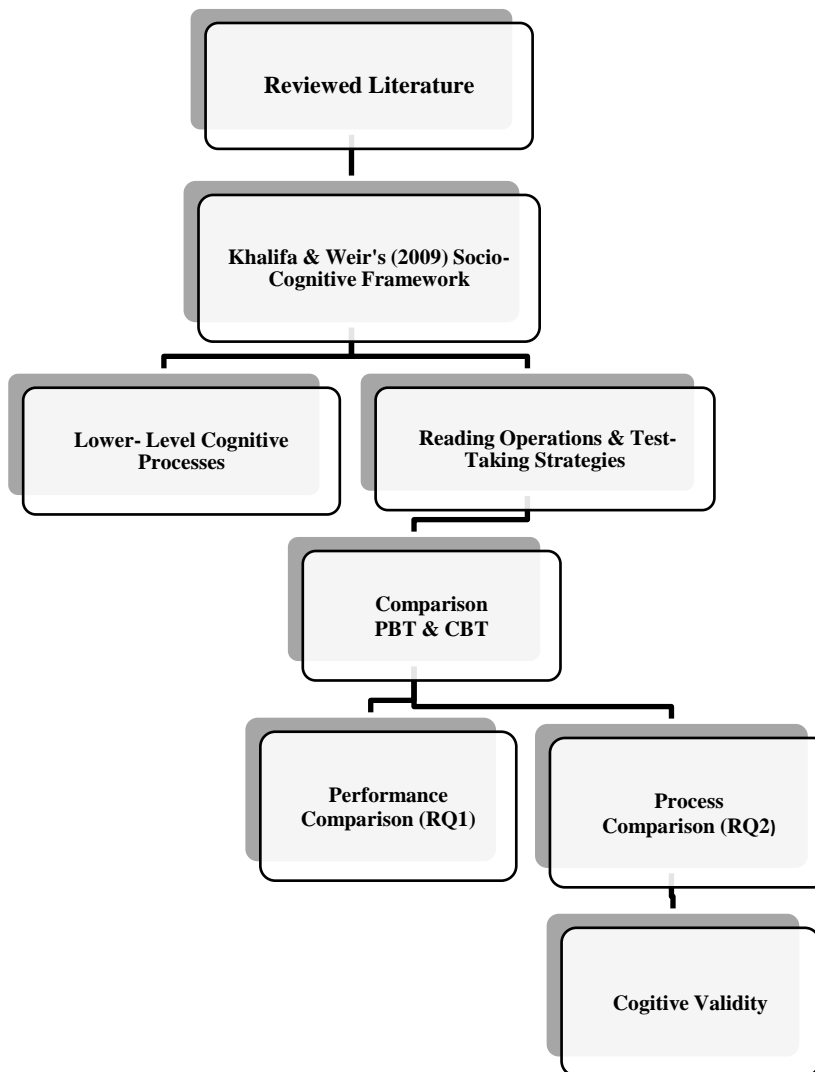


Figure 5. Reviewed Elements in Literature

Figure 5 above shows the reviewed literature and its relation to the research questions formulated to achieve this study's objectives. Khalifa & Weir's (2009) cognitive model of reading, which also assumes Weir's (2005) socio-cognitive framework for test validity, is the framework within which this study works. The two main elements of this framework i.e. reading operations to locate specific information within the text and the levels of cognitive processing dependent on the task at hand are firstly reviewed in order to further guide the research objectives (i.e. the type(s) of reading operations and the levels of cognitive processing expected to be executed by the test-takers). Because the same test was used in both testing modes, the reading operations and levels of cognitive processes could be compared between the PBT mode and CBT mode (i.e. RQ2) in addition to test-takers performance in both modes (i.e. RQ1). Furthermore, provided the processes comparison yielded no significant differences between the two modes, cognitive validity could be investigated for the CBT using Khalifa & Weir's (2009) framework as an anchor for establishing this. Furthermore, performance equivalence and process equivalence would provide further supporting evidence for the construct validity of this study's test in this particular context.

Another identified gap in the literature (as mentioned in section 1.3 of the introductory chapter) is the lack of research into the effect of a computer interface on test-takers' processes and performance when taking a reading test in CBT. There is no study to date that has synthesized the individual elements of an interface and investigated the possible effect of the elements together on test-takers. The interface which was developed in order to contribute to the field of language testing by addressing this gap by making it part of this study's independent variable (i.e. CBT) is comprehensively discussed in the section that follows.

2.13 Computer Interface Design

2.13.1 Introduction

The indications from the previous mentioned studies in the introductory chapter (section 1.1) that emphasized the importance of good computer interface design in computer-based testing show the need that exists for a study as the current one that investigates the possible effect of interface design on test-takers' processes and performance. In order to address this problem, literature related to computer interface design is reviewed focusing on elements of the interface that could affect this. By synthesizing the available literature on interface design, the optimal settings of the various elements of the interface are thought to be identified and could then be integrated into the development of the interface in this study in order to investigate its research questions and test its research hypotheses. An evaluation model is proposed through which the elements of the interface are reviewed based on various sources combined including previous reviews that have attempted to categorize interface design elements in their evaluations (e.g. Dillon, 1992; Muter, 1996; and more recently, Leeson, 2006).

2.13.2 Earlier Reviews

Two comprehensive reviews that have identified variables related to reading from screen are that of Dillon (1992) and Muter (1996). Dillon (1992) evaluated the literature on paper vs. screen reading from Schumacher and Waller's (1985) perspective emphasizing on *process measures* and *outcome measures*. He clarified the process as the way the reader uses the text whereas the outcome related to what the reader gets from it (i.e. the effect). The process measures consisted of eye movement, manipulation, and navigation, whereas the outcome measures covered reading speed, accuracy, fatigue, comprehension, and preference. Muter

(1996) identified variables that could aid in optimizing reading from screen based on studies dating from within the same timeframe. However, it should be mentioned that the majority of the VDU's in these reviews stemmed from the 1980s and included computer devices such as the IBM microcomputer, the Apple IIe and the II plus, which were the contemporary machines at the time. The quality of the screen-displays back then differed considerably from today's, and was likely to have affected the subjects' interaction with the VDU's and, therefore, study results. In addition, computer familiarity was generally lower back then compared to today, which could also have influenced the results obtained at that time (Belmore, 1985). This means that, applying any conclusions drawn from the aforementioned reviews to computer screens developed from the early 90s onwards may be questionable (Dyson, 2004) and would need critical evaluation before assumptions can be made based on them. Dillon (1992) himself shared this concern and further highlighted methodological shortcomings in a number of the studies he reviewed mentioning the limited scope, lack of controlling the variables included in the studies, vague criteria for selecting study participants, and the improbable nature of the reading tasks themselves (as a number of them involved proofreading). For these reasons, the variables addressed in the studies reviewed by Dillon (1992) are only referred to in this study when they show direct relevance to current interface design issues related to more contemporary computer devices (from mid-90s onwards) and/or when the lack of more recent studies addressing a particular variable necessitates this for contextualization purposes.

The overarching aim of this review is to look at the most significant relevant factors that could account for differences when taking a reading test on screen (i.e. through the interface). As a computer interface is made up of a combination of factors interacting with each other contributing to that interface, it is important that when developing one, these factors are

combined in their optimal settings to minimize mode-effect (Dyson, 2004).

Leeson's (2006) study is the most recent review of human computer related factors related to the field of language testing. The variables included in her review were selected from Muter's (1996) study on optimizing reading from screen. Although Muter (1996) identified 29 possible factors in his review, Leeson (2006) argued that she had made her selection based on the relevance of the variables to current screen technology and software. In her review she distinguished between presentation related factors, which she referred to as legibility and interaction related factors, as illustrated in figure 6 below.

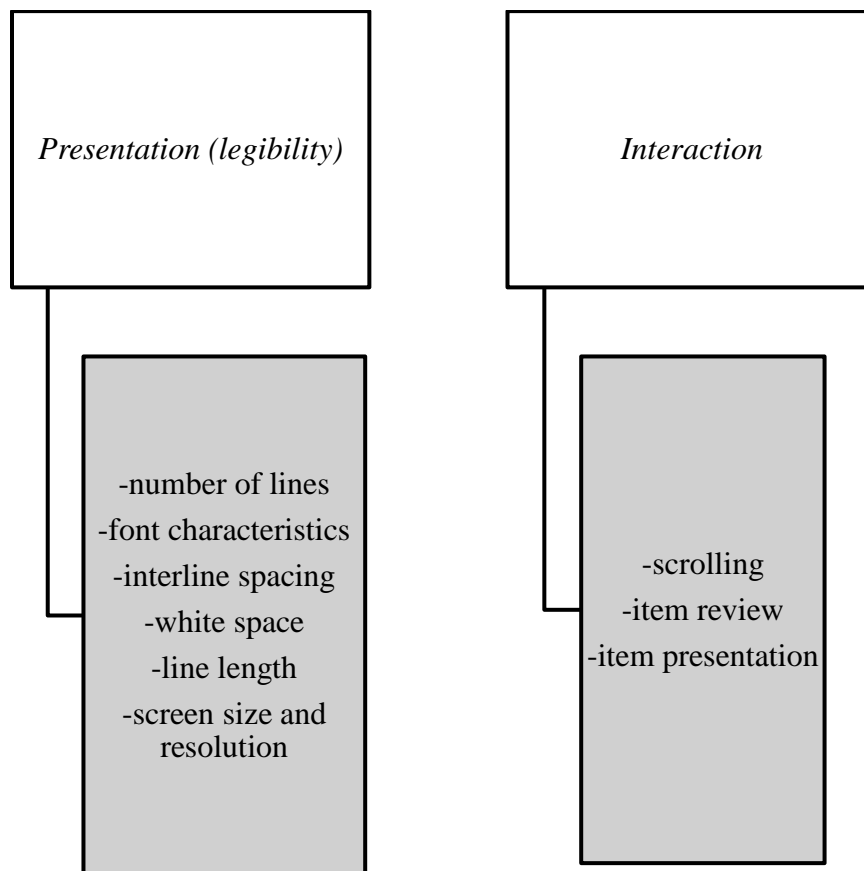


Figure 6. Identified Human Computer related Variables in CBT by Leeson (2006)

It is important to mention that the majority of the variables in Leeson's (2006) presentation column are either directly related to text presentation (i.e. font characteristics,

number of lines, line length) or indirectly (i.e. white space, interline spacing), which collectively are in the typography and HCI-literature referred to as *typographical factors* (e.g. Kahn & Lenk, 1998; Dyson, 2004). These typographical factors, however, do not function independently but are confounded as some factors directly influence others. For example, line length could refer to the physical length of the line or to the number of characters used. When it refers to the number of characters used, changing the type size could affect the number of characters per line directly. This means that when evaluating typographical variables, they are to be evaluated in relation to each other first in order to maintain internal validity (e.g. Lund, 1999). Lund (1999) argued that when certain variables are manipulated invariably, the ratio between the two differs when, for example, line length increases without a corresponding increase in interlinear spacing. Therefore, in evaluating interface design on presentation related factors in this study, typographical factors are discussed separately, making a distinction between typographical elements (text related) and graphical elements (format related). This does not, however, negate interactions between the two, nor does it negate interactions between presentation and interaction related variables in general, as it is the combination of these variables in particular settings in relation to each other that make up the interface with which the user (i.e. test-taker) interacts (Peters, 1992). The review below aims to address this and distinguishes between presentation related factors and interaction related factors as suggested by Leeson (2006).

2.14 Interface Evaluation Model

Leeson's (2006) review of human and technological (computer) related issues in language testing interpreted and incorporated two elements of a computer interface; a presentation related element and an interaction related element. This study will discuss human-

computer related variables building from Leeson's (2006) review and indicated in the two earlier comprehensive reviews by Dillon (1992) and Muter (1996) through distinguishing between presentation and interaction and reviewing the two independently. Based on both the theoretical understanding of a computer interface and its practical application by the aforementioned reviews (i.e. Dillon 1992, Muter, 1996, and more recently, Leeson (2006), and the purpose of this study (i.e. investigating the effect of interface design on test-takers' processes and performance) the following model shown in figure 7 below was developed for evaluating the elements of a computer interface that could possibly affect test-taker behavior and, subsequently, performance on computer.

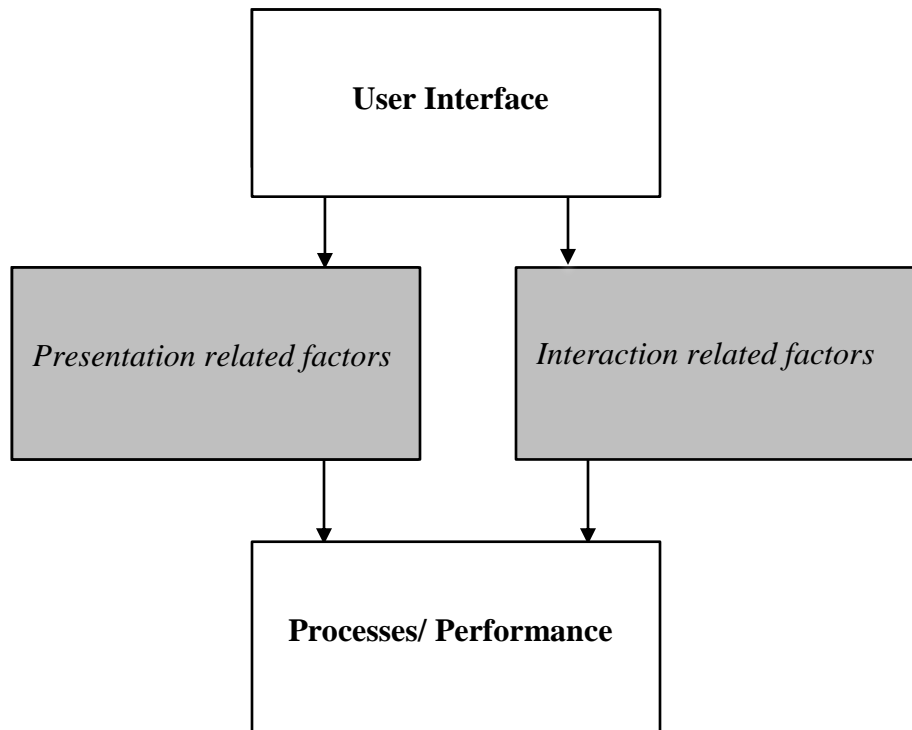


Figure 7. User Interface Evaluation Model for a Computer Based Language Test

In line with the model shown in figure 7 above, the interface will be developed according

to its optimally configured presentation and interaction related factors, which is therefore expected to minimize any possible effect on either test-takers' processes or performance. The review of the related literature is discussed in line with the model starting with presentation related factors, and, subsequently, moving towards interaction related factors from which an interface design template emerges comprising the optimal settings for assessing an L2 reading test in the target context to conclude this chapter.

2.15 Review Interface Design: Presentation (typographical factors)

2.15.1 Font Characteristics

Geraci (2002) reviewed thirty studies related to typography, layout, color, and screen density that were published during the, at that time, past decade dating from 1992-2002. Out of the thirty studies reviewed, four provided relevant information regarding font size and type (i.e. Bradshaw, 1998; Harrell, 1999; Horton, 2000; Skaalid, 2001). Although the mentioned studies were published within a three-year timeframe, results were inconclusive as to which typeface was most legible for reading from screen. For example, the first study of Bradshaw (1998) concluded that sans-serif typefaces were the typeface of choice to be used on computer. However, this appeared to be in disagreement with Harrell's (1999) study that found that the results from the subjects in his study suggested a strong preference for serif fonts. Geraci (2002) further referred to another study by Horton (2000) who concluded that the serif font Times New Roman were to be used for body texts, and Arial and Verdana, both sans-serif fonts, were preferred to be used for navigation links. The fourth reviewed study of Skaalid (2001) concluded that Georgia (serif) and Verdana (sans-serif) (both designed specifically for on screen texts) were the preferred typeface choices on screen.

Weisenmiller's (1999) study specifically addressed the two previously mentioned fonts (Georgia and Verdana) in his study as he aimed to investigate whether fonts specifically designed for on-screen text displays were more readable than fonts originally designed for printed texts. He compared a computer-designed serif font (Georgia) and sans serif font (Verdana) to the traditional serif (Times New Roman) and sans serif (Arial) (all at a 12-point type size) by measuring performance levels on reading rate and comprehension for which the Nelson Denny Reading Test (NDRT) was used. There were three testing conditions: two on computer screen (i.e. 1-bit rendering of onscreen text through Microsoft Office, and 8-bit rendered Adobe Portable Document format) and one on paper (i.e. 600dpi text rendered on paper). Both the paper condition and both screen conditions were displayed at a 100ppi resolution with a screen setting of 1280x1024 pixels on a 17" screen. A total of 264 University students of whom 95% majored in Industrial Technology were divided into twelve groups of twenty-two subjects to investigate the effects of the fonts in the three testing formats. 74% of the participants were male and 26% were female with a median age of 21.5 years. Although it was possible to navigate through the text by scrolling, participants were instructed to only use the paging option by pressing the up and down buttons in order to prevent possible inconsistencies between the paper and computer versions. The study results showed neither a significant difference in reading rate nor in reading comprehension between the different fonts. However, Weisenmiller (1999) did find significant differences between presentation modes. He found that 1-bit on-screen text presented through Microsoft Office was significantly less readable than both the 8-bit text presented through Adobe Reader 3.0, and the 600dpi paper version whereas the text in the 8-bit presentation, however, was not significantly less readable than the 600dpi text on paper.

Bernard & Mills (2000) compared font types Times New Roman font and Arial font in

Chapter 2: Literature Review

size (i.e. 10 vs. 12 point) and text format (i.e. dot matrix vs. anti-aliased) to determine which combination produced the highest readability (by testing accuracy). The thirty-five subjects included in the study were asked to read eight similar level passages each consisting of around one thousand words as quickly and as accurately as possible. The screen resolution was set to 1024x768 pixels to mirror contemporary devices' screen resolution. Fifteen words were substituted with context irrelevant words that rhymed to the original (correct) word. The participants were told to identify these and verbally communicate them when found. No comprehension differences were found between the types, sizes, and format. However, the participants gave preference to Arial over Times New Roman at 12-point type size. Times New Roman was read fastest at the 12-point type size compared to Arial font. Arial 12 was perceived by the subjects to be most legible followed by Times New Roman 12. In a following study, Bernard et al. (2001) compared twelve different font types including Sans-Serif fonts (N=5), Serif fonts (N=5), and Ornate fonts (N=2). Bernard et al. (2001) mentioned that according to a general web survey the text of most web sites consists of a 12-point type size. Therefore, they kept the fonts evaluated in their study at the same size except for the Agency font, which had to be increased to a 14-point size in order to reach the same height as the other eleven fonts. The fonts included in their study are displayed in table 4 below.

Table 4. Included Font Types in Bernard et al.'s Study

<i>Sans Serif Fonts</i>	<i>Serif Fonts</i>	<i>Ornate Fonts</i>
Agency FB (Agency)	Courier New (Courier)	Bradley Hand ITC
Arial	Georgia	(Bradley)
Comic Sans MS (Comic)	Goudy Old Style (Goudy)	Monotype Corsiva
Tahoma	Century Schoolbook	(Corsiva)
Verdana	(Schoolbook)	
	Times New Roman (Times)	

Bernard et al. (2001) used a 17” monitor with a screen resolution of 1024x768 pixels to administer 12 passages of approximately 1000 words each to twenty-two participants with an average age of twenty-five years. Each passage represented one font type. They found no effect on reading efficiency at the 12-point font size, which is in agreement with earlier studies such as Bernard & Morrison’s (2000). Furthermore, no significant effect on legibility was detected between the fonts examined although perceived legibility by the participants showed that Courier, Comic, Verdana, Georgia, and Times New Roman were found more legible than the other fonts.

In a follow up study, Bernard et al. (2002) compared four serif fonts (Courier New, Georgia, Century School Book and Times New Roman) to four sans serif fonts (i.e. Arial, Comic, Tahoma and Verdana). However, this study did not look at reading accuracy but rather included reading speed, perceived legibility and preference. As for the reading times, serif fonts were read significantly faster than the sans serif fonts. Verdana, Georgia, and Times New Roman were perceived as being most legible among the eight fonts included. The sans-serif fonts were preferred over the serif fonts.

Chapter 2: Literature Review

Geske (2000) conducted a study in which speed and comprehension were measured comparing serif and sans serif fonts in 10, 12, and 14-point settings. A total of 78 University students (56% Male and 44% Female) with a mean age of 21.2 were randomly assigned to read a paragraph of about 225 words either in sans serif or serif. After the subjects had read the paragraph, five multiple-choice questions were asked about the content to assess comprehension. All paragraphs had a Flesch readability of 7.5 to ensure appropriateness and fairness in difficulty level. Results showed no significant comprehension differences between serif and sans serif font in any of the type size settings (i.e. 10, 12, and 14-point). However, within font types, comprehension was significantly better with the 12 –point setting over the 10-point setting in both serif and sans serif fonts. For the serif font, this was also the case for the 12-point setting compared to 10- point and 14-point fonts. Geske (2000) concluded therefore that a 12-point type size was the best choice for text comprehension irrespective of the font used (i.e. serif vs. sans-serif), which seems to diverge somewhat from what had been suggested in the early literature assuming the legibility of type on screen increases by increasing its size (e.g. Griffing and Franz, 1896; Roethlein, 1912).

Morrison & Noyes (2003) compared a 12-point ornate sans serif font (Gigi) to a 12-point traditional serif font (Times New Roman) in their study. They used a 13.3” monitor with a 1024x768 screen resolution to administer four paragraphs of 140 words each to twenty-five participants (13 Male 12 Female) all having normal 20/20 vision. Morrison & Noyes (2003) further added that all participants were computer familiar, however, they did not mention to what extent. Like the Bernard & Mills (2000) study, recognition of context irrelevant words was used to test reading accuracy substituting ten words instead of fifteen. Results showed that comprehension was significantly better for the serif font over ornate sans serif font, which

differed from earlier findings (e.g. Bernard & Morrison, 2000; Bernard et al., 2001). Contrary to the Bernard & Mills (2000) study, the participants in this study favoured the serif font (TNR) over the ornate sans-serif font (Gigi) although they found the latter more attractive than the former. Morrison & Noyes (2003) concluded with stressing that the results in their study regarding font types and sizes should not be interpreted in isolation when evaluating online text, as readability could be affected by a combination of a number of factors among them being line length, word spacing, white space, and italics.

Chaparro et al.'s (2006) study examined the legibility of two Microsoft developed Clear Type serif fonts i.e. Cambria and Constantia, that were to be introduced on the new Vista operating system by comparing them to the more traditional Times New Roman font. They used a Dell Pentium IV laptop with a screen resolution of 1400x1050 and a 60Hz refresh rate to present twenty-six lower-case letters, digits (i.e. 0-9) and symbols in an 8-point font size at an exposure time of 34ms with 1.5 seconds blanking time. Each font was trialed five times with 230 characters presented per trial. Results showed that Cambria had the highest overall percentage correctly identified characters, which included letters, digits, and symbols (92.87 %) followed by Constantia (87.80 %) and then Times New Roman (87.55 %). The overall legibility was best when identifying letter characters for all three fonts, whereas the main differences between the fonts were found with the digit and symbol identifications at the 8-point type size.

Beymer et al.'s (2007) study used eye tracking to examine the effect of font size, font type, and pictures on online reading. In eye tracking, eye fixation points are registered by reading analysis software that identifies linear grouping within successive fixations. These are then analyzed by line matching algorithms that match these fixations to the actual lines in the text. Their study involved 114 participants of which seventy-four males and forty females with a good

Chapter 2: Literature Review

distributed age range between 20-60+ years. The computer used was an IBM T41 laptop where text was read from an Internet explorer browser. However, the researchers did not mention any screen resolution details. The fonts compared were serif font Georgia and sans-serif font Helvetica at the 10, 12, and 14-point type size. Although results showed no significant differences for comprehension, a 12-point font size produced a slightly higher retention rate (90.1%) compared to the 10-point font used (89.2%) and the 14-point font (88.9%). The font type comparison between sans serif and serif yielded identical results with both fonts producing a 75.6 % retention rate.

In a more recent study, Banerjee et al. (2011) examined the effect of typeface and type size on young adults reading on screen text. They compared three serif fonts (TNR, Georgia, and Courier New) to three sans-serif fonts (Arial, Verdana, and Tahoma) in 10, 12, and 14-point type sizes. A total of forty young adults (21=M, 19=F) with a mean age of 27.5 years all having 20/20 or better vision participated in the study. A 17" TFT-LCD monitor was used with a screen resolution of 1280x1024 pixels and a 60Hz image refresh rate. Participants were asked to read eighteen passages with an average of 657 words per passage, each representing a font type/size combination. The authors further mentioned that all passages were about the same difficulty level. The overall results indicated a better readability for Serif fonts compared to Sans-Serif fonts with Courier being read fastest at 14-point type size and Verdana had the least mental workload at a 14-point type size. The authors recommended, therefore, based on their results, a 14-point Courier New typeface/type size combination and a Verdana 14-point typeface/type size combination for on screen reading. Table 5 below summarizes the studies reviewed.

Table 5. Summary Reviewed Studies Addressing Typeface and Type Size

<i>Study</i>	<i>Typeface</i>	<i>Type Size</i>	<i>Comprehension</i>	<i>Legibility</i>
Weisenmiller(1999)	Serif vs. Sans-Serif	12	No significant difference	N/A
Bernard & Mills (2000)	Serif vs. Sans-Serif	10 - 12	No significant difference	Ariel 12
Bernard et al. (2001)	Serif (x5) vs. Sans-Serif (x5) vs. Ornate (x2)	12	No significant difference	Courier, Comic, Verdana, Georgia, TNR, 12
Geske (2000)	Serif Palatino vs. Sans-Serif Helvetica	10-12-14	No significant difference	Palatino & Helvetica 12
Morrison &Noyes (2003)	Serif TNR vs. Ornate Sans-Serif	12	Serif significantly better	TNR 12
Chaparro et al. (2006)	Serif Cambria & Constantia vs. Serif TNR	8	Clear Type Serif significantly better	Cambria 8
Beymer et al. (2007)	Serif Georgia vs. Sans-Serif Helvetica	10-12-14	No significant difference	Georgia & Helvetica 12
Banerjee et al. (2011)	Serif TNR, Georgia, Courier New vs. Sans-Serif Arial, Verdana, Tahoma	10-12-14	N/A	TNR, Georgia, Courier New 14

Out of the eight studies reviewed, five compared serif fonts to sans-serif fonts (i.e. Weisenmiller, 1999; Bernard & Mills, 2000; Geske, 2000; Beymer et al., 2007; Banerjee et al., 2011), one compared a serif font to an ornate sans-serif font (i.e. Morrison & Noyes, 2003), one compared serif, sans serif, and ornate fonts (Bernard et al., 2001), and one compared a traditional serif font (TNR) to serif fonts designed specifically for on screen (Cambria & Constantia) (Chaparro et al., 2006). Five out of the eight studies reviewed found no significant differences on comprehension whether the font size was kept the same (Weisenmiller, 1999; Bernard et al., 2001) or between different type sizes (Bernard & Mills, 2000; Geske, 2000; Beymer et al.,

2007). One study that compared a serif font to an ornate sans-serif font found the serif font to be superior to the ornate-sans serif font (Morrison & Noyes, 2003) and one study that compared a traditional serif font to specific serif fonts designed for computer found that the latter were superior to the former (Chaparro et al., 2006). Although the study of Banerjee et al. (2011) did not look at comprehension specifically, they did find a readability advantage by serif fonts over sans serif on all three type sizes and was therefore included in the review. As for legibility, the participants in three out of the four studies that compared typeface in different type sizes found that a 12-point type size was most legible (Bernard & Mills, 2000; Geske, 2000; Beymer et al., 2007) and found a 14-point type size to be most legible in one study (Banerjee et al., 2011). Out of the four studies where type size was kept constant, one study's participants found serif fonts Courier, Georgia, TNR, and sans serif fonts Comic and Verdana to be the most legible typefaces at a 12-point type size (Bernard et al., 2001). One study found a superior perceived legibility of the serif font (TNR) at the 12-point level (Morrison & Noyes, 2003), and one study found Clear Type serif font Cambria the most legible at the 8-point type size followed by Constantia compared to TNR at the same size (Chaparro et al., 2006). Worth to note is that Chaparro et al. (2006) did not investigate the mentioned fonts in relation to other type sizes and therefore results cannot be generalized to larger type sizes. The remaining study of Weisenmiller (1999) did not provide legibility information.

The results from the majority of the studies reviewed indicate that at the 10, 12, and 14-point type size, there are no significant comprehension differences between the serif fonts and the sans serif fonts examined. In addition, (albeit subjective) legibility results indicate that certain fonts at particular type sizes are perceived by study participants to be more legible than others. With this in mind, the following recommendations are made regarding optimal typefaces

and type sizes to be used in this study's reading test as a conclusion to this section.

Table 6. Suggested On-Screen Typeface/Type Size Settings

<i>Typeface</i>	<i>Type Size</i>
<u>Serif</u>	
Courier New	12-14
Georgia	12-14
Palatino	12
Times New Roman (TNR)	12-14
<u>Sans Serif</u>	
Ariel	12
Comic	12
Helvetica	12
Verdana	12

2.15.2 Line Length (characters per line)

Dyson & Kipping (1998) looked at the effect of line length on reading when scrolling vs. paging in two consecutive experiments. Twenty-four students participated in the experiment and they were asked to read six documents in 25 cpl, 40cpl, 55cpl, 70cpl, 85cpl, and 100cpl line length settings. They were further asked to compare every other document and report on which document they thought was easier to read. The typeface used was Arial at a 10-point type size and a 12-point interlinear spacing (i.e. the space between each line) with additional 12-point spacing between each paragraph. Results showed no significant differences in comprehension between the six line length settings. Furthermore, no signs of speed-accuracy trade-offs between reading rate and comprehension were found. As for perceived reading ease, a medium line length of 55cpl was reported easier to read than 100cpl and 25cpl, whether the participants scrolled or paged through the texts. In the second experiment, Dyson & Kipping (1998)

controlled for possible glare by replacing the bright white background with a gray background. However, the results of the second experiment were commensurate to the first despite the applied changes.

Dyson & Haselgrove (2001) looked at the effects of reading speed and line length on reading rate, scrolling patterns, and comprehension. They compared lines at 25 cpl, 55cpl, and 100cpl. The thirty-six participants they assessed were undergraduate and postgraduate students between the ages of 18-44 of which 68% were between 18-25. The majority of the subjects was familiar with computers and used a computer either at work, for leisure, or both. They were given eight articles of similar length from the National Geographic Magazine, each containing up to 1000 words of black text on a white background. The typeface used was Arial at a 10-point type size and a 12-point interlinear spacing (i.e. the space between each line) and additional 12-point spacing between each paragraph. A 0.5 cm margin was maintained on the left side as well as on the right side of the screen. Results showed that a 55cpl line length yielded the best comprehension compared to 25cpl and 100cpl. Furthermore, 25cpl lines were read slower than the longer line lengths. However, 100cpl did not increase reading rate compared to the 55cpl line length. The authors further concluded that comprehension differences were not cancelled out by reading rate, meaning that there was no trade-off observed between speed and accuracy.

Bernard et al. (2002) examined the effect of line length on the reading performance on screen of adults and children by measuring reading time and reading efficiency. The participants in the study were twenty adults with a mean age of 29 and twenty children with a mean age of 11 years all having 20/40 vision or better. The subjects were asked to read a passage in 132cpl, 76cpl, and 45cpl. Each passage consisted of an average number of 1028 words and the topics were psychology related. The text was presented in black on a white background and the

typeface used was Arial at a 12-point size. A Pentium II PC was used with a 17" monitor having a 1024x768 screen resolution and a 60Hz refresh rate. Results showed no significant performance differences between the three conditions either in the adults' group or the children's' group. Furthermore, reading rate did not show any significant differences either between the three conditions. McMullin et al. (2002) examined the effect of line length and white space on participants' comprehension. They assessed fifty-seven undergraduate psychology students (15 male, 42 female) in 115 cpl and 55cpl conditions. In addition, they added two conditions to examine the effect of white space by presenting the 115cpl and 55cpl with a paragraph adjacent to it in a foreign language that filled up the white space. The participants were instructed to read only the passage that was presented in English in these two additional conditions. Eight prose passages were used of approximately two hundred words each, which were adapted so that five multiple-choice comprehension questions could be asked about each passage. The eight passages and accompanying questions were piloted in advance to ensure equal difficulty between them. Each participant was asked to read two passages in each condition. Results showed no significant comprehension differences between the 55cpl and the 115cpl conditions. However, a significant comprehension difference was found between the passages with the added adjacent passage in a foreign language and the ones without with the latter being better comprehended than the former. As white space is discussed separately, these results will be discussed in more detail later on in this section. Shaikh (2005) examined the effect of line length on comprehension, reading speed, and user satisfaction when reading news articles online. He compared four different line length settings; 35cpl, 55cpl, 75cpl, and 95cpl. Twenty college students all having 20/40 vision or better were asked to read a different article in each line length setting with an average length of 375 words per article at a 12.0 grade reading difficulty level.

Chapter 2: Literature Review

The font used for the text was 10-point Arial and a 12-point interline spacing was maintained with an extra spacing between paragraphs. Each article was followed by nineteen comprehension questions, which the students completed directly after they had read the article. All participants were familiar with reading online as they all used the Internet regularly. Results showed that the articles read at 95cpl, were read significantly faster than the shorter line lengths. Although 95cpl was found to be more efficiently read than at 35cpl, no overall effect of line length on comprehension was found between the examined line lengths. The studies reviewed are summarized in table 7 below.

Table 7. Summary of Studies Reviewed on Line Length

<i>Study</i>	<i>Comparisons</i>	<i>Summary Results</i>
Dyson & Kipping (1998)	25cpl, 40cpl, 55cpl, 70cpl, 85cpl, 100cpl line length	-No comprehension differences. -No speed-accuracy trade-off -55cpl perceived easier to read than 100cpl, and 25cpl
Dyson & Haselgrove (2001)	25cpl, 55cpl, and 100cpl line length	-55cpl better comprehension than 25cpl, and 100cpl -55cpl, and 100cpl read faster than 25cpl -No speed-accuracy trade-off
Bernard et al. (2002)	45cpl, 76cpl, 132cpl line length	-No differences in reading rate and comprehension
McMullin et al. (2002)	55cpl, and 115cpl line length	-No comprehension differences
Shaikh (2005)	35cpl, 55cpl, 75cpl, and 95cpl line length	-No overall effect on comprehension -95cpl read faster

As table 7 shows, four out of the five studies reviewed did not show any significant differences in comprehension (Dyson & Kipping, 1998; Bernard et al. 2002; McMullin et al. 2002; Shaikh, 2005). Dyson & Haselgrove's (2001) study that did find a significant difference concluded that the medium line length of 55cpl lead to better comprehension of the text than shorter (i.e. 25cpl) and longer (i.e. 100cpl) line lengths. The authors mentioned that the reason for observing a comprehension difference contrary to Dyson & Kipping's (1998) study was that the comprehension test they used in their study was more elaborate and was therefore argued to be a better reflection of the comprehension construct. However, three more recent studies found no differences in comprehension, which is in contrast to their findings. The reason for the difference could have been that the scrolling patterns associated to better comprehension, which involved taking more time in between scrolling movements and increasing the number of scrolling movements had contributed to this (Dyson & Haselgrove, 2001).

The general conclusion based on the reviewed studies is that comprehension is not significantly affected by changing line lengths with no speed-accuracy trade-off. Some studies indicated that shorter lines are read slower than medium and longer lines (e.g. Dyson & Kipping, 1998; Dyson & Haselgrove, 2001; Shaikh, 2005) and that medium line lengths yield better comprehension (Dyson & Haselgrove, 2001). However, McMullin et al. (2002) who did not find any comprehension differences in their study reasoned that the reason for increased line lengths on screen not yielding comprehension differences in their study was mainly geometrical. They argued that because the distance between the reader and a computer screen when reading text is greater than when reading in print, the distance would automatically increase the visual angle of the reader, which would enable him to cover greater line lengths (assuming 20/20 vision or better for all participants). With this in mind the following recommendations are made for optimal line

length to be used in this study's reading test as a conclusion to this section.

Table 8. Suggested Line Length Settings

<i>Line Length</i>
55cpl - 115cpl

2.15.3 Number of Lines

Dillon (1992) reviewed four studies related to the number of lines on screen, which he categorized as *display size*. The first study by Duchnicky & Kolars (1983) looked at the effect of display size on reading speed and comprehension where subjects had to scroll continuously. They reported that increasing display size to more than four lines showed little gains on the reading speed and comprehension of the subjects in their study. The second study of Elkerton & Williges (1984) looked at different display sizes at the 1-7-13 and 19-line size display. They only found some speed advantages starting from the 7-line sized display.

The third study particularly looked at larger screen displays of twenty or forty lines in size (i.e. Dillon, Richardson and McKnight, 1989). The subjects had to locate specific information using an electronic book. Although they did not observe any performance differences between the two display sizes, they did report an overall preference for the larger display size over the smaller sized. In the fourth study, which involved a 3500-word text, Dillon et al. (1990b) compared display sizes of twenty, and sixty lines. In this study, they found a manipulation effect on the smaller display size, as the subjects manipulated the text more than on the larger display size. They concluded that the most likely explanation for this could have been that the subjects reread the texts or parts of it and skipped through the articles they were to read and therefore needed more manipulations on the smaller screen due to the size difference.

Despite the smaller display size increased the number of manipulations; it did not have any effect on comprehension of the reading passage itself.

De Bruijn et al. (1992) compared two screen sizes, namely 12 inch, which materialized in twenty-three lines per screen vs. 15 inch, which resulted in sixty lines per screen. A total of sixty-five subjects were asked to read a legal sociological discourse presented in both conditions. The text was assessed by summary in addition to multiple-choice questions. The study further looked at layout differences for which two additional conditions were administered to the participants, which are discussed later in the section addressing this variable. The results in de Bruijn et al.'s (1992) study showed neither differences in cognitive effort activated by the text in both conditions nor differences in retaining information. However, they found learning time to be less for the 15" condition (i.e. 60 lines per screen) compared to the 12" condition (i.e. 23 lines). The authors theorised that this could have been the case due to the 15" condition promoting a better integration process when constructing a mental representation of the text. However, no recent studies have further investigated this, which impedes verification of this variable in relation to contemporary technology. Although not stated explicitly, de Bruijn et al. (1992) mentioned that the two presentation conditions differed in terms of resolution, type size, and screen refresh rate. However, the effects of these differences were considered to be minimal by the authors.

The most recent study that compared screen size by comparing the number of lines presented on screen was that of Dyson & Kipping (1998b). They looked at differences between three different sizes, namely fifteen lines per screen, twenty-five lines per screen, and thirty-five lines per screen. Twenty-four students were asked to read a document in each condition of approximately 700 words, which was followed by comprehension questions about the content.

The font they used in their study was sans serif Arial font at a 10-point type size setting. The spacing between the lines was set at 12-points, which is in accordance to what Lynch & Horton (2002) recommended in their study as they suggested the ideal interlinear spacing to be 2-4 points greater than the actual font size. Results showed no significant differences in reading comprehension or reading rate.

In conclusion, the reviewed studies, although mostly performed under conditions using relatively dated computer devices, are in agreement that the number of lines per screen do not significantly affect on-screen text comprehension. The effect this has on the number of lines used in this study is that, in essence, the maximum number of lines a screen can support in the font settings and line lengths discussed previously, would be suitable to be used depending on the text length as it has been shown not to interfere with comprehension. This appears to be in agreement with one of the first studies addressing this variable, i.e. Duchnicky and Kolers (1983), who concluded that whether a full screen of text was read or only four lines on a single screen, efficiency was not affected. However, once the text its length exceeds that which the monitor is able to support, manipulation techniques such as scrolling and paging/page turning become relevant as they could possibly introduce interference with the reading construct and, therefore, comprehension. Hence, these manipulation techniques and their possible impact on comprehension are discussed additionally in the interaction related variables section. Only after that, a more informed decision can be made regarding the application of this variable for the reading test used in this study.

2.15.4 Interlinear Spacing

In addition to line length, and number of lines per screen, the white space between the

lines has been investigated as it is thought that this variable could possibly affect how text is read from screen. This white space is often referred to as *interline spacing* or *leading*. As mentioned earlier, Lynch & Horton (2002) stated that the ideal interline spacing would be 2 points greater than the type size of the text in print and suggested marginally increased generous spacing ranging from 2-4 points greater than the type size for on-screen text. However, this appears to be from a web designer's perspective, as Lynch & Horton do neither refer to any studies corroborating this nor do they clearly motivate why this would be the case. Therefore, verification is needed from studies that examined interlinear spacing of on-screen texts before applying their recommended interlinear settings.

Unfortunately, few studies have addressed this variable and the studies that did date back 10 years or more (i.e. Grabinger, 1993; Kruk & Muter, 1984). Kruk & Muter (1984) compared reading speed between single spacing and double spacing conditions on screen. They asked twenty-four undergraduate students all having at least 20/20 vision to read four sets of short stories, two in booklet form and two from a video screen. The researchers had the subjects read for five minutes per story and registered how far the student had read after the five-minute session. Each reading set was followed by a comprehension test, which lasted likewise five minutes. Results showed no significant effect on comprehension between the single space mode and the double space mode. However, reading speed was affected by interlinear spacing as the authors found that the single spaced text was read 10.9% slower than double-spaced. In response to an earlier study that argued single spacing to be negligible when used with particular displays (i.e. Kolars et al., 1981), the authors suggested that single interlinear spacing should best be avoided when this space is small in relation to the height of the characters of text (i.e. font type/size). This appears to be in line with Lynch & Horton's (2002) argument for having at least

the same size interlinear spacing as the font size of the on-screen text.

Grabinger (1993) included interlinear spacing as one of eight variables in his study as part of evaluating screen designs according to viewer arbitrations. Ninety-four participants were requested to judge samples of text on screen on readability and *studyability* using MDS (multidimensional scaling) as an evaluation tool to analyze paired comparison tasks carried out by the participants. Although results showed that single spacing elicited more positive responses from the majority of participants, the possible effect of single spacing vs. double spacing on reading comprehension has not been assessed. It is therefore difficult to draw any grounded conclusions from this particular study in this regard.

2.15.5 White Space

Bernard, Chaparro & Thomasson (2000) examined the effect of three different white space layouts on search performance, which they divided into low amount, medium amount, and high amount of white space. Each layout consisted of three columns of information with the low amount layout having the least white space between and around the columns and the high amount having the most. They asked sixteen participants to answer five questions in each condition where they had to search for particular information on a web page to answer each question. After the subjects had completed a question they were asked to rate the difficulty in finding the answer to that question (1=very easy, 5= very difficult) until they had completed all 15 questions in the three layouts. After that they were requested to rate their preference for each of the three conditions. Black text on a white background was used with serif Times New Roman being the typeface. The authors did neither provide details on the type size settings of the typeface nor did they provide computer screen details (i.e. screen size, resolution, etc.). Results

showed no influence of white space on the ability to find the information requested on screen to answer the questions correctly. However, the medium white space layout was preferred over the low and the high whitespace layouts. The authors concluded therefore that some whitespace may be better than none, or, too much. However, in order to establish the amount of whitespace that produces optimal results, and whether search performance would be affected when using, for example, multiple webpages, the authors stated that additional research addressing these areas is essential.

McMullin et al.'s (2002) secondary concern in their study in addition to line length (see line length section) was with white space to which they referred as *text density*. Their argument for naming it this way was that line length is confounded to white space i.e. when line length increases, white space decreases and vice versa. The participants were asked to read short prose passages in two cpl conditions i.e. 55 cpl, and 115cpl. Both conditions were either presented in a one-column format with white space adjacent to it or a two-column format with the second column adjacent to (i.e. 55cpl), or below it (i.e. 115cpl) with the second column containing irrelevant information in a foreign language. Comprehension of the participants was assessed by multiple-choice questions given immediately after they had read the passage. Results showed a five percent performance increase for the single column condition compared to the two-column condition where the additional passage was presented adjacent to or below the passage to read. The authors attributed this difference to possible distraction of the subjects by adding the second column. They concluded that although this difference is perceived to be small, it could prove decisive in a high stakes situation when it comes to passing or failing an exam. It is therefore important to consider these results when implementing whitespace into the interface to prevent possible interference in a testing situation.

Chaparro et al. (2004) investigated the effect of white space on reading speed and comprehension by comparing text layouts where margins were used to layouts without margins. They further compared the effect of interline spacing on speed and comprehension by applying optimal leading vs. sub-optimal leading to the texts. This led to the following four layouts; Margins with Optimal Leading, Margins with Sub-Optimal Leading, No Margins with Optimal Leading, and No Margins with Sub-Optimal Leading. Nineteen college students (10 Male, 9 Female) all having normal or corrected vision were asked to read two passages of approximately 800 words of text each taken from a retired SAT exam for each of the four conditions. After they had read two passages of one condition, eight comprehension questions were to be completed followed by a user satisfaction questionnaire. After this, the participants were given a short break before continuing with the second condition. This process was repeated until all four conditions and their accompanying comprehension questions and questionnaires were completed. Results showed an effect of manipulating the margin whitespace on reading speed as well as comprehension. The text containing the margins was read slower but was comprehended better than the text with no margins. Although there was no effect detected of optimal and sub-optimal leading on speed or comprehension, the participants did prefer the Margins with Optimal Leading condition to the other three conditions. Unfortunately, no recent studies addressing this variable seem to have been carried out to provide more insights into this issue. However, researchers did investigate the effect of the number of columns on users' on-screen experience, which is effectively another form of manipulating white space, which is discussed in the section that follows.

2.15.6 Number of Columns

Dyson & Kipping (1997) examined the effect of a three-column text format (i.e. 25cpl) compared to a single column format (i.e. 80cpl) on subjects when reading online text. Their study involved eighteen participants who were asked to read a text of around two thousand words in three different conditions; a single column condition where paging was the means of navigating through the text, a single column condition where scrolling was the means of navigating through the text, and a three-column condition where paging was used as navigation tool. Comprehension questions were presented to the participants after they had read each condition. The typeface used was sans serif Arial at a 10-point type size and the interline spacing was set at 12 points. Results showed that readers that were twenty-five years old or younger read the single column format faster; however, overall comprehension was not affected by the different formats used. Subjective judgments revealed that the participants found the three-column format easier to read compared to the single column format. Dyson & Kipping (1997) theorized that this could be due to a difference in reading patterns employed (particularly by faster readers) when reading a three-column format as opposed to a single column format. However, no research to date to the researcher's knowledge has explored this possible underlying cause, which leaves it difficult to draw any firm conclusions in this regard.

Baker (2005) investigated the effect of one, two, and three-column formats on reading speed, comprehension, and reader satisfaction. Sixty-six undergraduate students with a mean age of 22.8 were asked to read a 2191-word short story with a Flesch grade level of 9.6 in six conditions; the single column format, which had a line length of 90cpl, the two-column format, which had a line length of 45cpl, and the three-column format, which had 30cpl. These three conditions were presented in either full-justification or left-justification, which totals six

conditions.

Results showed that reading speed was significantly faster in the two-column format compared to the one-column format in the full justification condition. The one-column left-justified condition was read significantly faster than either the one-column full-justified condition, or the three-column full-justified condition. However, overall comprehension was not affected by any of the six conditions. This appears to be in agreement with the studies on line length reviewed earlier apart from Dyson & Kipping's (1998) study that suggested that longer line lengths up to 115 cpl, which effectively are single columns, do not affect overall comprehension, which, in turn is of importance to developing the interface for this study. Therefore, the conclusion drawn based on evidence from the results taken from these studies including Dyson & Kipping's (1997) and Baker's (2005) is that because reading rate appears to be faster for longer line lengths (which are effectively single columns), and comprehension is not affected by either single, two, or three-column formats, a single column format is recommended to be used in this study as there is no speed accuracy trade-off in this case.

2.15.7 Text/Background

The colour combination of written text and its accompanying background has been the focus of studies going back to the early 90s with regards to printed texts. One often quoted and well-known study that involved text/background colour combinations in print is that of Tinker & Paterson (1931). They compared ten different text/background colour combinations in their study in order to investigate the most legible colour combination for printed text. They found that black text on a white background was the most legible colour combination. One study that involved the legibility of text/background colour combinations that preceded Tinker & Paterson's (1931) was

that of Du Livre (1912). From this study, the Le Courier legibility table was devised and is often used as a reference point for printed documents. The Le Courier legibility table differed however from Tinker & Paterson's (1931) results, as black text on a yellow background was found to be most legible. Black text on a white background was considered to be the fourth most legible text/background colour combination.

What the findings in both studies do seem to have in common is that higher contrast yields superior legibility. Whether these findings are directly transferable to on-screen text is to be examined first due to the process of perceiving colour being different in the two presentation modes. In print, for example, creating new colours from the three primary colours (i.e. red, yellow, and blue) happens by adding one colour to another, which is known to be an additive process. On the contrary, colours presented on screen are created by mixing different colours of light, and is known to be a subtractive process. The difference is, therefore, that when you mix, for example, red, green, and blue in the additive process for printed colours it results in a colour that is nearly black whereas the same colour combination on screen would create the colour white in the subtractive process (Schaeffer & Bateman, 1996).

Because the two processes are different, it is plausible to assume that this could possibly affect the reader differently. For example, light waves perceived by the human eye are themselves not coloured. It is rather the appropriate receptors within the eye that eventually assign different colours to them (Galitz, 2007). This means that because the ways light waves are transmitted is different between screen and print (i.e. light waves from print are reflected light whereas the computer screen is illuminated light) it could influence the reader's perception of them and, therefore, his experience, especially when it comes to colour and contrast sensitivity as they could be directly affected by any alterations within these light waves (Kuehni, 2005, in his

book on colors, provides a comprehensive discussion on the essence of colour and how the human eye perceives it).

Studies conducted in the field of interface design (e.g. Brown; 1989; Faiola, 1989; Rivlin et al., 1990), typography (e.g. Keyes 1993), and reading on screen (e.g. Legge et al., 1990) in the late 80s and early 90s that discussed text/background colour/contrast on screen appear to be in agreement with Du Livre (1912) and Tinker & Paterson (1931) on printed text as they suggested a maximum contrast for optimal results in their respective studies. Later studies such as Chisholm et al.'s (1999) likewise propose a maximum contrast between text and background. Fulcher (2003) even referred to a maximum contrast between colours as a 'basic rule in colour design' (p. 393) for a user interface. Ling & van Schaik (2002) further added that a maximum contrast between text and background has a facilitating effect when performing search activities on screen.

However, care is to be taken with the sharpness of the contrast between text and background, as Galitz (2007) cautions that a harsh contrast between the two should be avoided when using today's high-resolution monitors. He therefore advised to use black text on a background colour of low intensity such as off-white or light gray instead of white to limit eyestrain and (therefore) possible fatigue on the user. This harsh contrast could very well have been the possible underlying cause for eye fatigue found in many of the more recent comparability studies between CBT and PBT where students had to read from screen for longer time periods. With this in mind, the following text/background colour settings are suggested for the interface used in this study in order to conclude this section.

Table 9. Suggested Text/Background Colour Settings

<i>Text Colour</i>	<i>Background Colour</i>
- Black	- Low Intensity Colours (e.g. off-white, light gray, lemon yellow etc.)

2.16 Review User Interface: Presentation (graphical factors)

2.16.1 Screen Size and Resolution

Ziefle (1998) looked at the effect of screen resolution on visual information processing in a two-experiment study. In the first experiment, she examined the effect of two Cathode Ray Tube (CRT) screen resolution conditions of 832x600 pixels (60 dpi), and 1,664x1,200 pixels (120 dpi), and a paper condition of 255 dpi on accuracy and proofreading speed. A 19” monitor was used to display black text on a white background. Results showed that participants on the paper condition outperformed the two CRT conditions. However, no performance difference was found between the two CRT conditions. In the second experiment, Ziefle (1998) compared the effect of a low-resolution condition of 720x540 pixels (62 dpi) and a high-resolution condition of 1024x768 (89 dpi) on reading performance by measuring eye movements of participants when completing a continuous search task. Fatigue was another variable addressed in this study. Results showed that reaction time and eye fixations increased significantly in the low-resolution condition, which resulted in a decrease in searching speed. Furthermore, an increase in errors was found after well over half an hour into the experiment. Based on these findings, Ziefle (1998) concluded that a high-resolution of at least 90dpi was to be recommended to achieve optimal visual performance on screen.

In a more recent study, Bridgeman et al. (2001) evaluated the possible effects of screen size and screen resolution on a verbal and a mathematics test in three different screen size conditions. The resolution specifications were 640 x 480 for both a 17-inch and a 15-inch screen and 1024 x 768 for the second 17-inch screen. They further examined the possible effect of presentation delay where they introduced a five second interruption between questions in order to imitate slow Internet connection. Despite the differences in resolution, the mathematics scores did not show any significant differences. The verbal tests, however, produced higher scores on the higher resolution, which Bridgeman et al. calculated to be around a quarter of a standard deviation. This appears to support the theory of Baudisch et al. (2003) who mentioned that larger screens and higher resolutions account for deeper immersion resulting in experience enhancement on part of the reader, which, in turn, could potentially affect test results positively. The aforementioned studies seem to contradict Muter & Maurutto's (1991) prognosis (cited in Muter, 1996), as they predicted that a modern computer system with a high-resolution screen, which presents the text in a positive polarity, could in terms of efficiency be equal to reading from paper. However, the two studies reviewed were conducted more than a decade ago and with the rapidly increasing changes in display technology within the past ten years, these conclusions could prove to be different today. However, no recent research to the researcher's knowledge has investigated this variable using current screen technology. Therefore, a conclusion is drawn based on the studies reviewed in this section of which the recommended screen size and screen resolution settings are mentioned below to conclude this section.

Table 10. Suggested Screen Size and Resolution

<i>Screen Size</i>	<i>Resolution</i>
17"	> 90 dpi

2.16.2 Icon & Button Design

Although several works on interface design addressed the use of icons and have made suggestions for the optimal settings for their design, in this study's interface design, no use is made of icons but rather (command) buttons are integrated to facilitate item review and question/item navigation. Therefore, only studies addressing these are reviewed in the following section and used as a guideline to integrating buttons in this study's interface.

Galitz (2007) mentioned three main types of buttons that comprise the interface, namely; *toolbar buttons*, *symbol buttons*, and *command buttons*. Toolbar buttons are generally squarely or rectangular shaped and contain an icon or graphic. Symbol buttons are shaped like the toolbar buttons and contain a symbol instead of a graphic or icon. Command buttons, also known as pushbuttons, are generally rectangular shaped and contain text that indicates the action that is to be taken when clicking on it (e.g. OK, next, previous, cancel, etc.). Command buttons are the type of buttons relevant to this study and will therefore be discussed in more detail below.

Fulcher (2003) argued that using buttons (i.e. navigation buttons) should be limited to an absolute minimum and if used, their location should be in the test-takers view range. This range would be either directly above the text/ items or directly below them. The reason for this is that the flow of a reader generally starts at the top and ends at the bottom of a page (Galitz, 2007).

Lee & Boling (1999) agree and further add that care should be taken with possible inclusion of colours and/or graphics, as they could become a source of distraction and therefore construct

irrelevance (Fulcher, 2003). Furthermore, anticipating cultural background is essential to design features of an interface such as icons and command buttons (e.g. Onibere et al., 2001).

Before this, other studies by Russo & Boor (1993) and Fernandes (1995) had already conceptualized that these aspects could be open to differential interpretation depending on the culture of the user population among other interface related aspects such as text & number format and textual representation. Other user interface researchers that were involved with the usability aspect in globalizing interface software acknowledged the difficulty of applying one specific interface format to different user populations due to cultural diversity (Nielsen, 1990; del Galdo, 1990; del Galdo and Nielsen, 1996).

Based on these indications, Evers (1997) suggested that in future research cultural background and user interface interaction should be investigated and specified in the studies' results. However, this does not mean that there cannot be a uniform set of general guidelines in terms of computer interface layout, as the rule of thumb in HCI (i.e. human computer interaction) encourages uniformity in interface design as the guiding standard. Manovich (2001) mentioned that 'One of the main principles of modern HCI is the consistency principle. It dictates that menus, icons, dialogue boxes and other interface elements should be the same in different applications' (pp. 96). He further mentioned the following about how this applies to the language used within them, 'Most of them contain the same set of interface elements with standard semantics, such as "home," "forward" and "backward" icons (pp. 96)', with slight variations from one application to another (e.g. Microsoft, Apple, IBM interfaces).

We can infer from the previous discussion that although the semantics within the command buttons appear to be standardized to a certain extent in general, the actual language/symbols used on them should be culturally appropriate to the user population by whom it is to be

used as suggested earlier. In addition to the language/symbols on the command buttons recommended to be culturally appropriate, it should also be at the right readability/word difficulty level, particularly in this study, as the participants' English proficiency is of lower-level, which could pose a potential problem to the test-takers interpreting the text on the command buttons used in this study's reading test. Manovich's (2001) example of *forward* and *backward* referring to navigating between applications (i.e. test items in this study) could possibly pose a problem for the test-takers in this study and may have to be simplified in order for them to instinctively interpret the commands correctly to prevent construct irrelevant variance from occurring. The pilot/usability study is expected to tackle these possible problems when integrating the buttons used in this study's interface design. The implications this has for the interface used in this study is that the number of icons/command buttons should be minimized by only including the ones that are necessary, which, in this case, will most likely be two command buttons where students are instructed textually to navigate from one question to another accompanied by two symbols representing arrows for visualization purposes. The third command button included in the interface used in this study enables students to confirm their answers, which is the OK button. By minimizing the number of icons/command buttons it is expected to prevent distracting the test-takers from their main task, which is correctly answering the test items in the reading test.

2.17 Review User Interface: Interaction

2.17.1 Scrolling

CBT's generally present in two ways; either test-takers page through the reading passage (that is, when the text is of such length that it does not fit on one page) by clicking on a button, or

test-takers scroll through a text passage by using a scroll bar (e.g. Piolat et al., 1997; Clough, 2008). Using this scroll bar when reading through a passage is argued to potentially affect reading performance as it could disrupt spatial layout and it could increase the load on working memory (O'Hara & Sellen, 1997; Piolat et al., 1997). Choi et al. (2003) posited that scrolling through a reading passage could have a negative effect on score outcomes and therefore needed further investigation. However, they were reasonably confident that issues related to scrolling (among others) would more than likely be solved in time due to advancement in technology: 'With computer and internet technology growing at an exponential rate, however, these problems may be solved easily' (p.300). However, they did not provide indications or suggestions on how this could possibly manifest in relation to the design of the computer interface.

Studies that investigated the influence of scrolling on reading performance have yielded inconclusive results. For example, Dyson & Kipping (1998), who compared scrolling and page turning on a CBT test in their study, found that reading comprehension was not affected by scrolling. What they did notice was that the paged documents were read faster than the ones that required scrolling in their experiment. Baker (2003) compared page turning and scrolling in his research and found no comprehension differences between the two. However, he did mention that the paged documents were read significantly slower than the scrolled documents, which contradicts earlier findings by Dyson & Kipping (1998). Choi & Tinkler (2002) concluded that scrolling was likely to have had a negative impact on item difficulty on their group of 3rd graders. However, no further research was carried out in order to verify whether scrolling in reality negatively impacts performance. Research conducted by Higgins et al. (2005) three years later compared the differences in score outcomes between a scrolling group, page turning group, and paper-based group taking a reading comprehension test. They did not find any significant

differences in reading comprehension between the three groups, either at the $p < .05$ level or at the $p < .01$ level. Similarly, more recent research conducted by Pommerich (2007) did not find any significant differences either between the scrolling group and page turning group examined in her study.

Unfortunately, none of the previously mentioned studies pinpointed the way in which scrolling could have affected their participants negatively. However, some theories based on reading behaviour could provide further insights into this. For example, one of the possible causes for scrolling negatively influencing test performance could lie in its effect on spatial memory (i.e. the ability to retrieve keywords/phrases/information in a reading text based on the location within it). This is based on earlier studies such as that of Lovelace and Southall (1983) that implied that readers visually establish the location of an item within a text and retrieve this then by memory when needed. Scrolling is then thought to affect this behavior, as it is believed by some theorists to weaken the relationship between the item itself and its location in the text (e.g. Dillon, 1994). This theory could possibly explain the reason for the 3rd graders' poorer performance on computer in Choi & Tinkler's (2002) study, particularly because the items that were found to be more difficult for them required interrelating of keywords in the question stem with keywords within the text, which is facilitated by spatial memory. The proposed weakened relationship between the (question) items and the location of the accompanying keywords caused by scrolling might have affected this negatively. However, to the researcher's knowledge, this has not been further investigated in other studies in support of the application of this theory to scrolling.

This study is expected to shed more light on this by recording the strategies used by test-takers in both testing modes through the think-aloud methodology, which might reveal a

difference in types of strategies used between the two modes due to this possible cause. Nevertheless, the general preliminary assumption with regards to the interface design development based on the previous reviews would be that, if scrolling cannot be avoided, for example, due to the text's length, the scrolling range should be minimized as much as possible as long as it does not cause the settings of the earlier discussed graphical and typographical factors of the interface such as line length, font size, interlinear spacing having to be compromised to the extent that it crosses their own specification boundaries.

2.17.2 Item Review

The ability to go back to review completed items or to skip forward to answer questions that you feel more comfortable with first has been an integral part of paper-based testing. The test-taker only needs to simply turn over to the next or previous page in order to achieve this. By doing so, it provides the opportunity to correct a mistake made earlier on in the test, as Dix (2005) mentions: 'It gives us official permission to reverse up the one-way street after we have taken the wrong turn' (p.40).

This permission or freedom to review items does not always necessarily transfer to testing on a computer. For example, computer-based tests such as the CAT do not allow for item review. Wainer (1993) mentioned that one of the possible reasons for not allowing test-takers to return to previous questions could be that experienced computer users could find ways to cheat the system, which would give them an unfair advantage. He neither specified, however, how this cheating of the system could materialize, nor did he provide any suggestions on how this could be prevented. However, this example is an exception related to the CAT testing method. Many studies that addressed this issue by allowing test-takers to go back to previous questions in order

to review and/or change previously given answers did not find any significant effects on performance between PBT and CBT (e.g. Lunz & Bergstrom, 1994; Zandvliet & Farragher, 1997; Mason et. al, 2001; Poggio et al., 2005). In addition, in line with Dix's (2005) argument, logic would assume that when the default reading test on paper (i.e. PBT) allows the freedom for item review, the same freedom should apply to a reading test taken on computer until proven that it would impede comparability. A similar consideration is found in the APA's (1986) item administration procedures guidelines as they mention 'test takers should be able to verify the answer they have selected and should normally be given the opportunity to change it if they wish' (p.17).

The discussion section of this study confirms this, as a significant number of students found the correct answer to a previously (wrongly) answered item later on in the test, which, consequently, could have had negatively affected the statistical results obtained from the CBT had they not been able to go back to change that previously answered item. Therefore, based on the previously discussed arguments, item review was made possible for the test-takers on the reading test used in this study.

2.17.3 Item Presentation

There are few studies that specifically looked at item presentation on CBT's and the majority of them were carried out around two decades ago. The main focus of these studies was to investigate the effect of items being presented individually on screen or items being presented grouped (e.g. Hofer & Green, 1985; Lee, 1986; Greaud & Green, 1986; Dimock & Cormier, 1991).

The findings in these studies produced mainly inconclusive results. For example, Hofer & Green

(1985) indicated in their study that when items were grouped together on the same page it might lead to rushed responses by the test-takers, which could affect performance. A year later Greaud & Green (1986) who administered a clerical skills test to the participants in their study concluded that grouping test items as opposed to presenting them individually had a facilitative effect, which was also the case for Lee (1986) who administered an arithmetic reasoning test where items were grouped on the CBT versus individually presented items on the PBT. However, as computer experience was the main independent variable in Lee's study and was thought to have had a significant effect on task performance, it weakened the initial argument for the grouping of items, as it should have been investigated in isolation in order to draw firmer conclusions from it. Furthermore, in both studies the type of item presentation was not counter balanced, i.e. grouped items on PBT were compared to individually presented items on CBT or (not *and*) vice versa. It would have been more useful when both had been compared within the same study, for example 25 % CBT grouped, 25% PBT individually, 25% PBT grouped, and 25% CBT individually in order to obtain more reliable results. A later study by Dimock & Cormier (1991), which likewise involved a verbal reasoning test, suggested that individually presented items had a negative effect on performance on computer compared to items presented as a group on the PBT format. However, this study showed the same weakness as the two previously discussed works for similar reasons. In addition, Dimock & Cormier (1991) attempted to 'simulate' the CBT mode by using index cards, which further takes away from the reliability of the results in addition to signifying the datedness of the devices used compared to current practice.

Theoretically, one argument in favour of grouping items could be when a reading text has subsequent items that build on each other, as knowing what is required from the item that follows could help in determining the search focus. However, in this study this would not apply,

Chapter 2: Literature Review

as the items included were stand-alone items and not related to each other in any way. For these reasons, the choice could be made to either group the items or display them one at a time individually depending on the consequences it may have on the other elements of the interface. The interface design section with its reviewed elements following the interface design evaluation model proposed in section 2.8.3 is summarized in section 2.8.6 on the following page and the model that displays the recommended optimal settings for this study's interface according to the reviewed literature is presented on page 94 followed by a summary to conclude this chapter.

2.18 Reviewed Elements and Recommended Settings User Interface

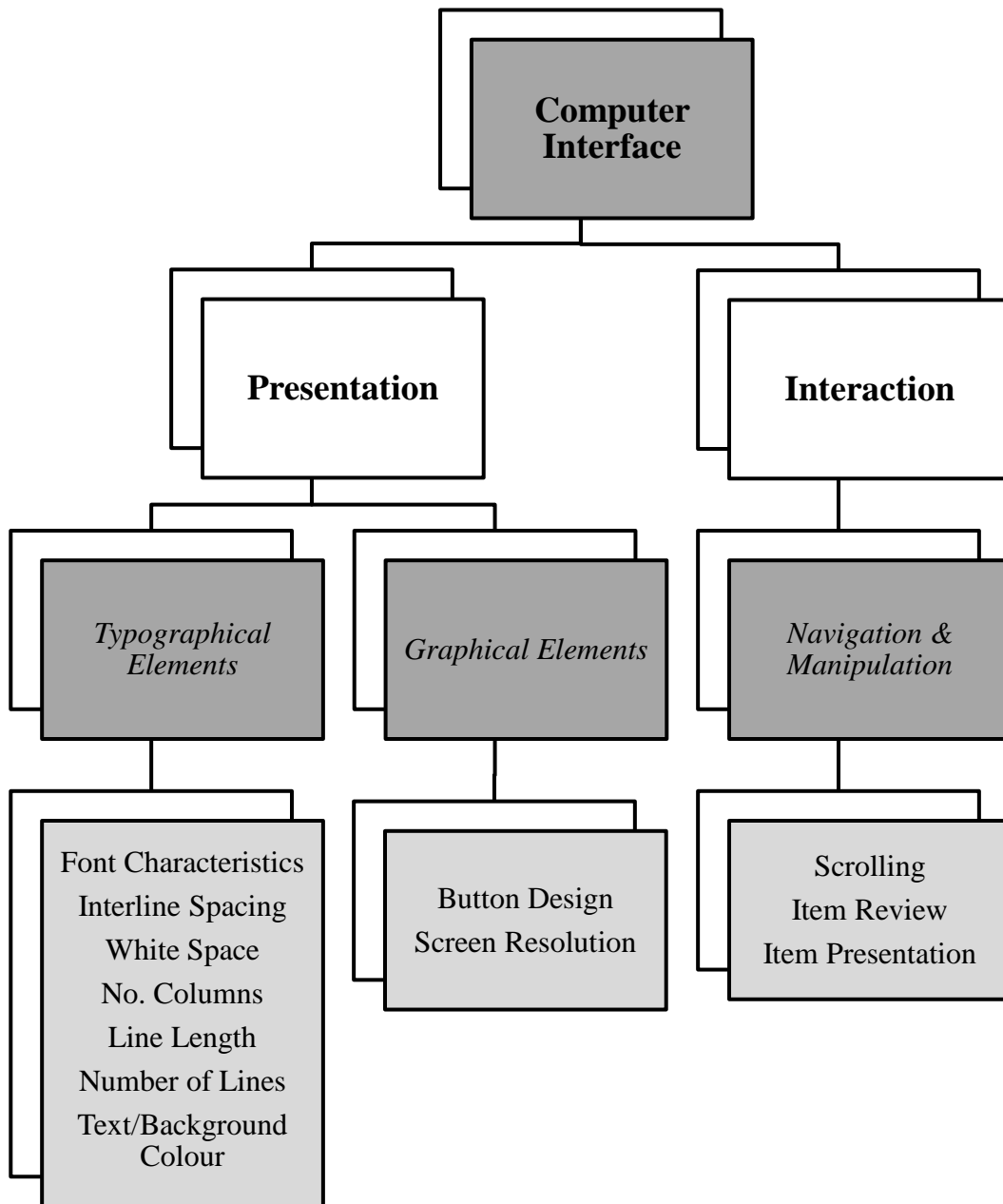


Figure 8. Worked out Interface Design Evaluation Model of Reviewed Elements

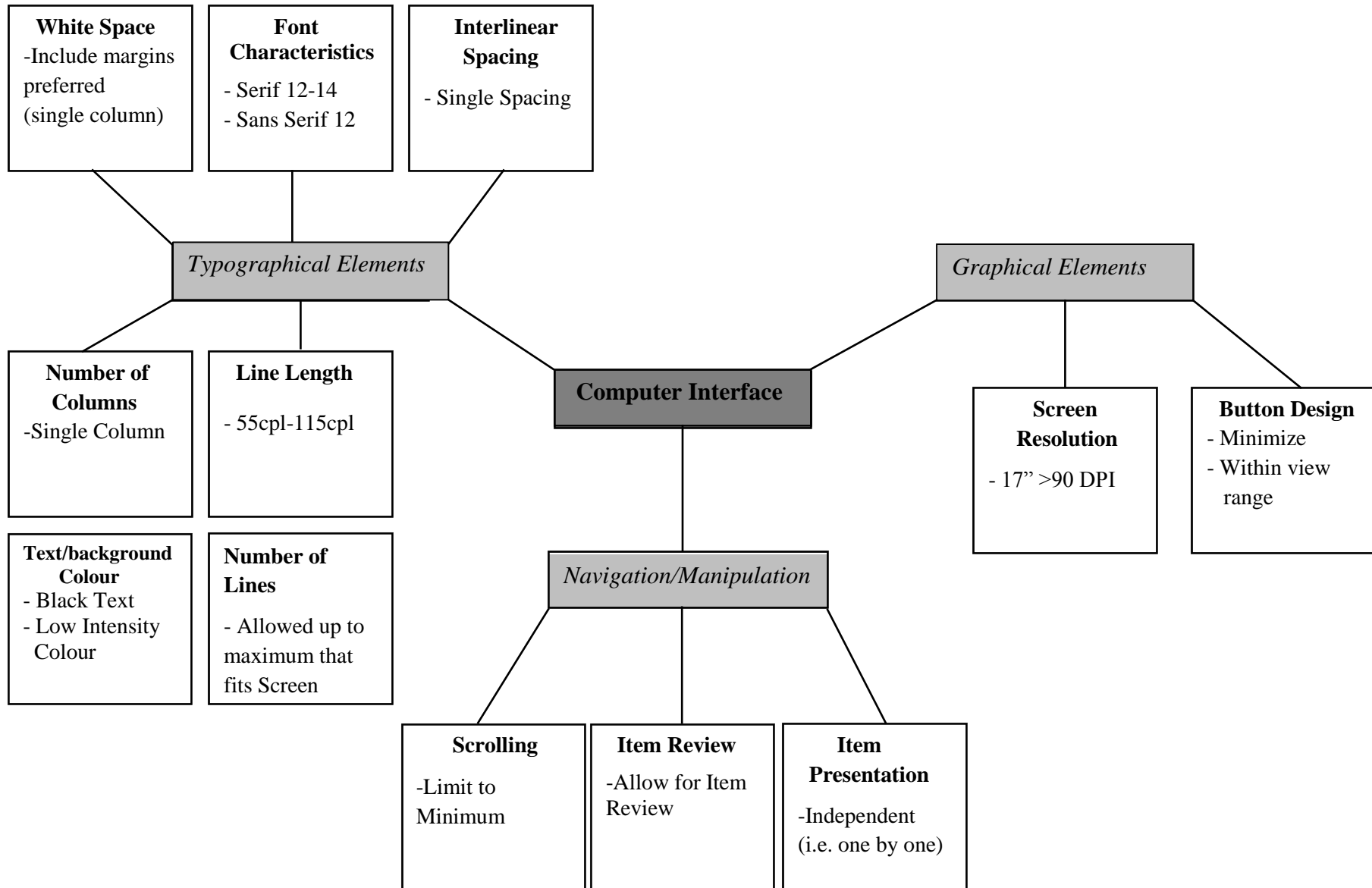


Figure 9. Overview Recommended Settings based on reviewed Literature through Interface Design Evaluation Model

Figure 8 on page 95 summarizes the reviewed element categories of the computer interface each with its included elements which led to the optimal settings for each of these reviewed elements to be integrated into the computer interface in this study's CBT displayed on page 96 (i.e. figure 9).

As for the category typographical elements, the ideal font characteristics were suggested to be 12-14 for serif fonts and 12 for sans-serif fonts. Interline spacing was found to be optimal at single spacing. White space was best using margins ideally held at 0.5/1 inch using single columns. The ideal line length was found to be between 55cpl- 115cpl and the total number of lines per screen was as much as the screen size would allow for, i.e. fit the screen. Black text on a background color of low intensity was found to be the ideal text/background color combination.

With regards to the category graphical elements, the use of buttons was suggested to be kept to an absolute minimum and, when used, should be held within view range of the test-taker. The ideal screen size/resolution was to be held at a 17" screen with a screen resolution greater than 90DPI.

As for the navigation and manipulation category, scrolling should be kept to an absolute minimum, item review should be possible/ allowed for, and items were suggested to be presented individually (i.e. one at a time).

2.19 Summary of the Chapter

This chapter reviewed relevant key areas to the theme of this study, which is the comparability of a reading test's PBT and CBT form, the CBT (its interface) being the independent variable that could affect processes and performance. This literature review began

by introducing the contemporary view of the reading concept where different types of reading are employed according to their underlying purpose. The main source of reference adhered to was Urquhart & Weir's (1998) model, which had been developed from a thorough review of established reading research in the field. It then further elaborated on these reading types by showing how these reading types interrelate with the cognitive processes involved when engaged in a reading activity through the comprehensive cognitive reading model devised by Khalifa & Weir (2009). After that, the transition was made from the reading process itself to the use of test-taking strategies introduced by distinguishing between the two. A brief overview was given on strategies involved in a reading testing context depicted by a two-stage model devised based on the assumed reading types, process-levels, and strategies. This was done to provide insights into the processes the test-takers in this study were likely to employ when taking an L2 reading test and to function as a rough guide for establishing this. Comparing these processes between their PBT and CBT form was expected to provide evidence towards the equivalence of both testing forms (i.e. PBT and CBT) and answer RQ2.

After that, the validity concept was introduced and the contemporary unified interpretation of validity exemplified through Weir's (2005) socio cognitive validity framework for language testing was reviewed, as this study worked within this framework and aimed to provide evidence for the cognitive validity of this study's test as described in this framework. Following this, a concise overview was given about the assessment method chosen for the reading test used in this study (i.e. SAQ's). A review was presented weighing out the pros and cons, and following justified the decision for using them for this study's purpose.

The final section of this chapter reviewed the literature that involved human computer associated issues related to the various components present on screen that could influence test-

Chapter 2: Literature Review

taker behavior (i.e. interface design characteristics). An interface design evaluation model was proposed to be implemented evaluating the different elements of the computer interface and concluded with a detailed worked out model that presented the optimal settings for the factors related to the interface design that could possibly affect test-taker behavior and performance. By implementing this worked out model into the computer interface used for this study's reading test, it was expected to minimize construct irrelevant variance from occurring. The next chapter sets out the methodology chosen and instruments used to collect required data in order to investigate the formulated research questions.

Providing qualitative insights into employed local expeditious reading types by this study's test-takers is expected to address the existing gap in the current literature (see Urquhart & Weir, 1998) as this has not been provided in the reading and language testing literature as of yet. Furthermore, the lower-level cognitive processes and the connection to expeditious reading operations to locate relevant information have likewise been underexplored. Providing these insights is of significant importance to the field of language testing for both educators and language test developers alike as for both of the aforementioned, this could aid in the construction/development test items based on this study's results.

Chapter 3: Research Methodology

3.1 Introduction

This chapter discusses the research methodology employed in this study. It begins with restating the research questions followed by the rationale for the overall research design leading up to the overall framework set out to address the research questions. After that, the data collection model that was devised to address the research questions as comprehensively as possible is presented and explained. Following this, an interface design framework is introduced illustrating suggested stages in the process of developing an interface for a language test and how this study integrated this into its interface development for the L2 reading test used. Then, the pilot study is described utilizing the main instruments, and feedback is generated following which implications for the main study are given. Finally, an overview of the main study is given and comprehensively discusses its data collection procedures, instrumentation, validity checks carried out on the study's test, and the final version of the interface that materialized from the previous piloting/usability testing stage. Screen shots of the interface design are shown at each stage of the development process in order to give an as detailed account as possible of the layout and amendments made at the different stages in the process of moving towards the final product, i.e. the main study's interface.

3.2 Research Questions and Hypotheses

Below are the (restated) RQ's and hypotheses that guided this study's investigation:

RQ1. *What is the effect of administration mode on test-takers' performance when taking a lower-level L2 reading test?*

H₀: *There is no effect of administration mode on test-takers' performance when taking a lower-level L2 reading test.*

RQ2. *What is the effect of administration mode on test-takers' cognitive processes when completing a lower-level L2 reading test?*

H₀: *There is no effect of administration mode on test-takers' cognitive processes when completing a lower-level L2 reading test.*

3.3 Research Design Rationale

The rationale of this study involves the ascription to a research (approach) philosophy, which in essence is a perception of how data about a phenomenon should ideally be used/treated. Epistemology, which includes the various research approach philosophies, refers to what is proven or established to be true (Carson et al., 2001), which is different from doxology, as this is the assumption or belief of something being true without having established it. Science in general is then concerned with establishing/proving (or disproving) what is believed to be true (i.e. from doxology into epistemology). Two main philosophies that are concerned with this in the realm of science are the *positivist* view (i.e. scientific), which views reality as being observable and describable from an objective standpoint (e.g. Levin, 1988; Lin, 1998), and the *interpretivist* view, which posits that a full understanding of reality is achieved through

subjectively interpreting phenomena instead of objectively and is also known as anti-positivist (Galliers, 1991).

Positivism emphasizes on the isolation of phenomena and their repeatability, which includes manipulation of reality by, for example, altering the independent variable in order to identify relationships, regularities, cause and effect, etc. Conclusions are drawn and/or predictions are then made based on the observed realities. An essential element in interpretivism is the investigation of phenomena in their natural environment, which, however, concedes the influence of the researcher on that environment. Each interpretation of the reality in interpretivism is considered a potential contribution to the new knowledge sought after (e.g. Hudson and Ozanne, 1988).

Both aforementioned views would suggest that the positivist view relates more to the quantitative aspect of seeking knowledge whereas the interpretivist view would likely relate more to the qualitative aspect of it. However, as Lin (1998) argues, it is possible for qualitative work to be positivist in essence too. One of the examples she gave was when practices were expected to lead to a certain set of outcomes, as with identifying strategic patterns across different venues (i.e. data collection instances) with different participants (or the same, depending on the research problem). An example of qualitative work from an interpretivist perspective is gaining the understanding of abstract concepts such as ‘poverty or race’ (Lin, 1988, p.162) through eliciting various explanations for them whether they are conscious or subconscious in nature.

This research falls within the overall view of positivism as it seeks to obtain observable and measurable knowledge from an objective standpoint (i.e. RQ1) and further aims to describe (observable) cognitive processes/strategies (RQ2) leading up to identifiable sets of outcomes on

both occasions. The section that follows discusses the methodology chosen in order to address the research questions as optimal as possible.

3.4 Frameworks and Design

This study consists of a qualitative as well as a quantitative element reflected in the main research objective, which is to investigate the effect of interface design on test-takers' performance (quantitative: RQ1) and its effect on test-takers' cognitive processes (qualitative: RQ2). In order to execute this, a mixed method approach in collecting and analyzing data was employed in order to address the research problem as adequately as possible.

Dörnyei (2007) stresses the importance of good research design and the potentially rich data it can generate in order to understand 'even subtle meanings in the phenomenon under focus' (p.127). A mixed method approach is often used when a research problem is to be viewed from different angles in order to understand it optimally. Johnson et al. (2007) define mixed method research as 'an approach to knowledge (theory and practice) that attempts to consider multiple viewpoints, perspectives, positions, and standpoints (always including the standpoints of qualitative and quantitative research)' (p.113). More specifically, Dörnyei (2007) describes a mixed method study in the following way: 'A mixed method study involves the collection or analysis of both quantitative and qualitative data in a single study with some attempts to integrate the two approaches at one or more stages of the research process' (p.163).

As mixed method research allows for collection of both qualitative and quantitative data as indicated, it is evident that this is reflected subsequently in the inferences made, which could then be of both qualitative and quantitative nature. The weight the qualitative and quantitative aspects of the mixed method approach have in a study varies according to the objectives of that

particular study (e.g. Creswell, 2003; Creswell & Clarke, 2007; Dörnyei, 2007; Johnson et al. 2007). A general framework was set out where these foundational principles of mixed method research were incorporated and adapted accordingly to address the research problem in this study, which is expanded on later in this chapter illustrating the data collection methods utilized to achieve these research objectives. The overall framework for this study's research design is outlined in figure 10 below.

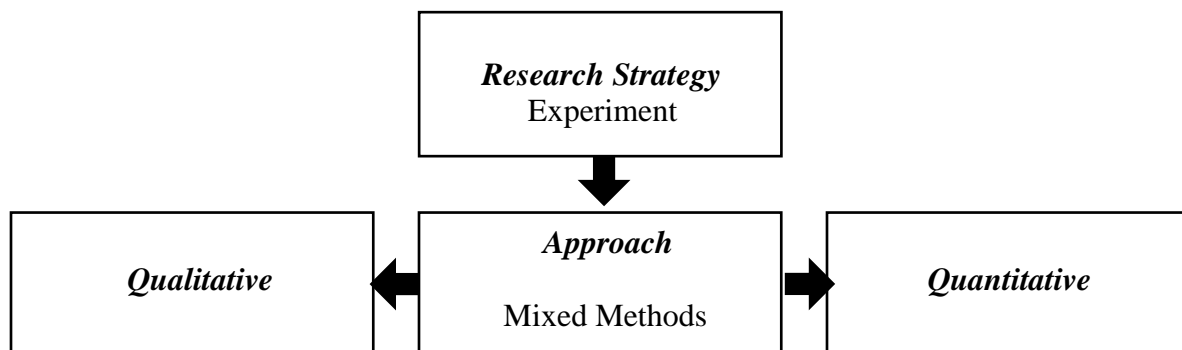


Figure 10. Research design.

As shown in figure 10 above, an experiment was chosen as this study's main research strategy, of which two of the main characteristics as mentioned by Denscombe (2000) are:

1. Controls. Manipulation of circumstances is put forward as the main characteristic (i.e. to investigate how subjects respond when the mode of testing is altered).
2. Identification of causal factors. Introduction or exclusion of factors to or from a certain situation (e.g. context) is identified as a possible influent on outcomes (i.e. the CBT included in this study as its independent variable).

Due to this study's aim, i.e. comparing test-takers' performance (RQ1) and test-takers' cognitive processes (RQ2) on the same test on two different occasions in two different modes of

testing, it is inevitable to change (manipulate) the context, as the requirement for having two different modes is the main constituent of the experiment (i.e. a CBT mode in addition to the PBT mode). Furthermore, because an experiment allows for data collection to be of quantitative origin as well as qualitative (Denscombe, 2000; Creswell, 2003, Creswell & Clarke, 2007) it enables the mixed method approach mentioned by Johnson et al. (2007) and Dörnyei (2007) to be used for generating data in this study. Consequently, a model was developed within the mixed method approach to address the stages, types of analyses chosen, and instrumentation used at each stage investigating this study's research questions. The model is presented in figure 10 below and is subsequently discussed.

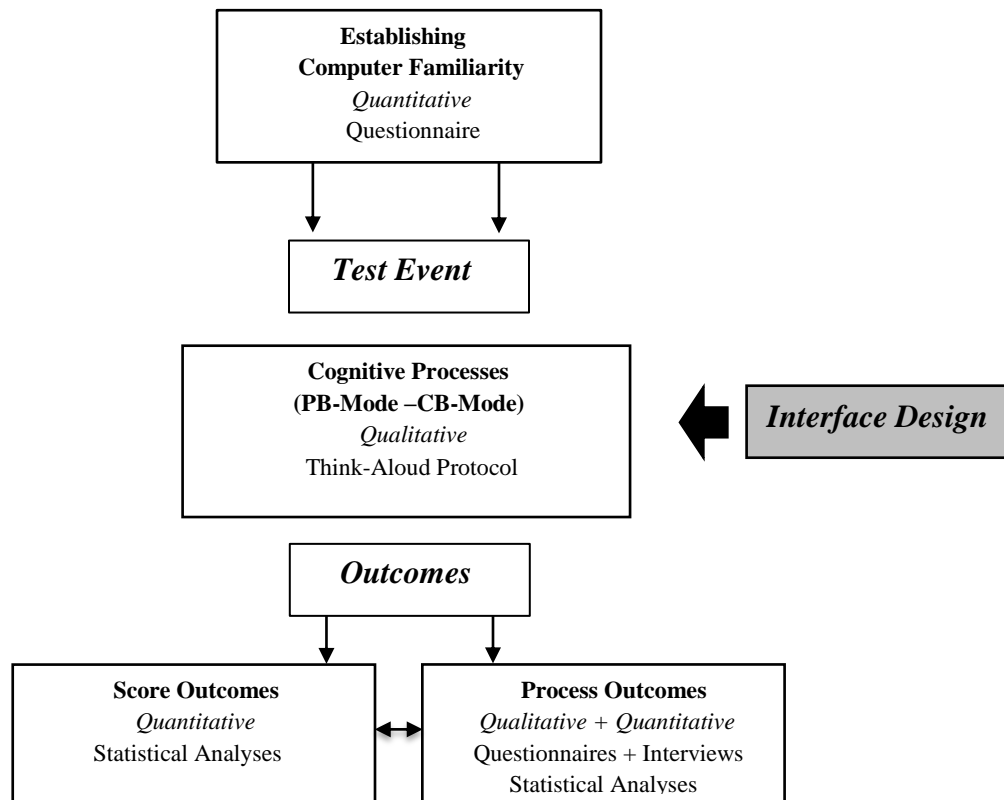


Figure 11. Devised Data Collection model.

As indicated in the top box in figure 11, establishing test-takers' computer familiarity was an essential first step to this study's main objective i.e., examining the effect of interface design on test-takers' performance and cognitive processes because students not being comfortable with using computers could have caused construct irrelevant variance from occurring through problems on the operational side (i.e. working/interacting with the computer itself). The box at the top signifies that, before the test event, sufficient computer familiarity was established for all participants. A computer familiarity questionnaire (henceforth, CFQ) previously validated and administered on a large scale in several studies (e.g. Eignor et al., 1998; Kirsch et al., 1998; Weir et al., 2007) was used to elicit information about the students' computer familiarity (see Appendix B and C for both a copy of the English version and Arabic version of the CFQ used). Each participant was given a factor code based on the mean of the total response scores embodying his familiarity ranging from 0-5. The higher the factor, the more familiar the test-taker was with computers. Three theoretical categories of familiarity established through Eignor et al.'s (1998) validated computer familiarity measure were referred to for deciding which test-takers would be included in the main study. These were low (CFQ-score from 0-2), moderate (CFQ-score from 2-3), and high (CFQ-score from 3-5). Any test-taker in the low familiarity range (i.e. a CFQ-score below 2) would not be included in the main study (section 3.8.3 discusses the selection, and process of validating the CFQ). The test event (i.e. the process of taking the test) was examined through recordings of test-takers' think-aloud reporting (or TA-reporting) whilst taking the test on the two testing occasions, one recording for each test-taker in each mode (i.e. PBT and CBT). The interface design box portrayed to the right of the mid-box reflects this study's independent variable as a possible influent on these processes in **CBT**. The verbalizations in both modes were then segmented and coded to enable frequency comparisons

between the CBT and PBT-mode, which would enable identification of any significant differences in test-takers' processes between the two modes, as shown in the bottom right box. Furthermore, post-test interviews were conducted as a supplementary instrument to attempt to help further interpret this data to illustrate any possible underlying reasons for observed cognitive processes differences between the CBT and PBT-modes that the recordings might have failed to identify, if found. Dörnyei (2007) refers to this type of mixed method approach as an experiment with parallel interviews. The score results were compared to see whether the newly introduced testing mode (i.e. CBT) significantly affected test-taker performance, as shown in the bottom left box supported by a post-test questionnaire (henceforth, PTQ) gauging overall experience with the CBT in comparison with PBT for illustration purposes. By employing this approach to investigate process and product, it was expected to gain a more in-depth understanding of the research problem under focus.

3.4 Interface Design

As mentioned in the introductory chapter, several researchers indicated that a poorly designed computer interface could be a serious threat to the construct validity of a language test (Choi et al., 2003; Pommerich, 2004; Higgins et al., 2005; Paek, 2005). By the same token, none have (yet) developed a framework or model that addresses these design problems or given a set of general guidelines to which one could adhere in order to develop a computer interface for language tests.

Fulcher (2003), however, has made a significant contribution towards understanding the process of developing a computer interface in a language-testing context. He devised a framework for designing a computer interface for language tests, which he said to have adapted

mainly from the available literature on interface design in the software industry. The stages of interface design that he set out provided a useful framework of reference when developing the interface for this study. Fulcher's (ibid) design process consisted of the following three main phases:

1. The planning and initial design phase.
2. The usability testing /or rapid reiteration phase
3. The field-testing & fine tuning phase.

As researching the effect of interface design on test takers is the overall objective in this study, an interface was developed based on a synthesis of related research on what has been established in the literature as good interface design from different areas of study (e.g. Dillon, 1992; Muter, 1996; Fulcher, 2003; Leeson, 2006). This was done to minimize the possible effect of the interface design itself on the test taker in terms of human computer interaction, which could affect the constructs measured. Fulcher's (2003) work was particularly useful to this study in terms of the reiterative process of developing the interface, which was therefore incorporated into this research and adapted to serve its objectives. A summary of the three phases set out by Fulcher (2003) is outlined in figure 12 below and subsequently discussed.

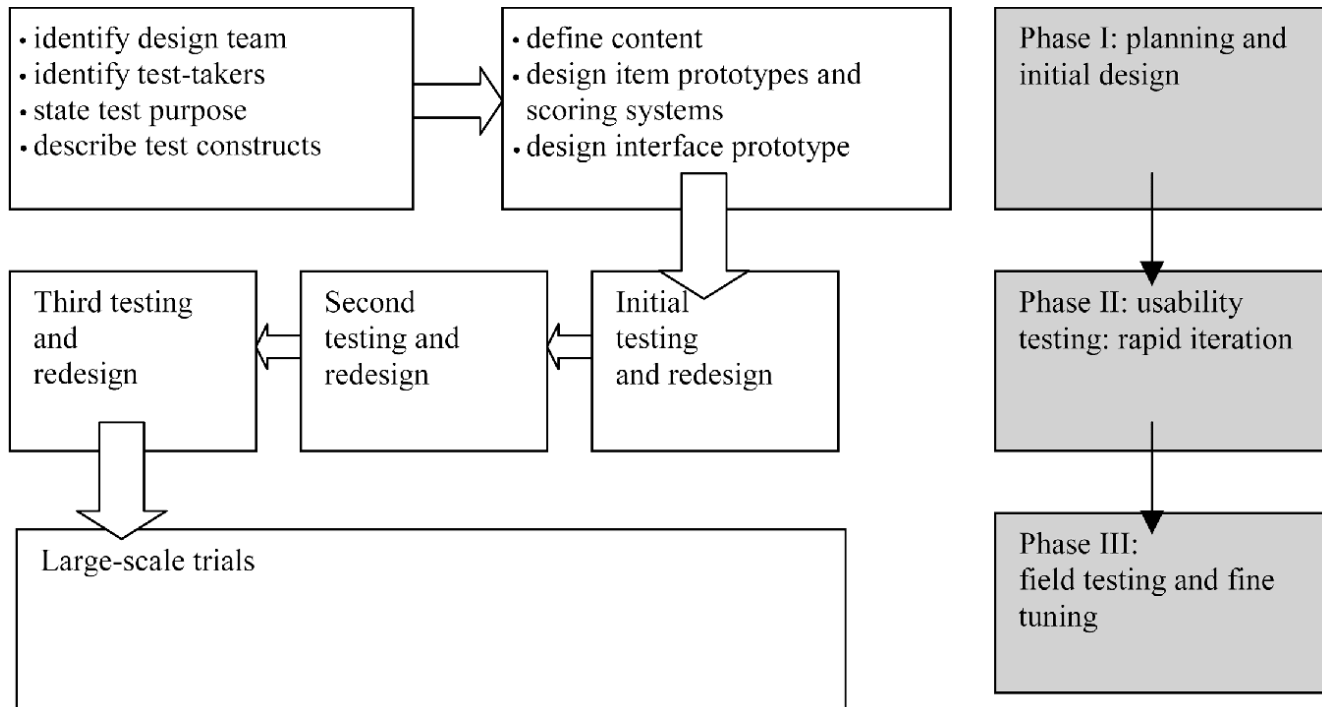


Figure 12. Essential components of a CBT interface design process (Fulcher, 2003).

As illustrated in figure 12, phase one in Fulcher’s (2003) process is two-fold; the planning stage and the initial design stage, which together consist of seven different elements summarized below.

The first stage of phase one is the planning stage (top left box), which is the stage where the design team is assembled, test-takers are identified, the test’s purpose is determined (high-stakes or low-stakes, placement test, final exam etc.) and the test’s constructs are described (comprehension, achievement, etc.). Fulcher (2003) did not describe these initial processes in his work as his main concern was with the development of the interface only. Likewise, these processes are not discussed here but are separately addressed in the main study section later on in this chapter, as they are not part of the operational part of the designing of the interface itself. This leaves *designing the interface prototype* as the main process to be discussed here in phase

one, as it is the only process directly involving interface development. The interface prototype can be considered as a preliminary version of the final product and generally only contains a small number of examples of possible item types to be used for the final product. The reason for this is that it allows for usability testing of the interface in its initial stages using relatively little funds in order to ensure interface appropriateness without aid from substantial human/financial resources (Fulcher, 2003). Once phase one is completed and the interface prototype has been found to be suitable, usability testing is commenced with in phase two where the interface is trialed. This process is reiterative where after each trial, feedback on the usability is generated and amendments are made accordingly for three consecutive intervals. Then, the interface design in its finalized form is subjected to large-scale trials in phase three, as the main issues with the interface have been addressed in phase two, which only leaves possible minor amendments to be made for fine-tuning in the final phase (if found).

In order to make use of Fulcher's guidelines optimally, I attempted to find a way to integrate the three phases into this study despite the various limitations of this research in terms of time, finances, availability of IT- technicians/ design experts. In order to achieve this, I tried to coincide the three phases of the interface design with the phases in this study to have them run parallel and therefore limiting possible delays or other problems as much as possible. It was inevitable to leave out a number of aspects of the phases, as designing an interface in the way described by Fulcher is exceedingly comprehensive as pointed out earlier. The result is shown in figure 13 below where the processes involved when designing the computer interface for this study and how it is interwoven with Fulcher's (2003) work are shown in order to serve the purposes of this study.

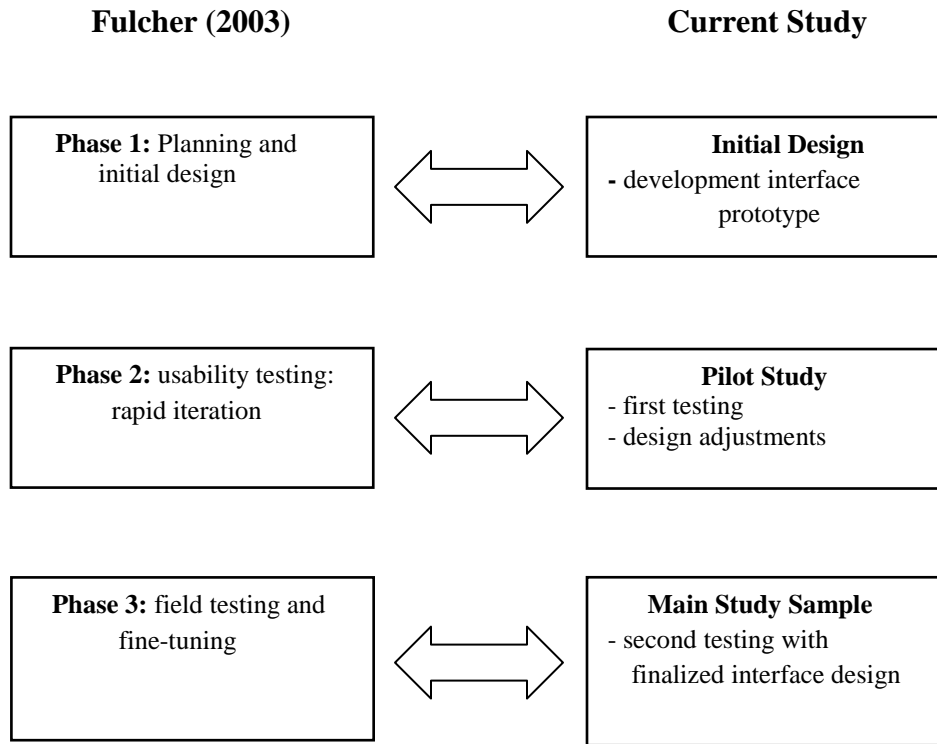


Figure 13. This study's interface design process adapted from Fulcher (2003).

Unlike Fulcher's design process, which assumes that an interface is developed from scratch, the hot potatoes software was used to aid in developing the interface for this study. This was done anticipating the absence of IT- technicians and design experts due to budgetary limitations on the researcher's part. This directly influenced the mechanics in phase 1, as the prototype was *further* developed from an *existing* template from the hot potatoes software instead of constructing the prototype from scratch. This template was then adapted, and text and test items were integrated. After that, the prototype was piloted to a small number of students in phase 2 where issues related to using the interface design in the target context were identified to be amended in preparation for the main study. Furthermore, feedback on interface/test usability was elicited from test-takers in the form of a questionnaire with open-ended items to gain

practical insights into problems related to the usability of the computer interface when completing the reading test at the piloting stage (Nielsen, 1990b). The design was then further adapted accordingly and the test was then administered on a large scale (i.e. main study) in its finalized form in phase 3. The pilot study is further discussed in the following section.

3.5 Pilot Study

3.5.1 Objectives

Pilot studies are widely used to gain preliminary insights in a variety of branches of research. Van Teijlingen & Hundley (2002) distinguished between two types of pilot studies; for the first type they cite Polit et al. (2001) who mentioned that ‘It can refer to so-called feasibility studies which are "small scale version[s], or trial run[s], done in preparation for the major study’(p. 1).

For the second type of pilot study they cite Baker & Risley (1994) who related that, ‘a pilot study can also be the pre-testing or 'trying out' of a particular research instrument’ (p.1). In this study a pilot study is carried out complying with interface design phase 2 as suggested by Fulcher (2003). A retired institutional test was obtained from the ELC and used for the first pilot study, as the main study tests had not been made available at that time. This was not a major issue at this stage as the main aim of this pilot study was to initially trial the interface design and simultaneously trial the study’s instruments. Its main objectives were as follows:

1. To initially trial the CFQ in the target context aiming to identify any problems with the question items in the questionnaire.
2. To obtain preliminary insights into test-taker behaviour taking the CBT and PBT.

3. To pilot the developed interface (Fulcher, 2003) in preparation for the main study and evaluate it from a usability aspect as indicated by Nielsen (1990a).
4. To trial the recording devices.

3.5.2 Participants

The participants were ten students enrolled in the Preparatory Year Program and were similar in educational background, cultural background, and English language level to the students who participated in the main study. They were taught the same syllabi throughout their education starting from primary school up to university (see section 1.4), were all from the same province, and shared the same cultural background. All students had to take the same placement test in order to ensure sufficient language proficiency in order to study in the preparatory year program. Ten students from the total intake of the student population studying in the Preparatory Year Program were sampled conveniently (i.e. from the same class) for the pilot study.

3.5.3 Instruments Pilot Study

The three main instruments trialed in the first pilot study were the CFQ, the reading test (i.e. in PBT and CBT), and the introspective think-aloud protocols. A brief description of each instrument and how it was used in the first pilot study is given below.

Instrument 1: Computer familiarity Questionnaire.

The CFQ was administered to the students in order to get an understanding of the subjects' level of familiarity with computers. The questionnaire was presented to the students in English and caused some problems at the operational side, which is discussed later on. A full account on the development of this questionnaire is given in section 3.8.3.1.

Instrument 2: Reading Tests.

The reading tests consisted of 3 short passages; each passage had five accompanying questions. The PBT and the computer-based test were administered subsequently without interruptions. The reason for this is that the tests for the main study were not available at that point, which made the researcher use retired institutional tests. This unexpected change shifted the focus from test content *and* interface design to interface design only.

Instrument 3: Introspective Think Aloud Protocol.

In order to measure the cognitive processes of the students when completing the tests, a think-aloud protocol was used throughout the test event (further discussed in section 3.8.5 in this chapter). The think-aloud reporting in the paper-based test was recorded with an mp3 player with voice recording facility. As for the think-aloud reporting in the computer-based test, this was recorded through a screen capture software program, which recorded audio and actions performed on screen simultaneously. Due to continuous problems encountered with the screen capture software, it was not used in the main study. Instead, the students were observed and notes were taken in addition to the voice recordings. Another reason for this change was that by using observations it created the possibility to gather detailed information on test-takers' behaviour other than what they verbalized in *both* testing modes as opposed to only the CBT-mode when using the screen capture software, which contributes to enriching the data collected.

3.5.4 Interface Design Pilot Study

Hot Potatoes, which is a product of half-baked software (Half-baked, 2004), was chosen for further developing the interface for this study's CBT as it was one of the few programs that was well-regarded and free to use at the same time (e.g. Chapelle and Douglas, 2006).

Chapter3: Research Methodology

Furthermore, the researcher was already familiar with using hot potatoes as he had successfully used it in an earlier study for similar purposes with a smaller sample where test-takers were likewise assessed on computer and on paper (i.e. Korevaar, 2008). Therefore, the decision was made based on the following three preceding premises:

1. Substantiation from the field: i.e. well regarded by researchers in the field of langue learning and assessment (Chapelle and Douglas, 2006).
2. Previous experience in similar context: i.e. its appropriateness for this study's purpose proven practically through an earlier study, which proved that it contained the core basic features needed to develop the computer interface for this particular purpose.
3. Convenience: i.e. it was free, easy accessible, and easy to use.

The first small-scale trial functioned as an initial usability test in order to work with the interface prototype using a small group before moving to Fulcher's (2003) field-testing/fine tuning phase in the main study after the amendments made based on the feedback received.

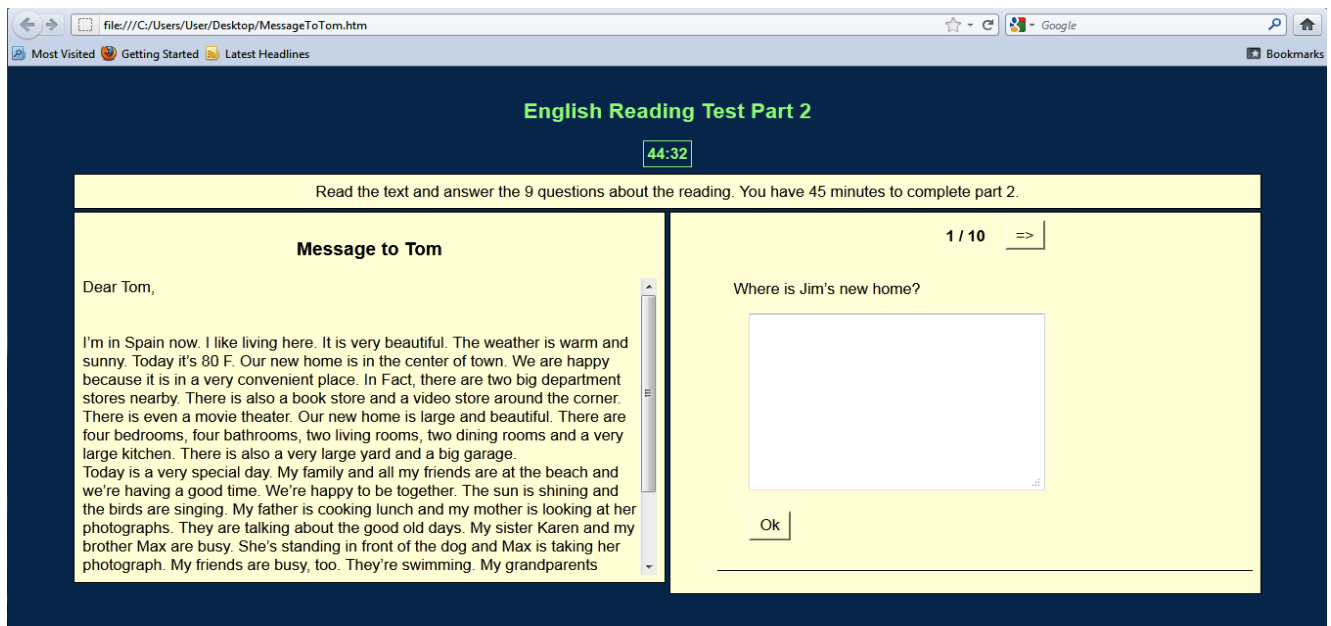


Figure 14. Screenshot Computer Interface Pilot Study

Figure 14 above is a screenshot of the interface that was developed based on the reviewed literature (as discussed in section 2.8). Based on this, initial amendments were made to the basic template provided by hot potatoes in order to improve the interface's usability for this study's purpose. Firstly, the screen contained an index button, which was linked to an empty page when clicking on it (i.e. outside the test) and had to be removed, as it cause construct irrelevant variance by creating possible anxiety on part of the test-taker.

Secondly, the word *check* on the button below the text-box would most likely have caused uncertainties among a number of the participants by not being familiar with the word in connection to its function. Furthermore, in a reading test (exam) students generally do not have the option to check their answers, particularly in high-stakes situations. Therefore, maintaining the *check* option in the reading test could have affected the way students approach the exam (i.e. cognitive processes), which in turn, could have had an effect on the validity of the study results. To eliminate this possible construct-irrelevant measure, the word *check* was replaced by the abbreviation, *OK*, which represents a confirmation of their answer given instead of giving the participants the idea that they would get something in return (i.e. an indication whether the answer was correct).

Thirdly, the scroll- bar was relocated from where scrolling coverage included the whole page to where it only included the passage's text in order to further minimize the scrolling range and to prevent other features of the interface from being scrolled outside of the viewing range of the test-taker while taking the test. Colour contrasting combinations were changed too according to the recommended settings discussed in section 2.7. Important to note is that the tests for the main study were not available at this time. Therefore, a full account of the interface design

features linked back to the worked out interface evaluation model on page 95, will be given in the interface design section of the main study (section 3.8.1).

3.5.5 Procedure Pilot Study

Initially, all students were instructed about the complete procedure in their L1. It was mentioned to them that they were to do a reading test and that before commencing the test they had to complete a questionnaire about their experience with computers. Before commencing with the study tests, all students completed the computer familiarity questionnaire. After the test event feedback was generated on any possible problems related to the CFQ in terms of interpretation of the questions. During the first session, five students began with the computer-based test and five started with the paper-based test. Three of the five students who did the computer-based test first did this while thinking aloud. One of the five students who started with the paper-based test thought aloud while completing the test, which brings the total to four TA- protocols. After all students had completed both testing modes (i.e. PB and CB), five students were given a usability questionnaire in order to report on any problems they encountered with the interface when doing the pilot study's test (Nielsen, 1990). Table 11 below shows the procedure for the pilot study.

Table 11. Design Pilot Study

Before Test Event				
CFQ (L1) & Instructions (L1)				
	3	2	1	4
Session 1	CBT (with TA)	CBT	PBT (with TA)	PBT
	5		5	
Session 2	PBT Usability Q. (Interview L1)		CBT Usability Q. (Interview L1)	

As shown in table 11 above, three out of the five students who did the CBT first in session one verbalized their thoughts whereas two did not. Of the five students that did the PBT first in session one, only one verbalized his thoughts whereas the remaining four did not. The second session was the same for both modes, i.e. none verbalized their thoughts when taking the test. As indicated, the intention initially was to interview the recorded students after their think-aloud sessions in order to get an initial sense of the reasoning behind their behaviour whilst thinking aloud. However, due to time constraints on the students' part, it was not possible to do so. Nevertheless, this will be covered in the main study when the number of subjects available will be greater in addition to the actual tests for the main study being available by then and will therefore provide more meaningful data in this regard. What I did try to gauge shortly after their second session was whether the students were able to accurately recollect what they had done during their first session in order to see whether to interview them after each session in the think-aloud study or after both sessions were completed by the test-taker. This check proved to have been essential as it turned out that it was difficult for students to recollect what exactly they had

done during their first session when asked after both sessions had been completed. For this reason, in the main study, interviews were done after each think-aloud session in order to maximize the potential of the interviews and the validity and reliability of the inferences made from the interview data.

3.6 Results Pilot Study

Due to an unexpected setback on the operational part from the University's side, the reading test that was meant for the main study had not been made available at the time of drawing the sample for the pilot study. In anticipation, the researcher (ad-hoc) selected reading tests that had been used in previous years and were readily available. No gaps were observed between the two sessions, as the main purpose of this pilot study was only to trial the instruments and acquire feedback on the usability of the interface. The results from the CFQ and usability questionnaires were promising and are further discussed below.

3.6.1 Computer Familiarity Questionnaire Results

Although the CFQ has been developed, validated, and administered in various contexts to large sample populations of various age categories (Eignor et al., 1998; Kirsch et al., 1998, Weir et al., 2007; Korevaar, 2008), the main concern was not whether the questionnaire accurately measured computer familiarity. Rather, its usability in the target context was of more significant importance, which, as further discussed in this section, revealed a number of problems that were amended in preparation for the main study. Section 3.8.3.1 further discusses the CFQ's reliability figures based on the main study's sample (n=102). Piloting the CFQ proved to be essential in

preparation for the main study as it highlighted issues that would otherwise have not been found before collecting data for the main study.

The first problem encountered by the majority of the students (9 out of 10) was related to Q1, items a,b,c,d and Q4, items a,b,c,d (appendix B/C). Items a-d of Q1 were related to the following question:

...How often is there a computer available to you to use at these places?

Items a-d of Q4 were related to the following question:

...How often do you use a computer at these places?

Item a of Q1 was parallel to item a of Q4, item b of Q1 to item b of Q4, item c of Q1 to item c of Q4 and item d of Q1 to item d of Q4. The majority of the participants reported that they perceived these two questions to be identical despite the fact that there was a clear distinction between the two in the questions. For example, looking at the main difference between the two questions above, it is clear that the former mentions *availability to* whereas the latter clearly focuses on *usage of*. This difference was accordingly made in the translated questionnaire as *لك توفر* referred to availability whereas *تستخدم* referred to usage. Apart from their different word classes (i.e. the former is an adjective whereas the latter is a verb) both are evidently two different entities, which should theoretically have been detected visually. The feedback received from the students revealed that they initially skimmed through the items related to the question quickly and subsequently read the question quickly. When they realized that the answer options were identical to what they had seen before in question one (item 1-4), they assumed that the question was identical to the formerly mentioned and therefore did not reread the actual question thoroughly but started answering the questions immediately.

Another issue was found with Q3, which was related to the following question:

...How would you rate your ability to use a computer compared to your peers?

The main issue the students had with this question was that the ‘peers’ with whom they were in class were only their peers for the English classes. It was therefore difficult for them to compare their own computer ability with their peers’, as they generally had not studied together before.

A further issue was found with item e of Q6 related to the following question:

...How often do you use each of the following kinds of computer software?

In one of the related options, the statistical analysis package SPSS was mentioned and none of the students knew what SPSS actually was/meant, which resulted in the majority leaving this item unanswered. However, what became apparent in the interviews when asked about this was that they did have experience with statistical analysis software although not SPSS in particular. The amendments made following the issues identified are discussed in section 3.7.1, which discusses the implications for the CFQ in the main study derived from this pilot study.

3.6.2 Usability Questionnaire

A usability questionnaire was devised in order to elicit information on different aspects of the user interface. Nielsen (1990) recommended that in order to identify the main problems with a user interface, a total of 5 participants should be given a usability questionnaire. Nielsen’s (1990) work was used as a foundation for the questions included in this questionnaire and adapted to fit its purpose. The selected questions that addressed the relevant usability elements in this study are discussed below.

1. Visibility of system status.

The following question was used to elicit information about the visibility of system status (i.e. whether the students were aware of what was happening on the screen at all times):

...Have you ever had the feeling that you did not know what was happening on the screen at a certain time during the test?

2. Match between the system and the real world & Consistency.

This question aimed to capture the appropriateness of matching between the system and the real world (meaning that words, phrases, concepts, language used, etc. are familiar to the test takers). Indirectly, it also checked whether the instructions and language used were consistent:

...Were you familiar with the language used throughout the test by the computer system (instructions, buttons, etc.)?

3. User control and freedom

In order to address whether the student felt he was in control and could freely move through the test, the following question was asked:

...Have you ever had the feeling that you were stuck in the system's interface and could not get out during the test?

7. Minimalist design

To verify whether any (to the examinee) irrelevant information was present within the test, the question below was used:

...Did you find any information that you thought was not really needed or you could do without?

8. Remaining Questions

To find out whether any problems were encountered in relation to navigating through the test, manipulating the text by scrolling or using buttons to move back and forth through the questions, the following general question was asked:

...If you have had any difficulties during the test, please describe them in detail below:

And more specifically:

...Did you have any difficulties in using the buttons presented on screen during the test?

In order to address the recognition rather than recall heuristic, the test instructions were continuously visible to the examinee placed right above the text and questions right below the timer.

3.6.3 Usability Questionnaire Results

Although the general attitude of the participants towards taking a reading test on computer through this particular interface was positive, the results from the Usability Questionnaires identified important points regarding human-computer related usability problems (two in particular), which would otherwise not have been discovered before commencing with the main study sample. This confirmed the importance of usability testing for this particular purpose (i.e. in preparation for administering a language test on computer). The results of the usability questionnaire are discussed below.

3.6.3.1 Buttons

Five out of the six participants had questions about the symbols used representing the two buttons used to navigate back and forth through the questions (i.e. \leftarrow \Rightarrow). It appeared that they were not completely sure what would happen when clicking on the actual button, as they were not familiar with the improvised arrow symbol itself (i.e. \leftarrow and \rightarrow together). After it had been mentioned that this symbol was meant to be an actual arrow, they immediately understood its function and therefore used the buttons confidently navigating through the questions. Another problem that was identified by three out of the six participants pertained the function of the button with the word 'check' on it right below the answer box. Some students were not familiar

with the connotation of the word ‘check’ to the actual option of checking an answer. Others who did not have difficulties with this issue were worried that once they had clicked on the ‘check’ button they would not have been able to change their answers as feedback to their answer would have necessitated completion of item, i.e. no chance to change the given answer later on.

3.6.3.2 Reading Passage Scrolling Feature

Neither the participants mentioned any problems with the scrolling feature, nor did the researcher observe any problems with it during the testing session itself. Therefore, the preliminary assumption was made that it did not pose any significant problems on the usability aspect. Whether it would affect the cognitive processing in any way is part of the main study and will be further discussed there where think-aloud protocols and complementing interviews are used to provide a better insight into this.

3.6.3.3 Test Timer

The timer was initially set to twenty-five minutes for each reading passage with its ten accompanying questions. This amount of time to complete each passage turned out to be insufficient, as a number of the participants did not manage to complete some of the passage’s items within the set time. Nevertheless, in the pilot study they still had the opportunity to answer the remaining questions of that particular passage although the timer would show ‘your time is over!’ on screen.

3.6.3.4 Screen capture software (SCS)

After I had recorded two short trials (around 1 minute each) without any problems, for unknown reasons, the screen capture file of student one's main recording failed to open after having saved it. The voice recorder I used as a backup *did* record the audio for the whole test, which prevented a complete loss of the data. As this was the first time it happened, and I initially thought that it could have been a matter of familiarizing oneself with the software, I continued using the same software for the screen recording. However, another instance occurred where the software again did not record the session (i.e. student four) and gave the same error as mentioned earlier without any clear indication on where the problem could possibly originate from. Assumable is that it had to do with either software related problems or an incompatibility issue between the computer's operating system and the software. Due to these recurring issues, the decision was made not to continue with the SCS and instead observe the participants during their TA-sessions in addition to voice recording.

3.6.3.5 Recording Devices

For the voice recordings I used two devices from Sandisk®; the Sansa e 250, which is an Mp4 player with a voice recording feature; the second one was the Sansa c 240, which is an Mp3 player with a voice recording feature. Both devices worked appropriately for all four participants although both Mp3 and Mp4 recordings were not sufficiently clear at times. To avoid this happening in the main study and therefore possibly losing valuable data, I purchased high-quality digital voice recorders to be used for the main study think-aloud sample.

3.7 Implications for the Main Study

A number of problems were identified with the research instruments and user interface design in the second pilot study. The actions taken for the main study according to the problems found with both the CFQ and Interface are discussed below.

3.7.1 Computer Familiarity Questionnaire

As mentioned in the results section, three main problems were identified with some of the questions in the CFQ. The first problem related to Q1 items a-d and Q4 items a-d being interpreted in exactly the same way while the vocabulary used in both questions clearly distinguished between the two. Because the interviews showed that this was more a matter of accurately reading the question, the decision was made to inform the students of the difference before completing the questionnaire. In addition, the conclusion (after having consulted bilingual language professionals) drawn was that the vocabulary used to distinguish between *usage* and *availability* could not be made much clearer through using different vocabulary.

Q3, which attempted to elicit information on the students' perceived ability to using a computer compared to their peers was amended by deleting the phrase *compared to your peers*. In this way, the focus was shifted from the peers that the student did not have information about, to himself. Q5 item e was amended by changing the abbreviation SPSS into 'statistics', as the students *were* familiar with this term, which became apparent during the interviews.

These findings show that, despite the questionnaire being validated previously in various contexts, it is essential to pilot it with a subsample for usability purposes before administering it to the main study sample. Had a pilot study not been carried out it would have had a detrimental

effect on the accuracy of the CFQ data in the main study as the problems would then not have been identified beforehand.

3.7.2 Interface Design

The improvised arrow symbols caused confusion among a number of the students, as they were not familiar with the symbol itself. Therefore, one of the suggestions made by a number of students during the interviews was to use words instead of symbols. In order to address this difficulty, the symbols were replaced by the words *back* (\Leftarrow) and *next* (\Rightarrow). The reason for choosing *back* instead of, for example, *previous* was that the word *back* was more likely to be more appropriate to their proficiency level and, therefore, more likely to be relevant to their *real world* (Nielsen 1990;).

The majority of the students completed the test within its time limit (25 min.) and only one exceeded the 30-minute boundary (31 minutes). Therefore, the time limit set was changed to 35 minutes for each passage in order to allow the students in the main study enough time to complete the test within reasonable limits.

3.8 Summary

The pilot study proved to be valuable as it identified a number of issues with the CFQ, the interface design, audio & video recording devices, which otherwise would not have been detected before commencement with the main study.

The pilot study implemented a usability study and elicited information from the students on how they perceived and understood the computer familiarity questionnaire, which led to a number of significant findings such as comprehension difficulties related to the CFQ of Q1 item

a-d and Q4 item a-d, Q3, and Q5 item e, which were all amended accordingly aiming to prevent further problems from occurring. The interface usability study revealed a number of problems with the navigation buttons, more specifically, the language and symbols used for navigation purposes. In addition, findings on the total time needed to complete the test resulted in increased time given to students to complete the test.

Instead of using the SCS to record students' activities on screen, they were unobtrusively observed in order to get more detailed information on behaviour in both testing modes instead of only the computer-based mode. As this method proved to elicit more valuable data than the SCS alone, it was used in the main study in addition to the TA-protocols. Furthermore, higher quality digital recorders replaced the recording devices used in the pilot study to ensure maximum recording quality for the TA-protocols in the main study.

3.9 Main Study

3.9.1 Target Population and Participants

This study's target population were Saudi male students aged 18-25, studying in a Preparatory Year Program (PYP). Not only students from the city where the university is located studied there but also from the surrounding villages. Nevertheless, their educational background is the same throughout the province as all students went through the same curriculum set out by the government from Primary School up until commencing studies at University and were evaluated using the same testing system at each level (see 1.4 and 1.5). In order to study in the Preparatory Year Program, students had to pass an admission test, which ensured the minimum English proficiency level required to study there. Apart from subtle regional differences in terms of dialect, Saudi culture is uniform throughout the country where the same overall norms and

values are shared. The total number of participants in the main study was 102 and, in addition, 20 students participated in the think-aloud study, which brings the total to 122 students. However, the twenty students who participated in the think-aloud took only one passage of the test in PBT and CBT with ten accompanying items and are therefore not included in the performance analyses of the 102 students who did complete all three passages in both modes.

3.9.2 Permissions

As this study was conducted in a university setting and required access to students, institutional tests, test venues, and computer labs, a number of permissions were needed in advance. Firstly, written permission was requested from the Head of the English Department for access to the institutional tests, the students, and classrooms to administer the tests in. After that, the request with written permission from the Head of English was taken to the Dean of the Preparatory Year Program for evaluation and final approval. A Copy of the request letter signed for approval by the Dean of the Preparatory Year Program is included in appendix G.

3.9.3 Informed Consent

The students were given a general introduction in their L1, explaining what the general purpose of the study involved. Subsequently, each participant was given a bilingual informed consent form in English and Arabic, adding to what had been explained earlier in accordance with the ethical considerations involved in this study. The form explained issues such as the purpose of the study, the sponsor (i.e. University of Bedfordshire), and confidentiality guarantees among other essential parts (Denscombe, 2001). Furthermore, as interviews and think-aloud sessions were part of the data collection including recordings of both, the participants were

informed about the reason for recording these, how they were going to be used, stored, and that they would be destructed after transcription, anticipating ethical requirements (Oliver, 2003). A copy of the informed consent form given to the students before data collection is included in Appendix H.

3.9.4 Instruments

3.9.4.1 Study Tests

The tests chosen for this study were institutional reading achievement tests provided by the English language centre of the university. The reason for using achievement tests in this study is that other than Al-Amri's (2008) work, the researcher does not know of any study that used an achievement test of English as an L2 in this context; i.e. lower level Saudi Arabian preparatory year students. Furthermore, this study used L2 tests of general English whereas Al-Amri's (2008) tests were L2 English tests for medical purposes. This is one of the additional reasons why this study makes a significant contribution to the field by investigating the effect of the interface design on test-takers through this particular type of test in this context. The test used in this study contained three reading passages, each passage accompanied by ten open-ended question items totaling thirty. In order to assure that the test used in this study was appropriate for its purpose, a number of validity and reliability checks were carried out beforehand, which are discussed in section 3.7.4.5. Before discussing these, the following section describes the test's level and its underlying reading types elicited in relation to the Common European Framework (CEF) and Cambridge ESOL levels for further contextualization purposes.

3.9.4.2 Types of Reading and CEF Level

As indicated in the literature review, this study’s test mainly elicits text processing at the local level covering either expeditious reading (i.e. to locate relevant information) or careful reading (when found relevant information). Looking at the Cambridge ESOL levels, these reading types are mainly found in either KET (A2 CEF), PET (B1 CEF), and to a lesser extent FCE (B2 CEF). Table 12 below illustrates how this study’s test compares to the aforementioned levels.

Table 12. Reading Types in relation to Cambridge ESOL Levels KET and PET

	KET A2	PET B1	This Study
<u>Careful Reading Local</u> Understanding propositional Meaning at clause & Sentence level	v	v	v
<u>Careful Reading Global</u> Comprehend across sentences Comprehend overall text Comprehend overall texts	v - -	v - -	- - -
<u>Expeditious Reading Local</u> Scanning or search reading	-	v	v
<u>Expeditious Reading Global</u> Skim for gist Search reading	- -	- v	- -

Table 12 above shows that as for the types of reading tested in comparison with the reading section of the ESOL test, this study’s test and test items involve 2 of the 4 reading types covered in the B1-level of the Common European Framework of Reference, i.e. expeditious reading at the local level and careful reading at the local level as discussed earlier in the literature

review. Careful local reading is covered in both A2 and B1 whereas expeditious local reading is only covered in B1, which places this study roughly between these two reading levels.

3.9.4.3 Reliability Reading Tests

As the reading tests for the main study had not been made available at the time of sampling for the pilot study, an additional reliability check was done before commencing data collection for the main study. Due to time constraints and lack of availability of computer labs at the time of drawing the sample, it was not possible to run a reliability check on both the PBT and the CBT-version of the study tests, which resulted in only the PBT being included. The sample consisted of 33 students conveniently sampled, from the preparatory year program of which the results are shown in table 13 below.

Table 13. Reliability Statistics Main Study Test

Cronbach's Alpha	N of items	N of Subjects
.773	30	33

As shown above, the reliability coefficient is slightly below the ideal .8 in a language testing context (Bachman, 2004), and could therefore possibly take away from the strength of the validity of the interpretation of the results based on the internal consistency of the scores. After having checked the items individually, a number of items were found to be loading negatively, namely item 3 (-.116), item 9 (-.079), and item 15 (-.050). I discussed this issue with my supervisors and the decision was made to first amend these three problematic items and then run the main experiment and check the reliability afterwards. Following this, I scrutinized the three items with the examination committee of the Preparatory Year Program who were responsible

Chapter 3: Research Methodology

for writing the test's items. After we had thoroughly analyzed the items and discussed possible ways to amend them, changes were made accordingly. The original questions are presented below followed by their adapted versions subsequently.

Item 3 original:

...What do Victor and Margaret do on Saturday?

Item 3 adapted:

*...What do **the Wilsons** do on Saturday?*

Item 9 original:

...How did Mr. Wilson get a headache today?

Item 9 adapted:

*...**Why** did Mr. Wilson **have** a headache today?*

Item 15 original:

...Where did Newman first meet Woodward?

Item 15 adapted:

*...Where did Newman first **know** Woodward **from**?*

The amendments made above proved to be of significant value, as they increased the internal consistency of the items (i.e. reliability) in the study tests significantly to an acceptable level, which encouraged me to continue with the experiment. The internal consistency statistics are presented and further discussed in section 4.2.1 (chapter 4).

3.9.4.4 Test Contents

The tests from the English language centre of the university that were used by the researcher were institutional reading tests of general English. The reading part of the tests included two sections; the first section had three reading passages, each passage having ten

accompanying test items, and the second section involved an additional vocabulary exercise, which was not used in this study. The reason for eliminating this section was that the main aim of this study involved reading comprehension whereas these particular exercises assessed pre-taught vocabulary items for which the students would have had to study in advance in order to prepare for the test, which would have put an extra burden upon them. Therefore, it is not directly of relevance to this study's purpose as far as cognitive processing is concerned. A copy of the tests used in this study is enclosed in appendix A.

3.10 Study Test's Validity checks

It is important to reiterate that the aim of this study was neither to see whether the test used in this study is a valid L2 reading comprehension test nor was it to validate an L2 reading test in its entirety including all validity elements, as this would be far beyond the scope of this study. It is rather to use a reading comprehension test representative of this particular context and to compare test-takers' processes and performance on two versions of it to each other (i.e. CBT and PBT). However, in order to ensure its appropriateness for this study in this context, I performed a number of validity checks before using the test for the main study, which are discussed below.

3.10.1 Face Validity

As the name suggests, face validity is interpreted in the literature as showing acceptability of a test by its appearance. Bachman (1990) mentioned: 'face validity is the appearance of real life' (p.307). Anastasi (1988) defined face validity as follows: 'Face validity pertains to whether the test "looks valid" to the examinees, who take it, the administrative

personnel, who decide on its use, and other technically untrained observers' (p. 144). Alderson, Clapham and Wall (1995) also argued that face validity is the: 'surface credibility or public acceptability' (p.172).

However, a test only looking valid being a sufficient proof of its validity has been met with stern criticism from researchers in the language-testing field. For example, Cronbach (1984) warned that implementing a test just based off the way it looks is unacceptable, as many tests that looked good in the past have been found to have questionable validity. Although face validity is not a guarantee for test validity, it does hold a degree of practical importance in language testing, which is recognized by various prominent researchers in the field (e.g. Bachman, 1990; Alderson, 1981c; Bachman & Palmer, 1996). Roberts (2000) mentioned that provided someone knowledgeable about the subject matter executes it, there is a good reason to implement face validity in validity checks. Furthermore, Hughes (2003) ascribes significance to face validity despite it being a non-scientific idea. Due to this, and because of its practicality, it was implemented in this study's test by the researcher who himself initially executed the face validity checks based on his seven years of teaching and examining experience in the target context. In addition, the test was presented to colleagues at the English language centre of the university where the test was administered. After having examined the study's test, the conclusion was drawn that the test used in this study appeared to measure what was anticipated (i.e. local expeditious reading and local careful reading) and therefore exhibited a sufficient level of face validity for the purpose of this study in its target context.

3.10.2 Content Validity

Another validity check I made involved the test's contents, which aimed to reveal

whether the contents of the test used for this purpose (i.e. achievement test) reflected what had been taught in the course. As Hughes (2003) said: ‘A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned’ (p. 26). For example, when a student has been taught during the semester how to skim a passage to look for gist, it is expected that the test items require of him to correctly use that particular skill. Hughes further indicated the importance content validity has towards the all-inclusive interpretation of construct validity (see validity discussion section 2.6): ‘the greater a test’s content validity, the more likely it is to be an accurate measure of what it is supposed to measure, i.e. to have construct validity’ (ibid, p.27). As indicated, the test used in this study was an achievement test, which is a test that assesses certain obtained knowledge or, in this case, skills such as (among others) expeditiously searching for information in a text and locating the referent of a pronoun. These were skills that were developed through classroom instruction during the course. For the reasons mentioned, the contents of the test were scrutinized to assure their appropriateness for its purpose by comparing them to the test specifications set out by the institution. Before discussing the test specifications, a brief account of the test development process in the target context is given below. Generally across the board the process is divided into three main stages:

1. First meeting

The research committee initially meets and discusses issues related to the test to be given such as the particular skill(s) to be tested, type of text(s) to be used, what topics will be covered, how many questions are to be included, what type of questions they will be (i.e. open-ended/MCQ etc.), how much time will be allotted to complete the test, weighting of the items for marking purposes etc.

2. Allocating Teachers

A minimum of three teachers are selected for each skill that is to be tested and are asked to write exam items based on the skill they were selected for according to the guidelines discussed in the initial meeting.

3. Evaluation and Test Finalization

The committee will meet again and evaluate the tests written by the teachers and from these tests devise an exam meeting the test specifications, which were set out initially by the committee involved with curriculum development.

The test specifications are generally the same across public universities in Saudi Arabia and consist of the following seven categories with regards to testing reading:

1. Skills and strategies to be tested

Examples of these are skimming, scanning, search reading, and reading to learn.

2. Passage content

If the passages in the course books are general reading passages this should be the same in the test. In the same way, if specific reading passages are included in the teaching of for example medicine or engineering courses, they should similarly be included in the test.

3. Text Length

The length of the passages included in the test should be in within the same range as the text length in the course books.

4. Number of Sections

The number of sections for a reading test in the target context is essentially two; one section focuses on reading i.e. reading passage and accompanying items, and the other section focuses on new vocabulary, which has been taught in the course.

5. Number of Items

For each section, the number of items should be fixed; the general range for items related to the reading passage in the test is between 7 and 10, however, in practice, 10 items are generally the norm.

6. Time Limit

Generally across the board in Saudi Arabia, for midterm exams the time limit is 2 hours and for final exams is 3 hours.

7. Marking Guidelines

Generally marking guidelines should be provided in advance in addition to answer sheets according to the format of the test (i.e. MCQ/ open ended/short answer questions etc.).

By matching the study tests with the main specifications as mentioned above it became possible to examine the degree of content validity. The three passages used in this study's experiment were scrutinized one by one to ensure their suitability. Firstly, the processes that the items elicited were examined and secondly the relevance of the passage contents were evaluated and discussed in the following section.

3.11 Processes

Before naming the processes elicited through the items in each of the three passages it is important to mention that the overall level of the strategies are at the local level due to the nature of the test mainly involving expeditious and careful reading at the local level mainly related to explicitly stated information in the text (Alderson, 2000). The test items are expected to either elicit scanning or search reading processes in order to locate the explicitly stated information most likely followed by careful reading of the sentence(s) containing the keyword(s) found.

Once the relevant information has been located through one of the previously mentioned reading types, careful reading is expected to follow for word, clause, and sentence comprehension purposes to ensure correctly answering the test item. Here is where mainly the levels of cognitive processes illustrated by Khalifa & Weir (2009) are expected to be identified. The think-aloud study is thought to shed more light on the nature of the processes test-takers employ in combination with expeditious reading operations to answer these task specific test items in this study's context. Text passage two was used in the think-aloud study (see section 4.2.5 for the justification for this decision) and therefore the specific details on item difficulty, student performance on these items, question types and the processes they are likely to elicit are presented together in the results section for the purpose of clarity in subsequent discussions.

The items coincided with the test specifications i.e. locating explicit information, answering item from context (local), and identifying the referent of a pronoun, as they were the targeted elements taught during the semester in addition to careful reading. When looking at the three passages there is a clear alternation between these item types, which further confirms the test's suitability for its purpose. All three passages in the test were general English passages, which was parallel to the texts used in the course books (see appendix I for course book samples). The question format for the test used was SAQ, which was one type of the questions used in the course book. As mentioned earlier, the vocabulary section of the test was left out in this study due to its diverging focus. The time given for the test-takers to complete the test was 2 hours, which is in accordance with the higher education regulations for university exams in the target context. Matching the test specifications in this manner further strengthened the acceptability of the content validity of this study's test to be used for its purpose. I performed a further validity check, which was a text analysis of both the test and course book, which is

discussed in the section that follows.

3.11.1 Text Analysis

For the text analysis, I checked the language content of this study's test's reading passages against two randomly selected samples from the textbooks used in the preparatory year program to assure its appropriateness (Appendix I). In order to achieve that, lexical profiling of the reading passages and the sample texts from the course book was done using an online available profiling program called lextutor¹, which is widely considered a reliable profiling program by researchers and educators alike (e.g. Almazova and Kogan, 2014; Fitzgerald, 2012; Simpson, 2010). Other programs such as RANGE (Nation and Heatley, 2002) were also considered to run the lexical profiles of the texts involved but the former was found to be easier to use and was therefore used in this study. To operate the lextutor program one simply copies the required text into the space provided of which an example is shown in the screenshot below.

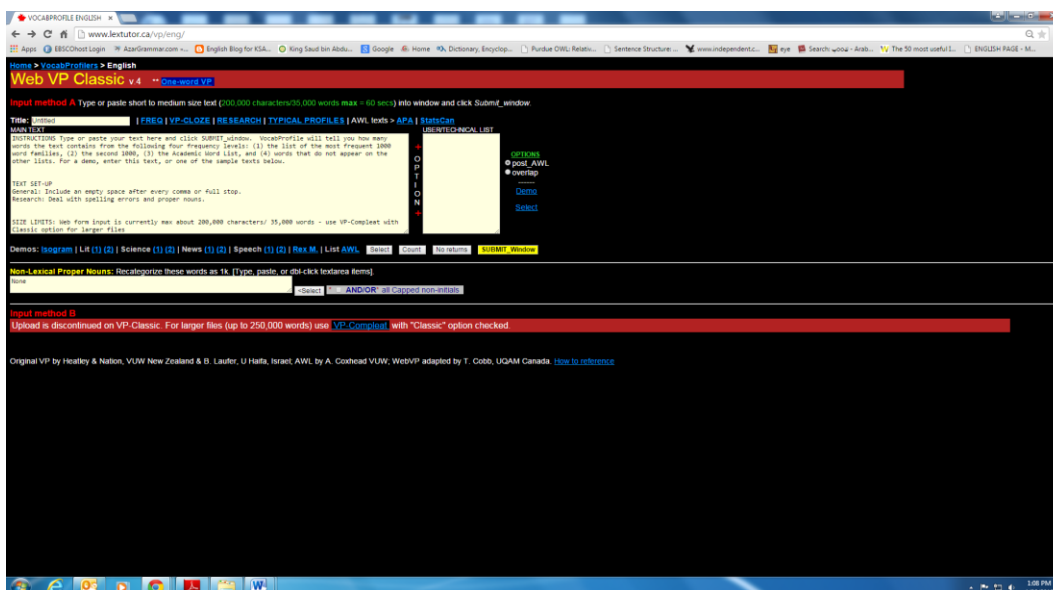


Figure 15. Screenshot Lextutor Input

¹ Available at: <http://www.lex tutor.ca/>

Chapter3: Research Methodology

After that, one clicks the ‘Submit_Window’ button as shown in figure 15 above and the program runs the lexical profile analysis. The program runs analyses on the following 4 frequency levels: (1) The 1000 most frequent word families (i.e. K1) (2) The second 1000 most frequent word families (i.e. K2) (3) Words from the academic word list (AWL) (4) Words that appear on lists other than the aforementioned (Off-list Words). Percentages are calculated based on a type/token analysis and based on these figures a comparison can be made between the text book samples and the reading passages used in this study. An example of output statistics is shown in figure 16 below.

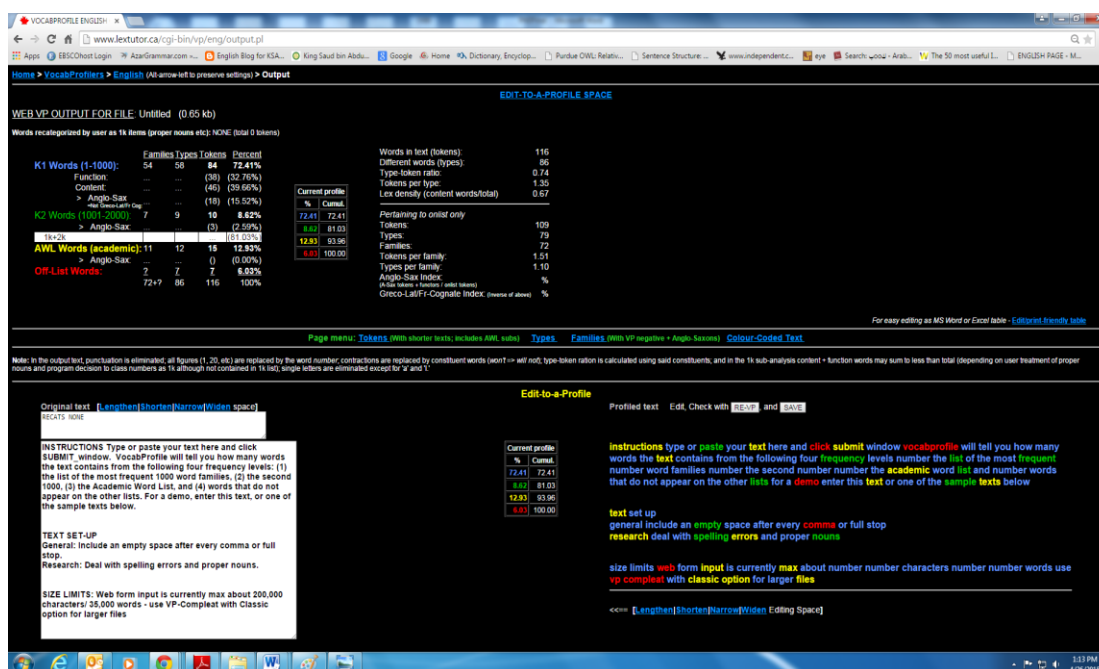


Figure 16. Screenshot Lextutor Output

In addition, readability statistics were calculated using a standard word-processing program (i.e. Microsoft Word). The lexical profiles of the three reading passages used in the study tests are shown below in table 14 followed by readability statistics of the passages in table 16.

Table 14. Lexical Profile Reading Passage

Lexical Profile	Passage 1	Passage 2	Passage 3
K1 Words 1-1000	87.47%	75.08%	78.99%
Function	46.01%	44.01%	48.61%
Content	41.46%	31.07%	30.38%
K2 Words 1001-2000	3.19%	6.15%	10.13%
AWL Words (Academic)	1.14%	1.94%	3.29%
Off- List Words	8.20%	16.83%	7.59%

As table 15 below shows, the reading ease of passage one is 77.5 and the reading ease for passage two is 65.2. Passage three appears to be easier than one and two with a 90.7 reading ease. These figures indicate an overall low difficulty appropriate to the lower language proficiency level of the students involved in this study.

Table 15. Readability Statistics Reading Passage 1, 2, & 3

Flesch Readability Statistics	Passage 1	Passage 2	Passage 3
Words	432	303	389
Passive Sentences	0%	0%	6%
Flesch Reading Ease	77.5	65.2	90.7
Flesch Kincaid Grade Level	4.9	8.1	2.6

The three passages were then compared to two passages' samples from the University course books (see Appendix I). As with the test passage of the main study, the lexical profiles of the two reading passages selected from the textbooks are shown first, followed by the readability statistics, which are illustrated in table 16 and 17 on the page that follows.

Table 16. Lexical Profiles Course Book Sample Texts

Lexical Profile	Text 1	Text 2
K-1 Words 1-1000	76.90%	88.48%
Function	44.83%	56.02%
Content	32.07%	32.46%
K2 Words 1001-2000	3.45%	3.66%
AWL words (academic)	2.07%	1.05%
Off-List words	17.59%	6.81%

Table 17. Readability Figures Course Book Sample Texts

	Text 1	Text 2
Total Words	288	188
Passive Sentences	0%	0%
Flesch Reading Ease	66.5	82.9
Flesch Kincaid Grade Level	8.1	6.3

Table 16 and 17 confirm that the language level between the study tests' reading passages and the course books are similar with a reading ease of the textbooks and the study tests ranging between 65 and 91. This is strengthened by the lexical profiles as they show similar figures

between the passages and the course books (e.g. K1 Words, Off-List Words, in addition to the readability statistics). This further confirms the appropriateness of using these reading passages for this study and its test-takers and adds to the validity and reliability of the tests used in the target context.

3.11.2 Passage Order vs. Item Order

The original item order was exactly the same as the order in which the answer appeared in the passage. As this could have affected the cognitive processing of the students (i.e. the students would expect to find the questions in the same subsequent order as the answers), the item order was switched around in order to control for this effect. Below is an overview of the item order, as they appeared in the passages in the main study.

Table 18. Question Order vs. Passage Order

Passage 1		Passage 2		Passage 3	
I.O.	P.O.	I.O.	P.O.	I.O.	P.O.
1	1	1	1	1	1
2	9	2	10	2	4
3	6	3	9	3	9
4	7	4	8	4	5
5	5	5	7	5	7
6	2	6	5	6	8
7	4	7	4	7	6
8	8	8	2	8	3
9	3	9	3	9	2
10	10	10	6	10	10

As table 18 above shows, the items for all three passages did not occur in their original expected order (i.e. chronologically). The expected result therefore would be maximization of relevant strategy utilization from the test-takers, which would likely result in a more accurate account of strategies elicited for each test item.

3.12 Test Administration Procedure

The administration procedure of the study tests for the quantitative element of main study took close to six weeks to complete, which was equally the case for the think-aloud study. Table 15 below outlines the timeline for the data collection processes in the main study.

Table 19. Timeline Data Collection Main Study

Quantitative data collection Session 1		Quantitative data collection Session 2
Week 1, 1 st week of March, 2011	5-week intermittent gap	Week 6, 2 nd week of April, 2011
Data collection Think-Aloud Session 1		Data collection Think-Aloud Session 2
Week 7, 3 rd week of April, 2011	5- week intermittent gap	Week 13, 1 st week of June, 2011

The students had completed the CFQ's before and all were sufficiently computer familiar scoring 2 or above on the familiarity scale. The number of students was split up in sections of

twenty-five due to the limited availability of computers in the computer labs. By the same token, it was easier to control and observe one section in one classroom/computer lab at a time. Before starting the test (i.e. either PBT-mode or CBT-mode), all students were informed about the purpose of the test and confidentiality was reassured in retrospect to the informed consent. In addition, the participants were informed about the amount of time they had to complete the test, which was thirty-five minutes for each reading passage as discussed in the previous section. The section's teacher was present during the exam until the students had all completed the reading tests and no assistance was further required. The study test was administered to the students in a counterbalanced order to control for mode effect, and a five week gap between the two testing sessions was included in order to minimize memory effect on student processes and performance. An overview of the administration procedure is presented in table 20 below.

Table 20. Test Administration Procedure Quantitative Study

All students CFQ (100%)					
	Session 1		5-Week Intermitted Gap*	Session 2	
	Day 1	Day 2		Day 1	Day 2
Group	25%** 1 CB	25%** 2 PB		25%** 1 PB	25%** 2 CB
No.	25%** 3 PB	25%** 4 CB		25%** 3 CB	25%** 4 PB

* In order to minimize effect of memory on test-takers' processes/performance

** Approximate estimate, due to limitations on the operational side

As table 20 shows, the examination divided over two days for each session with each session being divided over two days, i.e. 50% of the sample on each day. After the first session, a 5-week intermittent gap was maintained in order to control for memory effect, as the same test-takers were to do the same test on two occasions, i.e. within-subject repeated measures design. Although this method has been found potentially problematic for establishing test-retest reliability by a number of language testing researchers in the field (e.g. Anastasi, 1988; Alderson, 1991a; Weir, 2005), by employing the within-subject design using the same test on two occasions with the same test-takers, I could control for participants' individual differences, which possibly affected the participants in Kobrin's (2000) study who used parallel tests to investigate cognitive processes employed by her study's participants when taking a reading test in both modes. She reported that her students found one of the passages more difficult than the other, which could have caused the difference found (though in this case non-significant) between the two modes in her study. Furthermore, despite its potential problems, Anastasi (1988) pointed out that a carefully estimated intermittent gap, i.e. not too short for memory related reasons and not too long for the possible influence of environmental factors over time, potentially limits these possible confounding effects. In addition, higher-level global processes such as main idea extraction, cross-text or intertextual inferential processes

3.13 Interface Design Main Study

The interface used in the main study was the final version of the interface that had gone through different stages where elements were removed and amended in order to minimize possible construct irrelevance to be introduced by it, which could potentially skew implications drawn from results obtained in this study. It further contained the tests that were meant to be

included but due to problems at the operational part had not been used in the pilot study. This section shows the decisions made at the design level justified by a combination of previous research from the field, logic, and convenience. A screenshot is used as an illustration at various stages in order to give an insight into how the researcher put the obtained theory from the literature into practice at the design stage. Furthermore the final product of the interface used for the reading test in this study visualizing the accumulated optimal settings with reference to the interface design evaluation model in section 2.8.3 of chapter 2, is shown in figure 25 to conclude this section.

3.13.1 Interface Design: Presentation

The hot potatoes software was used to further develop/amend the interface to suit this study's purpose implementing the optimal settings gathered from the discussed synopsis of the existing literature. A full workshop on how to amend features of the interface by altering html-coding provided by hot potatoes is available online, which is particularly useful to practitioners when using this interface in their respective settings². The more basic amendments in terms of interface features can be made in the program itself without having to change html-codes. A screenshot of this is given in figure 17 below.

² https://hotpot.uvic.ca/howto/hacking_workshop/index.htm

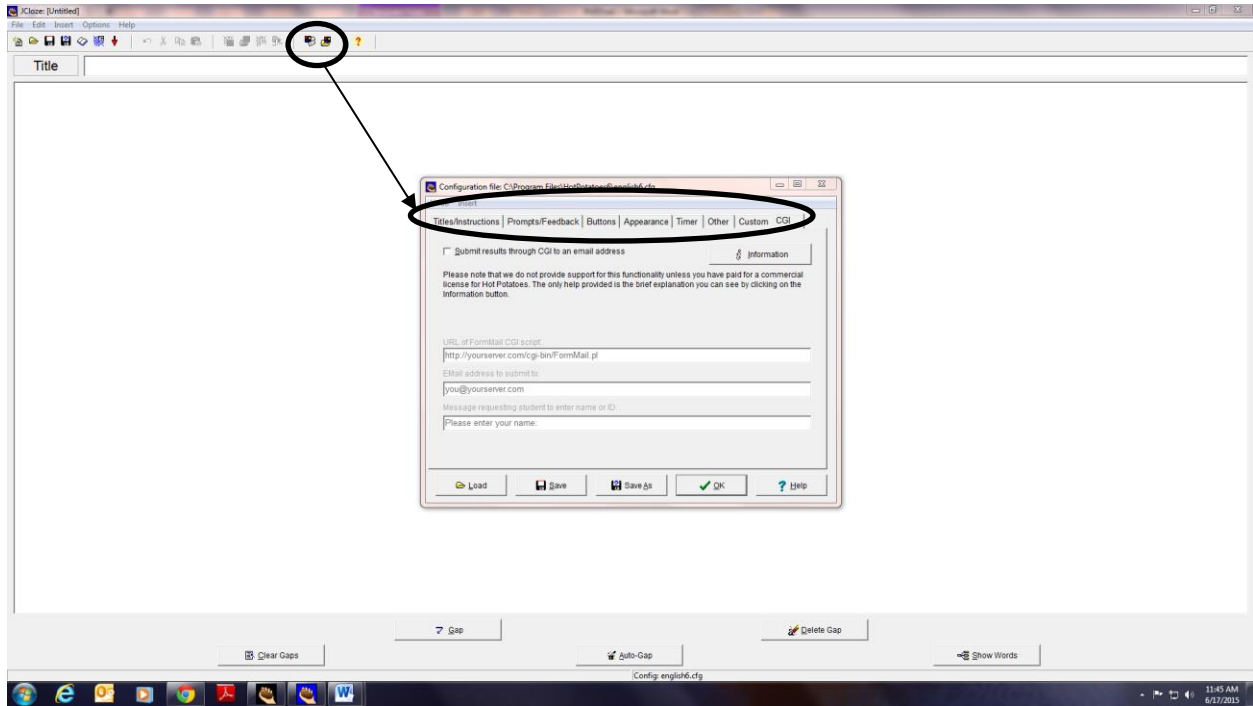


Figure 17. Screenshot Basic Configuration Hot Potatoes

As encircled in figure 17 above, a pop-up menu that allows the user to amend basic features of the interface is given by selecting the relevant icon on the main page. Amendments can be made to the title, feedback/prompts, buttons, overall appearance (i.e. fonts, colours etc.) and the timer (if used). The ‘other’ option provides the user with a number of other amendments that are shown in figure 18 below.

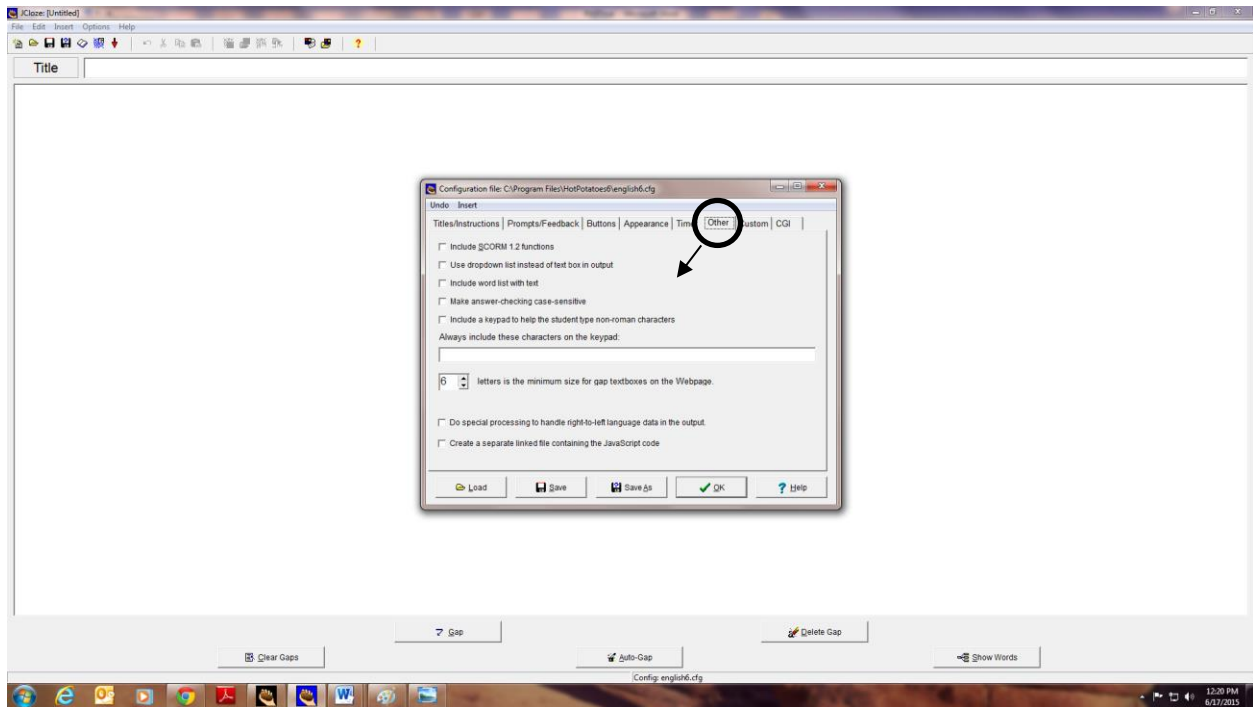


Figure 18. Screenshot Other Optional Amendments

Here choices can be made involving case-sensitivity, keypad inclusion, JAVA script coded titles, gap size between words etc. The decisions on the included elements of the interface in this study in their optimal form are discussed in sections 3.13.1.1, 3.13.1.2, and 3.13.2 in the same order of occurrence as discussed in the literature review illustrated with screenshots as found appropriate.

3.13.1.1 Typographical Factors

1. Font Characteristics

The font characteristics chosen for the interface design used in the main study were sans-serif (i.e. Arial) with a type size of 11. Although the type size recommendations based on the literature were slightly higher (i.e. type size 12), I decided to reduce the size slightly for the following reason:

Chapter3: Research Methodology

By reducing the type size it would decrease the amount of space occupied by the passage's text, which because of this would reduce the amount of scrolling required by the test-taker. This, in turn, would further decrease the possible construct irrelevant variance introduced by scrolling as suggested in some of the earlier discussed earlier studies that addressed this feature. I could afford doing this, because although recommendations based on the literature recommended a type size of 12 for sans-serif fonts, the studies that included smaller type sizes up to as low as 10 and the more recent ones in particular did neither find a significant difference between the fonts (i.e. sans-serif vs. serif) nor did they find a significant effect on performance (i.e. Beymer et al., 2007; Banerjee et al. 2011). A visualization of this amendment in hot potatoes is given in figure 19 below.

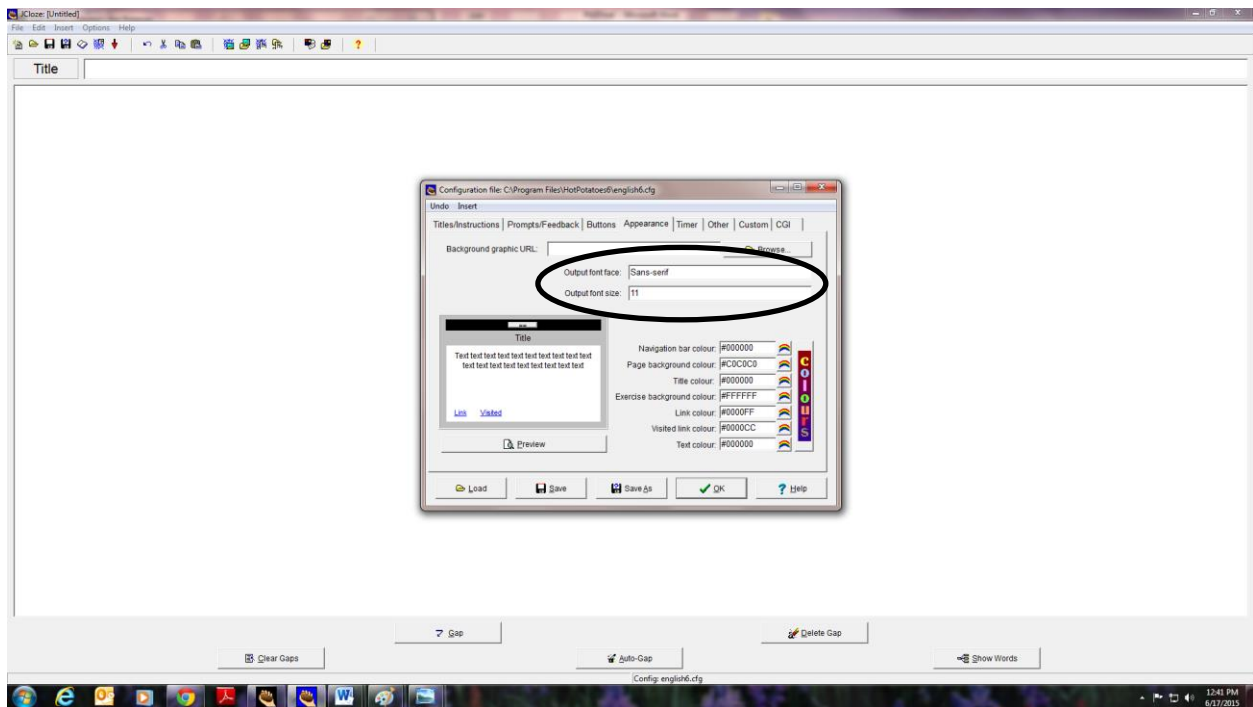


Figure 19. Screenshot Font Amendments

2. Line Length (i.e. characters per line)

The consequence of reducing the type size slightly from 12 to 11 automatically increased

the characters per line to 85 cpl, which is still well within (i.e. exactly in the middle of) the suggested range based on the reviewed literature (i.e. 55cpl – 115cpl).

3. Number of Lines

Conclusions drawn based on the (limited) literature available on the effect of number of lines on comprehension showed no significant effect irrespective of the number of lines on screen. The only findings were related to speed differences and were from studies that worked with dated computer devices (i.e. 1980s). Therefore, the assumption with regards to number of lines for this study's purpose was up to the number of lines allowed that fit the screen (i.e. 13", 15", 17" etc.). However, to actually allow a screen filled with lines would indirectly affect the amount of scrolling depending on the size of the text passage (i.e. number of words) and should therefore be considered in relation to the scrolling range, which is related to line length and so on, as discussed in the previous section.

4. Interlinear Spacing

As discussed in section 2.15.4 of the literature review, there seemed to be no solid indication regarding the ideal interlinear spacing settings based on either strong theories or evidence from relevant studies suggesting any. The only studies that addressed interlinear spacing were carried out over two decades ago and used technology appropriate to that era (i.e. Grabinger, 1993; Kruk & Muter, 1984).

However, although there appeared to be some effect on reading speed, the fact that no effect on comprehension between different spacing options (i.e. single and double spacing) was found in these studies and that single spacing was met with more positive responses in the more recent of the two (i.e. Grabinger, 1993) is encouraging to say the least, as the technological evolution over the past twenty years would only have been expected to have further minimized

any discrepancies in this regard. For these reasons, single spacing was chosen for the reading text in this study, as by doing so, the indirect effect on other features of the interface such as scrolling (i.e. decreasing scrolling range) would be minimized and therefore limit possible construct irrelevant variance from being introduced.

5. White Space

The white spacing options, which are essentially the number of columns as discussed in section 2.15.5 of the literature review, indicated that a single column layout produced optimal results. Furthermore, standard margins were indicated as acceptable for this study's purpose. Therefore, a single column layout was used with a minimum of a 0.5/1.0 margin. Again, increasing the margins on the left and right sided could have affected the amount of scrolling required and were therefore kept to a minimum as there was no accuracy trade-off.

6. Text/Background

The ideal contrast settings (i.e. text/background) that were identified based on the review of the literature were a combination of black text and low intensity background colours. Subsequently, black text was used in this study in combination with a creamy (light yellow) background to minimize eyestrain (i.e. Galitz, 2007), which was likely to have been the cause of reported eye-fatigue as a consequence in various earlier studies (e.g. Kirsch et al., 1998; Choi et al., 2003). A visualization of this amendment in hot potatoes is given in figure 20 below.

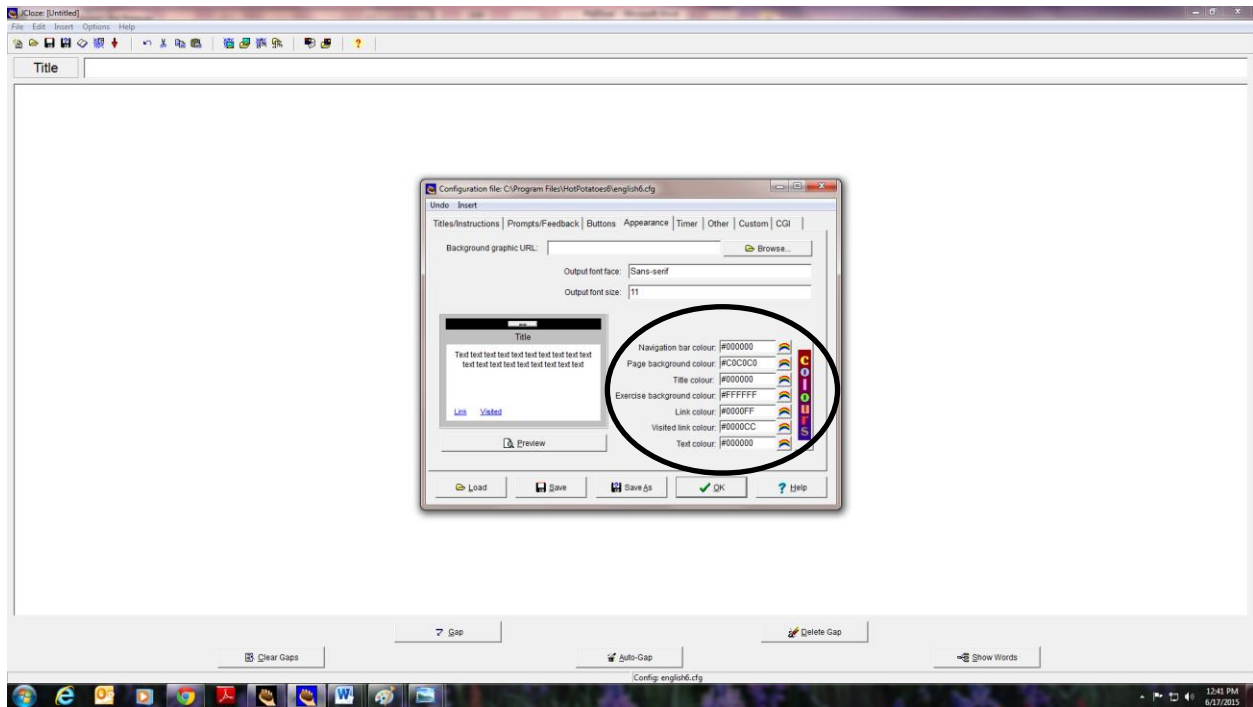


Figure 20. Screenshot Text/Background Colour options in Hot Potatoes

3.13.1.2 Graphical Factors

1. Screen Size and Resolution

The recommendations for optimal results based on the literature were a 17” screen size with a screen resolution greater than 90 dpi (i.e. > 90 dpi). The monitor used in this study was a 60Hz, 17” monitor with a screen resolution of 1920x1080, which was well within the recommended settings.

2. Icons and Button Design

As suggested in the literature review (i.e. section 2.16.2 p. 85), no use was made of icons but rather command buttons were included in order for the test-takers to navigate through the test items. In addition to suggestions from the literature, usability tests aided in optimally configured command buttons, which turned out to be a combination of arrow symbols (i.e. >) and modified

Chapter3: Research Methodology

text to suit the test-takers' language proficiency and (cultural) background knowledge. The text eventually decided upon was *next* for navigating forward to subsequent items whereas *back* was chosen for navigating to previous items. A visualization of this amendment in hot potatoes is given in figure 21 below.

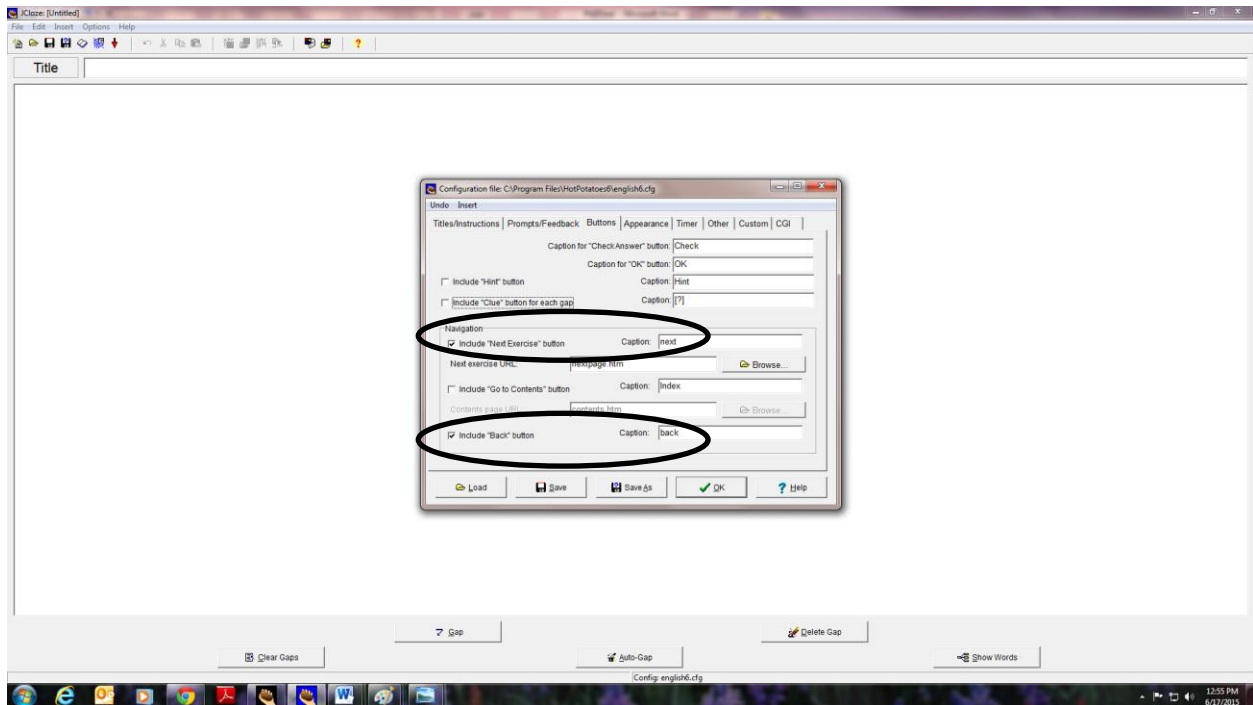


Figure 21. Screenshot Button Amendment Hot Potatoes

As shown encircled in figure 21 in the navigation section, both boxes depicting the inclusion of forward and backward navigation were selected and the text chosen to appear were 'next' and 'back' as discussed earlier. The 'include hint' button and the 'include clue' button were left out for obvious reasons and the 'go to contents' button was left out due to there not being a contents page, which was irrelevant to the test's task.

3.13.2 Interface Design: Interaction

1. Scrolling

Based on the discussion in the literature review (see 2.17.1 p.88) the decision was made to keep the amount of scrolling required when reading the passage to an absolute minimum in order to prevent construct irrelevant variance as much as possible from occurring. The actual amount of scrolling (i.e. scrolling range) was kept to approximately 30% of the total text. In order to have the text scroll independently from the screen, a change had to be made in the html-coding for which hot potatoes provided a solution as shown in figure 22 below.

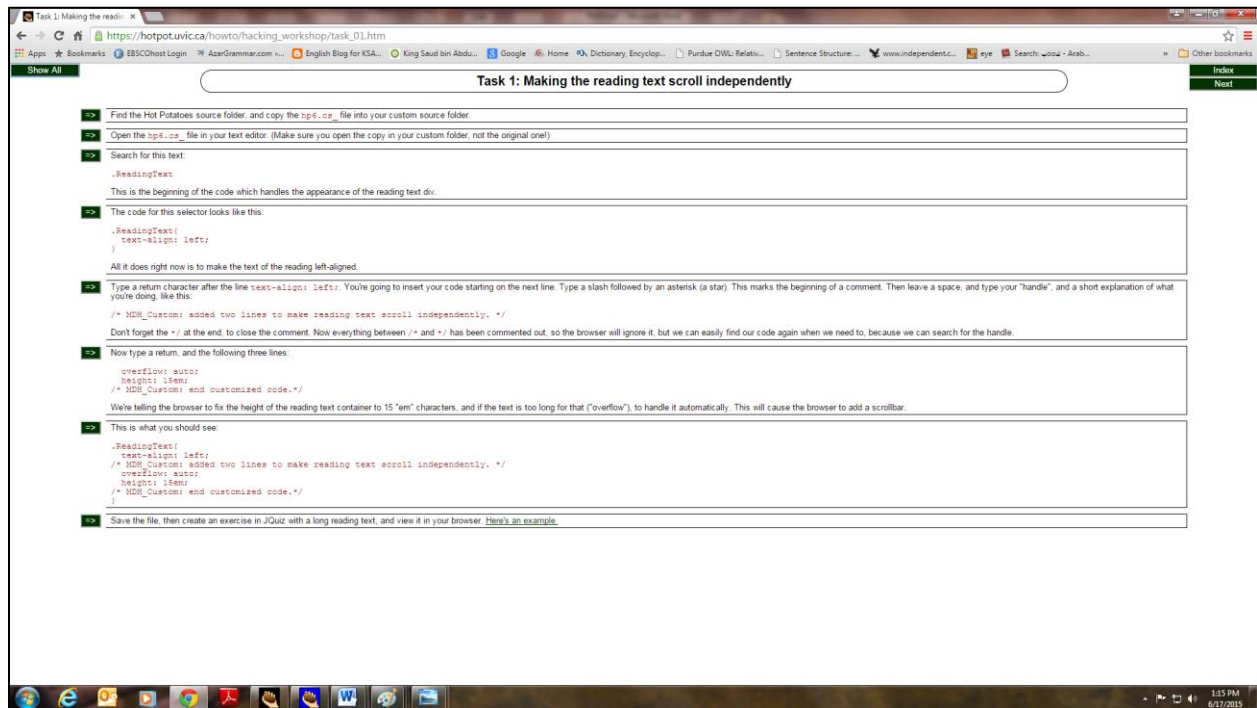


Figure 22. Screenshot Instructions Scrolling Amendments

2. Item Review

Taking into account arguments based on logical assumptions from researchers such as Dix (2005) who denoted the essentiality of the aspect of freedom to correct mistakes made at a

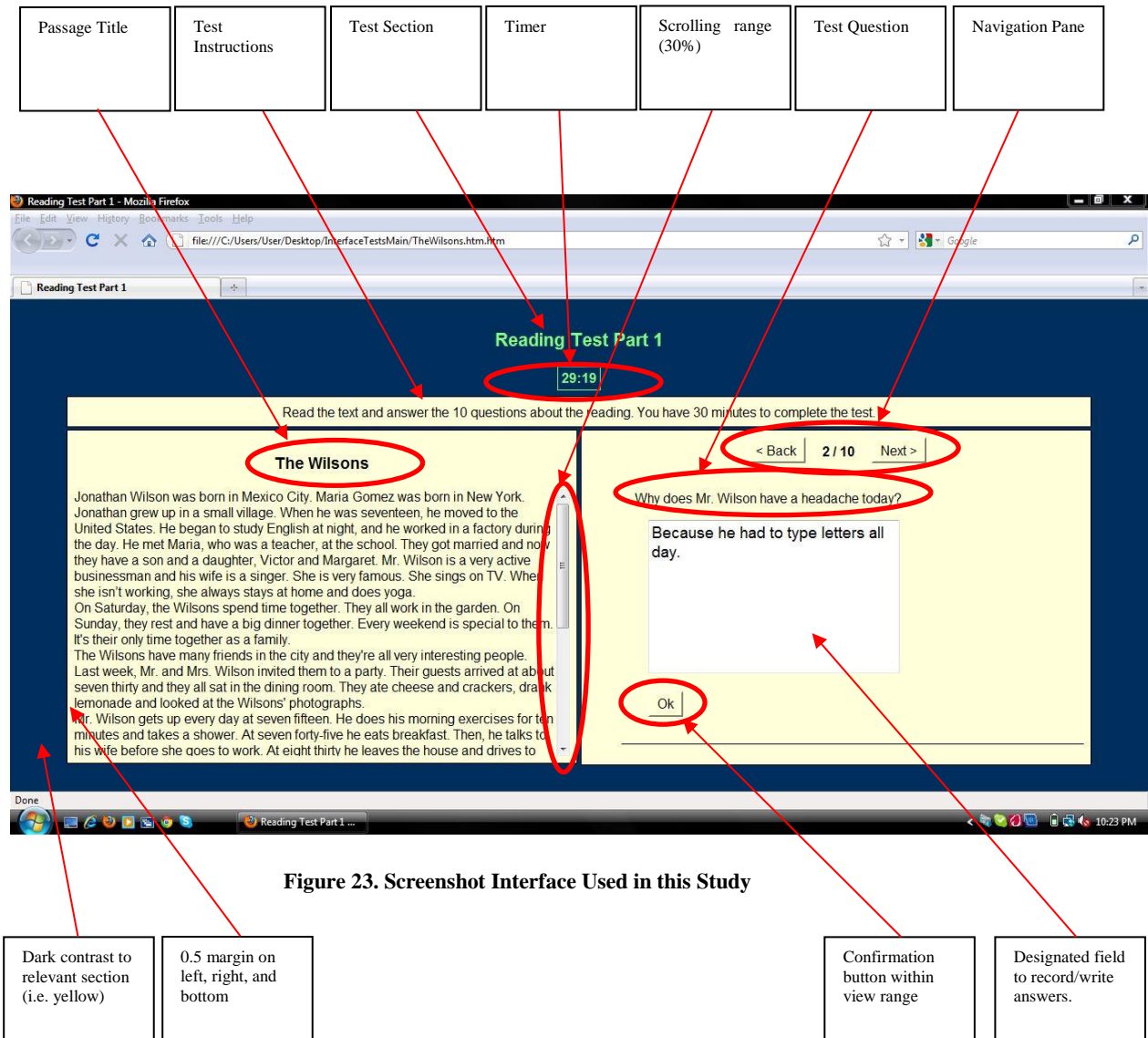
later stage in addition to other studies that investigated the effect of item review and did not find any negative effects on allowing for it (e.g. Lunz & Bergstrom, 1994; Zandvliet & Farragher, 1997; Mason et. al, 2001; Poggio et al., 2005), and, because of the default position being to allow item review in paper-based tests, the settings for this study's test likewise allowed for item review.

3. Item Presentation

The discussion on item presentation in the literature review showed that for the purpose of this study, there is no favorable setting between the grouping of items and presenting items one at a time. However, due to the assessment format involving open-ended items, it was more suitable in this case to present items individually, as it would automatically allow for enough space below the question item to type the answer. Therefore, the reasoning behind choosing one item to be presented at a time was a matter of practicality in this case.

3.13.3 End Product Computer Interface

Figure 23 below shows the visual accumulation of the optimal settings of a computer interface proposed in section 2.18 of the literature review (i.e. figure 9, p. 96) for an L2 reading test as it would appear in a testing situation. For clarity in reviewing the elements, various annotations have been made depicting the settings for the particular elements based for which can referred back to page 96.



3.14 Test Analysis

This study's test consisted of three reading passages and each passage contained ten open-ended, short-answer questions. The test was administered to the students in two modes, i.e. PBT and CBT. The test items were marked through assigning a zero for the wrong answer and a one for the correct answer, which makes maximum score to be gained for the thirty-item reading test thirty. After marking the items and double-checking of the marked items, the answers were subjected to statistical analyses using SPSS statistical package, which was used for virtually all statistical analyses in this study. Descriptive statistics were utilized in order to get an overall view of the samples score distribution characteristics in both modes. Further analyses such as group comparisons were applied to look at the significances of performance differences and correlational analyses were run to further look at the strength of the relationship between the scores on CBT and PBT (further discussed in chapter 4).

3.15 Study Questionnaires

One of the main reasons why questionnaires are popular and often used for data collection purposes is because they are: 'easy to construct, extremely versatile, and uniquely capable of gathering a large amount of information quickly in a form that is readily process able' (Dörnyei, 2003, p.1). On the other hand, incomplete or poorly completed answers, the inability to check truthfulness, and poor response rate are a number of disadvantages of using questionnaires for data collection purposes (Denscombe, 2003). It is therefore upon the researcher to arrive at a correct judgment with regards to the suitability of using a questionnaire in accordance with the given context.

3.15.1 Questionnaire Design

As the data collection in this study by way of questionnaires aimed to elicit quantifiable information about students' computer familiarity, structured, closed-format questionnaires were the most likely type of questionnaires to be used to gather this type of data with (Walliman, 2006). When respondents had been given too much freedom in answering the questions by using an open format, it would have led to predominantly qualitative answers. As the responses for the computer familiarity questionnaire were solely subjected to quantitative analyses, it would not have been suitable for this particular purpose. However, the second questionnaire consisted of a mix of closed and open format questionnaire types as several closed questions were followed up by an open question investigating the *why* and *how* of the answers given (Oppenheim, 1992). This semi-structured method allowed for more freedom on the participant's part in answering the questions, which aimed to lead to more qualitatively enriching data, which was intended to add to the data gathered from the analysis of the score outcomes and in support of the think-aloud sessions. The CFQ and the Post Test Questionnaire are used in this study to serve the following main objectives:

Computer Familiarity Questionnaire

1. To gather Demographic data about the participants (i.e. Gender, Nationality, Institution).
2. To gather information about participants' computer familiarity in its unified form as based on Kirsch et al. (1998) to ensure suitability for this study's purpose.

Post-Test Questionnaire

1. To obtain information about students' attitudes towards the main features of the Computer-Based Test compared to the Paper-Based.

2. To gauge which mode was favorable for the test-takers in terms of the features mentioned in the first point.

The following subsection discusses constructing questionnaire one and questionnaire two in more detail. The first questionnaire was administered before the study tests whereas the second questionnaire was given to the subjects after having completed both tests (i.e. computer and paper-based).

3.15.2 Computer Familiarity Questionnaire (CFQ)

The CFQ was given to the test-takers before the test event and consisted of twenty- five items that collectively addressed four different aspects of computer familiarity which are computer access, attitudes, experience or use, and related technology. The reason for including these four aspects of computer familiarity is that Kirsch et al. (1998) gathered the different types of computer familiarity from what had already been established in the literature and formed them into one computer familiarity scale in their study, which was validated by Eignor et al. (1998) in their study. Furthermore, the questions in the questionnaire were used by the researcher in a previous study (i.e. Korevaar, 2008), and were taken from Weir et al., (2007) who had formed it from two previously validated instruments in the literature. The first instrument was originally developed for PISA (Program for International Student Assessment) who observed 15-year-olds from the principal industrialized countries for three consecutive years. They administered a paper and pencil test to 265,000 students from 32 different countries and had them fill out specifically designed questionnaires which measured a number of issues related to computer familiarity such as; *perceived ability & comfort, interest in computers, affect and computer usage* (Weir et al., 2007), which is in line with Kirsch et al.'s elements comprising computer familiarity. The second

instrument was the computer attitude questionnaire developed by Knezek & Christensen (1995, 1997), which focused on young learners. The complete instrument consisted of eight different categories. The anxiety category was modified and added to the final instrument. To ensure that the changes made did not affect the information sought, Weir et al. (2007) trialed and revised the questionnaire before using it for their research to ensure the appropriateness of the amendments that had been made. As investigating computer anxiety is not part of this study’s objective but rather establishing computer familiarity, this category was left out in this study. The questionnaire sections, their related items, and sources are summarized in table 21 on the page that follows.

Table 21. Overview Division CFQ Questions

Name of Section	Item Numbers	Origin
Demographic Data	A, B, C, & D	Researcher
Computer Access	1-4	Kirsch et al. (1998), Weir et al. (2007)
Attitude & Ability (Keyboard)	5-8	
Computer Use/Experience	9-16	
Related Technology	17-21	Kirsch et al. (1998), Weir et al. (2007)
Attitude & Ability (Keyboard)	22-24	Weir et al. (2007)
Computer Use/Experience	25	Kirsch et al. (1998) & Researcher

Question A-D collected general demographic data about the participants such as, the student's name, gender, nationality, and institution of study. As all students came from the same province (i.e. Hail) no information was collected with regards to region or province. Although all students in this study were of Saudi nationality, theoretically it was still possible to have one or two students with a different nationality. However, even if this had been the case, the student was most likely born and raised in the same region, and had therefore gone through the same educational system as the Saudi nationals. Nevertheless, I maintained item C, which required the students to specify their nationality. The function of the questions used is discussed below maintaining Kirsch et al.'s (1998) view, which is a single, comprehensive measure of computer familiarity. Items 1-4 elicited information about students' place of access to computers, or where the participants can use computers (i.e. at school, at home etc.). Items 5-8 were a self-assessment of attitude and ability of the students in relation to using computers. Items 9-16 aimed to gather information about students' computer use and experience with computers whereas items 17-21 addressed students' use and experience with *related technology* such as word processors, games and others. Items 22-24 served as a self-assessment of attitude and ability of using a keyboard, which is in essence similar to Items 5-8 yet specifically aiming at the keyboard instead of the computer in general. Item 25 was added following Kirsch et al. (1998) and amended to fit the purpose of this study, as Kirsch et al.'s (1998) study focused on the TOEFL exam whereas this study involved institutional reading tests. Therefore, *TOEFL test on computer* was replaced by *reading test on computer*. Furthermore, instead of looking at the number of times a student had completed a reading test on computer, for the purpose of this study, the sole interest was in

whether the student had taken a reading test on computer before. Therefore, the question was changed into: *Have you ever taken a reading test on computer?*

3.15.2.1 Administering the CFQ

The CFQ was given to the students some time before starting their reading test to allow for the analysis of their responses to ensure sufficient familiarity. All participants were given 15 minutes to complete the CFQ simultaneously in their respective classrooms. Before filling out the questionnaire, the students were instructed how to respond to the questions in order for the procedure to run as smoothly as possible. After the students had completed the questionnaire, they were collected and subjected to further analyses.

3.15.2.2 Data Analysis CFQ

As for data analysis of the CFQ, factor analysis was employed firstly in order to create a reliable computer familiarity scale. After that, correlational analyses were used to further confirm the reliability of the computer familiarity scale created by the factor analysis. A factor was calculated for each student based on his answers given on the CFQ items on their computer familiarity (ranging from 1-5) meaning the higher the assigned computer familiarity measure, the more familiar the test-taker was with computers. All students scored 2 and above and therefore participated in the main study.

3.15.2.3 Computer Familiarity Scale Development

As the number of participants in the two pilot studies was not sufficient to create a reliable computer familiarity scale (i.e. only 4 in the first and 10 in the second pilot study), the

main study's participants (n=102) had to be included in order to achieve sufficient number of test-takers for the reliability figures related to the scale to become meaningful. The initial item total for the CFQ was 25, which was later reduced to 18 items as a result of factor analyses that were carried out on each of the questionnaire's items. This was done in order for the scale to be as comprehensive as possible following Eignor et al.'s (1998) approach creating a single computer familiarity scale encompassing the four aspects of computer familiarity. As recommended by Eignor et al. (1998), a categorical principal component analysis was run with the items loading on only one dimension since only one all-inclusive measure of computer familiarity was required. The results are shown in table 22 on the page that follows.

Table 22. Loadings of Computer Familiarity Questions on Familiarity Scale

Q. No.	CFQ Questions	Component 1
Q1	How often is there a computer available to you to use at home?	.212
Q2	How often is there a computer available to you to use at University?	.481
Q3	How often is there a computer available to you to use in the library?	.559
Q4	How often is there a computer available to you to use at another place?	.516
Q5	How comfortable are you with using a computer?	.590
Q6	How comfortable are you with using a computer to write a paper?	.647
Q7	How comfortable would you be taking a test on computer?	.579
Q8	How would you rate your ability to use a computer?	.607
Q9	How often do you use a computer at home?	.252
Q10	How often do you use a computer at University?	.559
Q11	How often do you use a computer in the library?	.678
Q12	How often do you use a computer at another place?	.508
Q13	How often do you use the internet?	.188
Q14	How often do you use a computer for email/chat?	.432
Q15	How often do you use the computer for school/studies?	.609
Q16	How often do you use the computer for programming?	.650
Q17	How often do you use each of the following kinds of computer software (games)?	.136
Q18	How often do you use each of the following kinds of computer software (Word)?	.562
Q19	How often do you use each of the following kinds of computer software (Excel)?	.563
Q20	How often do you use each of the following kinds of computer software (Graphics)?	.493
Q21	How often do you use each of the following kinds of computer software (Statistics)?	.512
Q22	How do you feel about using the keyboard? I can type as fast as I can write	.564
Q23	How do you feel about using the keyboard? I do not think it is a problem for me	.499
Q24	How do you feel about using the keyboard? I find using the keyboard difficult	-.259
Q25	Have you ever taken a reading test on computer?	-.259

* Item 4 was also deleted as it loaded below .4 after the second trial

The loadings of each item on the one-dimensional computer familiarity scale, accounted for 25% of the total variance. Reliability analysis of the 25-item questionnaire showed a reliability of .837. In order to increase the reliability of the scale, I deleted the items that loaded below .40 on the single dimension and the component analysis was run again as shown below.

Table 23. Loadings of CFQ Questions on Familiarity Scale after item omissions

Q. No.	CFQ Questions	Component 1
Q2	How often is there a computer available to you to use at University?	.476
Q3	How often is there a computer available to you to use in the library?	.577
Q5	How comfortable are you with using a computer?	.579
Q6	How comfortable are you with using a computer to write a paper?	.666
Q7	How comfortable would you be taking a test on computer?	.584
Q8	How would you rate your ability to use a computer?	.576
Q10	How often do you use a computer at University?	.568
Q11	How often do you use a computer in the library?	.705
Q12	How often do you use a computer at another place?	.464
Q14	How often do you use a computer for email/chat?	.402
Q15	How often do you use the computer for school/studies?	.630
Q16	How often do you use the computer for programming?	.689
Q18	How often do you use each of the following kinds of computer software (Word)?	.572
Q19	How often do you use each of the following kinds of computer software (Excel)?	.606
Q20	How often do you use each of the following kinds of computer software (Graphics)?	.500
Q21	How often do you use each of the following kinds of computer software (Statistics)?	.557
Q22	How do you feel about using the keyboard? I can type as fast as I can write	.531
Q23	How do you feel about using the keyboard? I do not think it is a problem for me	.466

Table 23 shows the loadings of the 18 remaining items on the single scale, which increased the overall reliability from .837 to a Cronbach's alpha of .87. The variance accounted for increased from 25% to 32% after item omission with the remaining 18 items all being above

.4 as shown in table 23. These results gave confidence to use this questionnaire as a single, one-dimensional, reliable computer familiarity measure for the main study. Then, a factor score was created for each student based on the results of his questionnaire response ranging from 1-5. Worth to note is that Eignor et al. (1998) categorized the participants in their study into three computer familiarity groups; lower familiarity, moderate familiarity, and high familiarity whereas in this study, the scale was kept continuous, as the purpose for using the questionnaire was merely to ensure sufficient computer familiarity and not to treat it as an independent variable. The lower the scale-score assigned to a participant, the lower his familiarity with computers would be, and vice versa. All students' computer familiarity scale scores were above 2 (i.e. at least moderately familiar), which indicated adequate computer familiarity for the purpose of this study and were therefore included (n=102). Any score below 2 would have meant exclusion of that participant from the main study sample.

3.15.3 Post-Test Questionnaire (PTQ)

The Post-Test Questionnaire was for a large part adopted from two previously administered questionnaires by Boo (1997) and Gorsuch (2004). The questionnaire intended to elicit information about students' attitudes towards completing the reading test on computer and on paper *after* they had taken both. This included questions about various features of both tests in comparison to each other as well as features specific to the computer interface. The researcher added a number of questions to make the questionnaire more comprehensive in achieving its objectives. An overview of the questions with references to their sources is given in table 24 below and discussed subsequently.

Table 24. Overview Division Questions Questionnaire Two

Item Numbers	Origin
PQ1, PQ2, PQ4	Gorsuch (2004)
PQ3, PQ5, PQ6, PQ7, PQ8,	Researcher
PQ9-PQ13	Boo (1997)

PQ1 and PQ2 elicited information about the reading ease of the questions in the PBT-mode and the CBT-mode. PQ3 and PQ5 intended to gather data about differences in cognitive processing when reading the text and questions on computer and paper. These questions were both followed by an open question (why?) in order to give the student the opportunity to elaborate in detail on the answers given. PQ4, 6, and 7 aimed to gather information about students' perceptions about some of the most significant features of the CBT. Questions were asked on the degree of easiness in using the scrolling feature, navigation buttons and about the appropriateness of the screen size. PQ8 attempted to obtain students' views about on which test they thought to have performed better (i.e. PBT or CBT). Question PQ9-PQ13 targeted mode preferences and elicited information comparing various characteristics of the CBT and the PBT. For example, readability of the text, easiness of writing down and changing answers and navigation were covered. Question P10, focused on the test taking preference i.e. whether the

subject preferred to take the CBT or the PBT. The complete questionnaire can be found in Appendix D.

3.15.3.1 Administering the PTQ

The PTQ was administered to the participants after they had completed both the computer-based form and the paper-based form of the reading test. As with the CFQ, the PTQ was given to the students at the same time during class hours and were allowed thirty minutes to complete it. As the emphasis of this questionnaire was to elicit qualitative data, as opposed to the CFQ, students were advised to try to answer the questions as comprehensively as possible when given the opportunity (i.e. open-ended follow-ups to several questions). This further justifies why more time was given to the participants in order to complete the PTQ. After the PTQ's had been completed by all participants, the questions were analyzed in order to look for patterns in behaviour and attitudes, which could be then further explored during retrospective interviews. The complete Post-Test Questionnaire can be found in Appendix D/E in English and Arabic.

3.15.3.2 Data Analysis of the PTQ

The PTQ elicited information about students' experiences with features of the computer interface and how they compared to the paper-based test. This questionnaire was mainly used for further illustration in support of the results of the performance analyses. Therefore, test-takers' responses were described in percentages for each question in the questionnaire. Another function of the post-test questionnaire was to use the information given to support findings from the think-aloud data and their subsequent interviews. It also formed a basis for questions to be further explored in the post-test interviews. For example, if a substantial number of test-takers

had answered that they found it easier to read the text on screen than on paper, interviews could be used to gauge the underlying cause from their perspective. How the interviews were held with this study's think-aloud participants is discussed in the following section.

3.16 Interviews

3.16.1 Instrument Rationale

Interviews are and have been integral part of qualitative data collection in L2 reading and assessment research and are used for gathering data in various types of research approaches. Kerlinger (1970) advocated that interviews could be very helpful as a way of following up on unexpected results in, for example, an experiment, or to go deeper into the motivation of participants behind certain answers they have given on a questionnaire. It could further function as a validation tool for other methods employed in the same study or help develop another method of data collection by means of triangulation (Green, Caracelli, and Graham, 1989). The function of the interviews in this study is in line with the previous mentioned theories, as it seeks to get a deeper understanding of the reasoning behind the (cognitive) behaviour of the test-takers when taking a reading test (i.e. their strategies) on the one hand and serves as a tool to partially validate what has come up in the post-test questionnaires filled out by the test-takers on the other hand. Furthermore, the way of probing made possible by way of interview is difficult to achieve when using a questionnaire only for various obvious reasons, and complements therefore one of its limitations (Walliman, 2006). The three most common types of interviews used in applied linguistics research are structured interviews, unstructured interviews, and semi-structured interviews (e.g. Creswell, 2003; Walliman, 2006; Dörnyei, 2007). For this study's purpose, semi-structured interviews were used, as they aimed to validate responses of the test-takers to a

previously given post-test questionnaire. Like in unstructured interviews, open questions were formulated, however, they were based on the answers given in the previously given post-test questionnaire. In this way, it guided the interviewee (test-taker) but gave him the opportunity to respond freely to the particular question in an exploratory way at the same time, hence, semi-structured (Dörnyei, 2007).

3.16.2 Procedure

After I had gone through the answers given by the students on their questionnaires, I noted any responses that I thought were unusual and formulated a question based on the identified variability and also did this with the think-aloud recordings. Generally, it involved the type of questions that elicited explanations from the test-takers on why they had performed a certain action or why they had employed a certain strategy or why a combination of certain strategies was used as opposed to what was normally to be expected when completing that particular item. The students that were interviewed were the same students that made up the think-aloud sample, which was a total of twenty test-takers at the outset. However, due to limitations on the operational part, only nine students were eventually available for interviewing. As mentioned earlier, semi-structured interviewing was chosen as the most appropriate interview technique for this study's purpose. Dörnyei's (2007) proposed questions when conducting interviews in applied linguistics guided the approach taken in this study and consisted of the following question types:

1. A number of opening questions
2. A number of (open) content questions
3. Probes (i.e. why, what do you mean by that, how, etc.)

4. A final closing question

Leading questions were avoided as suggested by Patton (2002), and the number of content questions was limited too, as they were effectively probes in part, due to the interview mainly being a follow-up to a combination of earlier questionnaires and think-aloud data. During the interviews, naturally there were additional probes introduced based on the answers given in order to increase the fruitfulness of the interview data. The interviews were conducted in the participants' preferred language (i.e. English or Arabic). When a participant had chosen Arabic as his preferred language, the interview was translated into English afterwards before being transcribed. The majority of the participants felt comfortable enough with interviewing in English, however, whenever a difficulty arose in explaining a particular point in English, participants were free to switch to Arabic in order to provide an optimal platform for free expression during the interview. Each interview lasted around 5-10 minutes for each participant.

3.16.3 Interview Data Analysis

After all participants were interviewed, the recordings were saved on an external hard-drive for safekeeping. The interviews that were conducted in English were directly transcribed verbatim whereas the interviews that were conducted in Arabic were transcribed in Arabic first and translated into English afterwards. The aim was to interview significantly more participants in both modes than eventually turned out to be the case (only nine instead of the eighteen that had been anticipated initially). Nevertheless, the data obtained from the interviews still proved to be a worthwhile addition to illustrating test-takers' cognitive behaviour when taking a reading test, as will become clear in the discussion section.

3.17 Think Aloud Protocol

3.17.1 Instrument Rationale

The think-aloud method has been used for data collection in various research disciplines for many years before being introduced into language research, particularly in the field of psychology. Karl Duncker (1945) was one of the early psychologists who conceptualized think-aloud reporting by labeling it a way of *productive thinking* and also as an aid to understand the development of thought of the participants in his study at the time. Green (1998) described verbal reporting as: ‘a special label used to describe the data gathered from an individual under special conditions, where the person is asked to either ‘talk aloud’ or to ‘think aloud’ (p.1). It is thought to give insights into learners thought (i.e. mental) processing when performing various tasks, which is formulated by Cohen (1998) as a: ‘stream of consciousness disclosure of thought processes while the information is being attended to’ (p.34).

Several studies were conducted using the think-aloud method about fifteen years later, mainly focusing on problem solving-strategies that involved non-verbal tasks (e.g. Gagné & Smith, 1962 and later by Davis, Carey, Foxman, & Tarr, 1968). About two decades later, the think aloud method was proposed as a valid way to investigate cognitive processing signified by the often-quoted seminal work of Ericsson & Simon (1993) who synthesized earlier work on think-aloud reporting and developed a model based on STM (short-term memory) and LTM (long-term memory). During the mid-80s through the early 90s, a number of studies were carried out using think-aloud to collect data on reading behaviour in L1 (e.g. Cohen 1986, 1987; Gordon, 1990; Earthman, 1992), and likewise in L2 (e.g. Cohen, 1984, 1986, 1987; Cohen & Cavalcanti, 1987; Pritchard, 1990; Pressley & Afflerbach, 1995). Green (1998) further argued think-aloud reporting to be a valid method in aiding to establish test validity and reliability. In

this study, the think-aloud method is being used to provide evidence for the cognitive validity of the CBT through the recording of the cognitive behaviour of the participants (e.g. the strategies used when completing the reading test) compared to their cognitive behaviour on the PBT.

Think-aloud reporting can be done either retrospectively or introspectively (i.e. concurrent think-aloud reporting). Retrospective think-aloud reporting refers to verbalizations after the subject has completed the task whereas introspective or concurrent think-aloud reporting refers to verbalization whilst performing the task at hand. Cohen (2006) further distinguished between two types of verbalizations in introspective/ concurrent thinking aloud; *mentalist*, where test-takers describe what they are thinking/doing in order to complete the task at hand, or *non-mentalist* verbalizations, where the test-taker does not explain what he is thinking but rather speaks his mind without explaining it. Cohen (2000) also referred to these as self-observational for the former and self-revelational for the latter. Non-mentalist verbalizations are generally preferred, particularly in research in language learning and testing as they are thought to more accurately reflect thought processes (e.g. Green, 1998; Cohen 2000; Cohen & Upton, 2006) For these reasons, introspective thinking aloud was chosen for this study's purpose in combination with non-mentalist verbalizations from the test-takers to increase the validity and reliability of the interpretation of the results obtained from the think-aloud reports.

3.17.2 Sampling of Think-Aloud Participants

The aim for the number of students to participate in the think-aloud sessions was a total of twenty-five students, which would account for approximately 25 % of the total of 102 students that partook in the main study. However, five students were not available for the second

think-aloud session after having completed the first as we were towards the end of the semester at the time, and the students had already made holiday plans. This left me with a sample of 20 students who had completed the think-aloud sessions in both the PBT and CBT. Furthermore, after the data collection was completed, two of the students only verbalized one of their two sessions sufficiently for transcription purposes despite repeated prompting on the researcher's part during the think-aloud sessions. By that time it was too late to reschedule any students for re-sitting the think-aloud sessions, as I was at the end of my scheduled data collection time limit on the one hand, and the majority of students had already left or had made plans for the holidays on the other. In total, out of the twenty-five students anticipated, eighteen students eventually partook in both think-aloud sessions.

3.17.3 Instruments used in TA-Sessions

3.17.3.1 Reading Passage

The reading passage to be used for the think-aloud sessions was chosen based on a number of issues. Firstly, I chose one passage out of the three used in the main study, as introspective think-aloud is known to significantly slow down the test-taking process, which was the reason for Green's (1998) suggestion of allowing more time for the student to complete a task in order to counteract this problem. So the time to be spent on one passage whilst thinking aloud had to be estimated for the subjects taking this into account and was estimated to be between 35-50 minutes for each session. This was based on averaging out the total time spent by the students in the main study on the CBT and the expected extra time needed due to introspective verbalization. The average time spent on the chosen passage when verbalizing

turned out to be between 35-45 minutes (outliers on either side excluded), which was within the expected range established in advance.

Secondly, choosing only one passage with ten accompanying items was anticipated to be sufficient as the items assessed mainly the same reading types, i.e. local expeditious and careful reading. Therefore, ten items would be sufficient to get a good insight into the processes test-takers employ when answering these items.

Thirdly, the second passage was chosen for the think aloud based on the item analyses' results of the main study due to the significant difference found on that item favouring PBT, which gave the opportunity to further explore this significant difference qualitatively through examining the underlying processes when answering this item in PBT compared to CBT.

3.17.3.2 Training Materials TA-Sessions

The reason for using training materials before the think-aloud sessions was based on recommendations from researchers such as Green (1998) who stressed the importance of having subjects train in advance on the new method to prevent collecting inaccurate data due to participants' unfamiliarity with the mechanics of thinking aloud itself. Other researchers such as McDonough (1995) agreed and argued that doing this is expected to enhance data obtained from participants as a result. For the training session in this study, students were asked to complete a mathematical problem whilst thinking aloud. The reason for choosing a different discipline than in the main study is following Scholfield's (2006) criteria, which iterated not to choose training materials in the same skill/area as the study to be done by the participants. Two simple mathematical multiplication problems were selected for the participants to solve using think-

aloud to see whether they were likely to correctly verbalize their thoughts in the think-aloud study.

3.17.3.3 TA Training Session

The researcher introduced the students to the concept of think-aloud in the beginning in order to assure everyone was familiar with it before starting to practice it independently. Then, before the participants started their training session, the researcher thought aloud in front of the group, as to give the participants an idea of how thinking-aloud looks in practice and to get a grasp of what was being required of them. After that, the participants thought aloud solving the two mathematical problems while the researcher went around to make sure all did exactly what was required of them. All in all, the session proved to be an efficient way to ensure all participants were comfortable with putting the think-aloud method into practice, which gave me confidence to proceed with the qualitative data collection sessions using the think-aloud method.

3.17.3.4 Procedure Think Aloud

The twenty students that participated in the think-aloud sessions were scheduled at their convenience two at a time per session. This meant a total of ten sessions for the PBT's and another ten for the CBT's. The participants were given forty-five minutes to complete each version of the test, which added up to around ninety minutes of recording for each student. Each session lasted between 45-55 minutes in total. The participants were reminded to think aloud and to try to maintain a flow talking before they started each session. The digital recorder was put in place (out of sight from the participant) and the researcher sat down unobtrusively in the room to observe one of the two the participants' behaviour. The participant that was not observed w

verbalized alone in a separate room. In each session, two students were thinking aloud at the same time albeit in different locations (i.e. language labs) to prevent disruption. One student was observed in each session. The freedom was given to the students to verbalize in both their L1 and L2 as seen fit, as other studies that have done the same did not find any effect on either cognitive processing or performance in the case of Saudi students, which is the target context in this study (i.e. Addamegh, 2003). The two versions of the test were the **same**, however, the participants were not made aware of this fact in order to aid in control for memory effect, which would increase the reliability of the obtained data. They were only told that there was going to be a second session where they had to do another exam, but no information about its nature was given. Furthermore, a 5-week gap was maintained between the participant's first and second session, which further aimed for memory-effect control. The mode order was counter-balanced as was the case in the main study, i.e. nine students did the PBT first and CBT second and nine did the CBT first and the PBT second. The participants that were observed were interviewed after they had finished the test. The TA-recordings were firstly transcribed verbatim, and the interviews were subjected to further analysis.

3.18 Think Aloud Data Analysis

3.18.1 Protocol Transcription

The researcher and a colleague, who is an academic expert in the Arabic language and translation studies and bilingual (i.e. Arabic and English), transcribed the think-aloud data together verbatim. This was done in order to ensure accurate transcription in both Arabic and English, and to ensure accurate translation from Arabic *into* English for the parts of the protocol that were verbalized in Arabic after the protocols were transcribed. A list of transcription

Chapter3: Research Methodology

conventions was developed by the researcher in order to create an as transparent, clear and easily readable transcription as possible, through the following list of abbreviations and varieties of text/font display:

[Strategy]: indicates the assigned strategy, e.g. [OS1]

Underlined: the student is reading the actual text of the reading passage

Italics: the test-taker is reading the question

Normal font: verbalization in English

(normal font): verbalization in Arabic

(.) short pause (i.e. > 10 sec)

(...) longer pause (i.e. < 10 sec)

word* indicates that the word was misread/mispronounced by the student

(= text) correction of the mispronounced word

{UV}: unclear verbalization from the participant

<text>: explanation/illustration by researcher

The recording devices used when playing the recordings when transcribing were the same high-quality digital audio recorders that recorded the think-aloud verbalizations. Earphones/headsets were used to listen to the audios and the researcher later segmented the transcriptions. Each protocol was listened to multiple times in order to assure accuracy of transcription and segmentation. When there was doubt about a particular segment, it was replayed until certainty had been achieved. The coding and classification methods and their underlying theory are discussed in the section that follows.

3.18.2 Development of the Coding Scheme

3.18.2.1 Segmentation and Coding Stage

The reason for using think aloud protocols in this study was to answer RQ2, which was expected to give insights into cognitive behavior of test-takers when taking a reading test in PBT and CBT and to generate supporting evidence for the cognitive equivalence of the two confirming/validating hereby the appropriateness of the developed interface for this study's purpose and providing supporting evidence for the test's cognitive validity contributing to its construct validity.

Based on the reviewed studies that evaluated reading and/or test-taking strategies in CBT and PBT, a preliminary coding rubric was generated from these studies to serve as a guide and aid in segmenting and coding the first think-aloud recordings. It was expected that unlike cognitive behavior would be found in certain cases to some extent due to the nature of this study's *test* (i.e. mainly assessing text processing at the local/text level), and *test items*, which was different from the aforementioned studies from where the coding schemes were taken (i.e. open-ended items vs. MCQ's and gap-fill items). For the coding of the initial think-aloud protocol, the longest think-aloud recording was chosen based on the consideration that it potentially contained the greatest number of strategies. This study's reading/test-taking strategy identification process shared Cohen and Upton's (2007) theory, which compared identifying strategies to identifying moves in a discourse genre. When a genre move referred to a specific communicative function within the genre's overall communicative purpose, a strategy referred to a specific choice made by the test-taker in order to facilitate the reading/test-taking activity. Cohen and Upton described the parallelism between the two as follows: 'While a genre move represents a recognizable communicative event characterized by a communicative purpose, a

reading or test-taking strategy represents a specific and recognizable strategic choice made by the subject that is deliberate and purposeful and is intended to facilitate the reading or test-taking task. Furthermore, just as a genre move is identified and mutually understood by members of the professional or academic community in which it regularly occurs, a reading or test-taking strategy can also be identified and mutually understood by expert readers and test-takers' (p.38). Van Someren et al. (2004) agree as they emphasize the importance of the coder having knowledge about the task at hand and its underlying theory. Following this assumption, the strategies in this study were identified based on the strategic function they had in the test, which could be both explicit (i.e. explicitly stated) and implied (i.e. only when obvious), and could be both reading related strategies and test-taking strategies as indicated in section 2.4. The effect this had on segmenting the verbalizations was that a combination of complete thoughts, segments marked by pauses, single words, sentences, and even multiple sentences could signify a single strategy or strategic move. This variation in strategy length and the intermitted jumping between text and questions pointed out by Bax (2013) necessitated episodic coding rather than coding each segment (i.e. complete thought). With this in mind, every segment/combination of segments that appeared to contain a strategy or strategic *move* in the initial think-aloud protocol were then coded with the guidance of the reviewed relevant studies that informed the strategy list of the preliminary coding scheme (see Appendix J, O, and K for the strategies that initially guided this study's coding). The 2 excerpts below are a worked out examples of the think-aloud segmentation and coding process to illustrate how this was done essentially. The first excerpt is a fragment of student number 16 (i.e. S16) initially reading the passage whereas in the second fragment the same student reads and answers item 17 of the reading test.

Excerpt 1

OS1: Paul Newman was born in Cleveland, Ohio, in 1925, and did some acting in high school

OS1: and college, but never seriously (.)

R5: What does this mean? (.)

OS1: never seriously considered making it his future career (.)

OS1: However, after graduating (.)

R8: (graduating)(.)

OS1: he started working in the theatre (.)

R6: What is a theatre? (.)

R11: Is it a museum? (.)

R9: I don't know what it is (.)

OS1: And on several TV shows in New York (.)

OS1: When he was thirty, he went to Los Angeles and made his first film, OK (.)

OS1: It was what he called an uncomfortable start in the movies (.)

R7: This is the film, Good (.)

<END>

As shown in the excerpt above, OS1, which is the overall strategy 'reads passage first and then answers questions' (see Appendix P for the full list of strategies identified in this study), has been repeatedly coded, i.e. whenever the test-taker read the text. This was done for illustration purposes only to indicate that the reading of this test-taker amounted to reading the whole passage and after that started answering the questions. The strategies preceded by R all belonged in the category *initial reading of passage*. [R5], for example, is an exemplification of the strategy 'pauses and thinks about reading', which, as illustrated in the excerpt above, the test-

taker did by pausing and thinking about what the part he had just read meant. [R8] exemplifies the test-taker (correctly) translating the word ‘graduating’ in his L1. Further down, the test-taker read the word theatre and repeated the word theatre to aid in understanding of it. After that, he guessed that it could have meant ‘museum’, by using background knowledge (synonym matching) [R11]. He then indicated that he did not understand the meaning of the word theatre he had just read [R9] after having applied the previously mentioned strategy. As shown by [R7], the test-taker summarized the sentence he had just read by indicating that this was ‘the film’ to aid/confirm comprehension of what he had read. The excerpt below similarly illustrates the strategies used by the same test-taker when answering one of the test items during the reading test (i.e. item 17).

Excerpt 2

Which film made Newman a star? (.)

T7: He was living in Los Angeles when he became engaged to Joanne Woodward. Newman

T7: has been interested in car racing, and in 1979 he came second in the twenty-four hour Le

T7: Mans race. He has a strong social conscience, and has supported causes (.)

EV2: This doesn’t have anything to do with the question (.)

T26: *Which film made Newman a star? (.)*

T6: <scans passage> (.)

T9: The next film he chose was his big break. He played the role of the boxer, Rocky

T9: Graziano in the film ‘Someone up There Likes Me.’ (.)

TW1: It seems that this is the answer because it said ‘big break’, Good (.) <writes answer>

EX1: Let’s go to the next question (.)

<END>

Test-taking strategy 7 [T7] shows that the test-taker started search reading the passage in order to find relevant information/clues to the test item. He realized after having read part of the passage that his search was unsuccessful by indicating that it was not relevant to the question. He then reread the question again [T26], started scanning the passage, most likely for the keyword ‘film’ [T6], and, when found, read the sentence containing the keyword [T9]. He then used a test-wiseness strategy to answer the question as he based his answer on the same possible keyword ‘big break’ relating to the film that made Newman a star [TW1]. After he answered the question the test-taker verbalized his new target, i.e. going to the next question to answer it. The total strategy tokens in both of the excerpts above are seven based on the episode break indications, i.e. (.). For excerpt one they were: OS1, R5, R8, R6, R11, R9, R7 and for excerpt two they were: T7, EV2, T26, T6, T9, TW1, and EX1. As shown in excerpt two, T7 is tagged on each of the first three lines but, as indicated by the first episode break (.), is a single strategy that ends on the third line. The reason for tagging each of the three lines with T7 was for segmentation illustration purposes for the reader only. This coding scheme as illustrated through the two examples above was maintained throughout the segmentation and coding process for all 18 test-takers in this study. It proved to be a solid and reliable scheme, as is further supported through the intra and inter-judge reliability checks later in section 3.18.3.

After all protocols were coded, any identified discrepancies between strategies within the coding were discussed with two colleague until agreement was reached; one being a PhD with similar research interests as the researcher (i.e. cognitive processing in reading), and the other being a vastly experienced ESOL lecturer whose research interests included reading. Cohen & Upton (2007), whose RAs (i.e. research assistants) followed the same procedure when identifying and coding strategies in their study, suggested this to be done to ensure consistency

and accuracy in coding. This process led to a comprehensive list of strategies that was partially established based on the strategies found in the initially coded protocol in both its PBT and CBT combined with that which was discovered in the initial think-aloud report. This list was then used as a template for coding the remaining protocols in both PBT and CBT produced by the remaining eighteen test-takers in this study. As subsequent think-aloud recordings were coded, the list of strategies underwent changes as a number of preliminary found strategies had to be amended, certain strategy combinations ended up having to be collapsed into a single strategy, and a number of strategies anticipated initially from the other studies were not found in any of the protocols. The final comprehensive list of strategies produced in PBT and CBT by the test-takers in this study is enclosed in appendix B. Further reliability checks such as intra-coder reliability and inter-coder reliability were carried out to ensure consistency in segmenting and coding of the verbalizations for this study's think-aloud data to increase the validity of the inferences made from the think-aloud data. The results of these coding validity checks will be discussed in section four on page 202 below.

3.18.2.2 Categorization/Classification Stage

After segmenting and coding of the think-aloud protocols was completed, the coded strategies were assorted into a number of categories. The three overall strategy categories for this study's think-aloud protocols were:

1. Overall test-level strategies (n=6, coded as **OS**)
2. Initial reading of the passage (n=9, coded as **IR**)
3. Test taking strategies (n=30, coded as **TS**)

The reason for initially dividing these categories into three overall categories was that, particularly category one and two were interrelated and could therefore significantly affect each other. For example, if a test-taker would decide to read the passage completely initially in PBT but would go straight to answering the questions in CBT it could lead to a significant difference in strategies employed between the two modes. This could mean that in PBT this test-taker employed, say, 14 strategies when initially reading the passage whereas in CBT he would have none because he did not initially read the passage. By categorizing it in this manner, the two categories could easily be left out (if necessary) to avoid skewing the data in this regard.

After the categories were finalized, they were discussed with fellow PhD-colleagues who had similar research interests and were proficient in the coding and categorization process. Some elements of the categorization in this study were similar to the studies reviewed that focused on cognitive processing in the language learning/testing literature (see section 2.4 of the literature review), however, it is unique due to the nature of the reading tests used (i.e. open-ended questions), the nature of the item types (i.e. inducing expeditious reading behaviour), and the level of test-takers (i.e. lower-level students) in this study, which is different from what has been investigated in the available literature so far. The complete taxonomy for this study's strategies illustrating examples from either the think-aloud reports or accompanying interviews divided into their final categories is presented and discussed in section 5.3 of the results and discussion 2 chapter.

3.18.2.3 Strategy Counting

In order to systematically count the occurring strategies in each protocol, a template was developed including all coding schemes in rows and a column for each item in which they

occurred (see Appendix J). As the passage included 10 items, a colour was assigned for each item to ensure clarity when counting the occurred strategies later on. E.g. Item one was red, item two blue, item three green, etc. Simultaneously, the strategy was coded according to its sequence of occurrence (i.e. 1, 2, 3, etc.) to identify any structural strategy order differences between PBT and CBT that could be of significance. The strategies were then counted after each recording had been reviewed at least two times. All items were local text processing items and largely assessed similar reading skills and strategies (i.e. expeditious reading operations followed by careful reading processes when information had been located). If a strategy was used more than once when solving one item, this strategy was only counted once. This was done to prevent skewed data display on the totals due to multiple strategy occurrences for the same test-taker on that item only. The total strategies for each test-taker in PBT and CBT were calculated afterwards and subjected to further quantitative analyses in SPSS, which is discussed in the second results and discussion chapter (i.e. section 5.3).

3.18.3 Reliability Checks

Scholfield (1995) suggested the following reliability checks to ensure validity and reliability of coding think-aloud protocols: *Intrajudge Reliability*, where the coder/researcher segments and codes the same subset of think-aloud data, and *Interjudge Reliability*, where a second coder/researcher codes the same subset of think –aloud data that the first coder initially coded. In this study only the former was carried out, as at the time, I could not find a fellow researcher/ PhD-student/colleague who was fluent in both Arabic and English *and* had sufficient knowhow in the researcher’s field to warrant for context relevant, reliable coding results. The procedure for carrying out the intrajudge reliability check and its results are described below.

3.18.3.1 Intrajudge Reliability

For this reliability check, the researcher coded two random subsets of think-aloud data (one PBT and one CBT) initially and recoded the two again after a month had gone by without referring back to the initially coded subsets. The counted the number of strategies that were found the same between the two coding sessions (1st and 2nd session) and divided this by the number of strategies found in the first coding session (Scholfield, 1995). Table 25 below shows the total strategies found in the first coding session in the first column (i.e. two subsets), and the number of strategies out of the first coding session that agreed with the second session in the second column. The third column shows the agreement between the two coding sessions in percentages.

Table 25. Intrajudge Coding Reliability Figures

TA Subset	1 st Coding Session	Agreement of Strategy Tokens 2 nd Session	Percentage
1	76	67	88%
2	56	45	80%
Totals	132	112	84%

Table 25 shows that out of 132 strategy tokens, a total of 112 were counted the same over both sessions. This brings it to an overall agreement percentage of 84%, which is an acceptable figure for intrajudge reliability (Scholfield, 1995). This gives additional supporting evidence for the reliability and consistency of the coding procedure and the results obtained from it for this study's think-aloud protocols.

3.18.3.2 Interjudge Reliability

The Interjudge reliability check was the second step to ensure validity and reliability of the coding of the think-aloud protocols. It appears that, from the many PhD-theses the researcher read, normal procedure is to have the researcher and one independent coder code the subset selected by the researcher for this purpose. However, due to the novelty of this study, and the genre theory this study's coding is based on, which proposes that strategic moves are best interpretable by academics in the field of study they occur in (Cohen & Upton, 2006), one additional independent coder coded the chosen subset bringing the total to 3 (i.e. the researcher and two independent coders) in order to increase the reliability of the coding scheme applied in this study. Table 26 below shows the result of the coding agreement.

Table 26. Interjudge Coding Reliability Figures

TA Subset	No. Strategies Coded by Researcher	No. Strategies Coded Same by Independent Coder 1	No. Strategies Coded Same by Independent Coder 2	Total Agreement Researcher & Coder 1	Total Agreement Researcher & Coder 2	Total Agreement Coder 1 & 2
2	56	45	48	80%	85%	93%

As shown in table 26 above, the coding agreement between the researcher and first independent coder is not very high but acceptable at 80% using Scholfield's (1995) calculations. A slightly higher agreement was found between the researcher and the second independent coder (i.e. 85%). Interestingly, the agreement between the two independent coders was significantly higher at 93%. The reason for this would likely be that, although the independent coders' research interests involved L2 reading, they were not as immersed in the realm of strategy

utilization in reading tests as the researcher, which likely resulted in more strategies being identified by the researcher. Nevertheless, the fact that the agreement between the researcher and the two independent coders is acceptable (i.e. 80% and above), and the two independent coders reached an even stronger agreement, further substantiates the suitability and reliability of the coding scheme used in this study.

3.19 Chapter Summary

This chapter discussed the methodology used to address the proposed research questions in the literature review. This study employed an experiment as its main approach in order to investigate the effect of the newly introduced testing mode (i.e. CBT) on test-takers cognitive behavior and performance. Firstly, the study instruments were developed and piloted in two consecutive sessions. The analyses from the first and second session yielded useful results that shaped the eventual instrument materialization for the main study. Each instrument was tested on its reliability as far as possible *a priori*, but also *a posteriori* the main study when required and was comprehensively discussed in order to demonstrate its suitability for this study's context. Furthermore, each stage of the process of analyzing the think-aloud reports from its underlying theoretical model, the process of segmenting the protocols, the coding of the protocols, and the intra and inter-coder reliability checks were discussed sequentially.

The next chapter will consecutively present the results according to its underlying research question (e.g. RQ1, RQ2) to maintain clarity and uniformity in further discussions. RQ1 was addressed using quantitative instruments (i.e. computer familiarity questionnaire, ease of use questionnaire, and the test used in this study) whereas RQ2 was addressed through think-aloud

Chapter3: Research Methodology

recordings and accompanying interviews. The results of the analyses for RQ1 and RQ2 are presented in Chapter 4 (i.e. results chapter) and a discussion of both follows in Chapter 5.

Chapter 4: Study Results Part 1

4.1 Introduction

The purpose of this study was to investigate the effect of interface design on test-takers' performance and their cognitive processes employed when taking this study's reading test in its PBT and CBT form by comparing their (score) outcomes and on (cognitive) processes in both modes. This chapter presents and discusses the analyses and results related to test-taker performance in order to answer the first research question (RQ1) and its accompanying hypothesis:

RQ1. *What is the effect of administration mode on test-takers' performance when taking a lower-level L2 reading test?*

H₀. *There is no effect of administration mode on test-takers' performance when taking a lower-level L2 reading test (PBT=CBT).*

In order to address the research question and its accompanying null-hypothesis above, a number of statistical analyses were carried out sequentially. Firstly, reliability figures of the total test scores on both PBT and CBT are presented. After that, descriptive statistics, score distribution comparisons, correlational analyses, and item-level analyses are given for both modes. The results are anticipated to either confirm equivalence or show discrepancies between the two modes. In addition, if the scores and shapes of the scores are found to be equivalent (i.e. show no significant differences), these outcomes will provide (in part) evidence in support of the validity of this study's test by demonstrating that the test is measuring the same or similar constructs. Whether these constructs are the *appropriate* (reading) constructs is to be further explored in RQ2 through the examination of test-takers' cognitive behavior in PBT and CBT.

This means that RQ2 complements RQ1 by further qualitatively exploring possible significances found through RQ1 in addition to qualitatively describing the underlying processes of test-takers when taking an L2 reading test in both modes.

4.2 Testing Mode Effect on Test-Taker Performance

4.2.1 Reliability Figures PBT and CBT

One of the basic elements that contribute to establishing equivalence between PBT and CBT is showing comparability of reliability figures between the two modes as mentioned by the International Testing Commission (ITC, 2005). In order to explore the possible mode effect on the reliability of the test scores, the internal consistency of both the items on the paper-based test and the computer-based test were measured again by using Cronbach's Alpha after having amended the problematic items (i.e. item 3, 9, and 15) that were found in the pilot study. As shown in table 27 below, the reliability results of the main study sample in PBT exhibits a considerable improvement for overall reliability compared to the initial analysis, as it increased significantly from .773 to .911, which is a .138 difference.

Table 27. Reliability Coefficient PBT Main Study

Cronbach's Alpha PBT	N of items	N of Subjects
.911	30	102

Furthermore, at the item level, item 3, item 9, and item 15 showed significant improvements after amendments were made. Item 3 increased from -.116 to .380, item 9 increased from -.079 to .418, and item 15 increased from -.050 to .433, which reflects the positive impact of the adaptations made to the test's items in addition to the increased sample size. The fact that test-takers from the same sample (i.e. preparatory year students) were used in both reliability checks and the reliability figures of the three problematic items increased in a similar manner in both modes further supports this. The reliability figures in CBT were in line with the PBT as shown in table 28 below. Cronbach's alpha increased to .879, which is also a considerable improvement, compared to its initial reliability measures.

Table 28. Reliability Coefficient CBT Main Study

Cronbach's Alpha CBT	N of items	N of Subjects
.879	30	102

A further reliability check was to calculate the standard error of measurement as suggested by Brown (2005) who proposed the following formula to achieve this:

$$SEM = S \sqrt{(1-r_{xx})}$$

SEM = Standard Error of Measurement

S = Standard Deviation

r_{xx} = Reliability of the Test

The calculated SEM for both PBT and CBT are as follows:

Table 29. Standard Error of Measurement for PBT & CBT

SEM (PBT) $.24194 \sqrt{(1- .911)} = 0.0722$
SEM (CBT) $.22824 \sqrt{(1- .879)} = 0.0794$

As shown in table 29, the SEM for the PBT and CBT are very similar indicating an absence of testing mode effect on SEM when repeating the reading test in a different mode and further supported the initial reliability measures of the PBT and CBT. Establishing an acceptable reliability across modes is essential for the further course of statistical analyses, as the absence of it could indicate serious issues with the items, which would then reduce the validity and reliability of the results of the statistical analyses performed, and the inferences subsequently made from them. Therefore, at this point, the acceptable reliability figures after item amendment and the supporting similar SEM figures for both modes gave confidence to proceed with further describing the test results obtained in both modes by calculating descriptive statistics to describe the score distribution characteristics of both PBT and CBT, which will then form the basis for further statistical procedures such as establishing relationships and/or differences among these score distributions (Bachman, 2004). Either the statistical analysis software package SPSS or a combination of Microsoft Excel and SPSS was used for all quantitative analyses carried out in this study.

4.2.2 Descriptive Statistics PBT

Table 30. Descriptive Statistics for the PBT Reading Test

Descriptive Statistics PBT		
N	Valid	102
	Missing	48
Mean		14.4216
Std. Error of Mean		.71867
Median		14.0000
Mode		8.00 ^a
Std. Deviation		7.25823
Variance		52.682
Skewness		.042
Std. Error of Skewness		.239
Kurtosis		-1.242
Std. Error of Kurtosis		.474
Range		27.00
Minimum		1.00
Maximum		28.00
Sum		1471.00
	25	8.0000
Percentiles	50	14.0000
	75	20.2500

As table 30 above indicates, the test shows an acceptable difficulty level with the mean being around 50% of the total possible score (i.e. $x = 14.4216$ out of 30). The median is slightly lower than the mean (i.e. 14), due to the distribution being slightly positively skewed as the skewness figure shows (i.e. .042). The spread of the scores ranges from 8 (Q1) to 20.25 (Q3) of which 50% has a range of 12.25 (IQR), which would make the semi-interquartile range 6.125 for the PBT. Both the skewness and kurtosis values are between -2 and +2 indicating a reasonably normal distribution (Bachman, 2004).

4.2.3 Descriptive Statistics CBT

Table 31. Descriptive Statistics for the CBT Reading Test

Descriptive Statistics CBT		
N	Valid	102
	Missing	48
Mean		15.0784
Std. Error of Mean		.67798
Median		15.0000
Mode		5.00 ^a
Std. Deviation		6.84725
Variance		46.885
Skewness		.037
Std. Error of Skewness		.239
Kurtosis		-.989
Std. Error of Kurtosis		.474
Range		28.00
Minimum		1.00
Maximum		29.00
Sum		1538.00
Percentiles	25	9.0000
	50	15.0000
	75	21.0000

Like the descriptive statistics of the PBT, table 31 above indicates an acceptable difficulty level for the CBT with a slightly higher mean than the PBT (i.e. 15.078). The median is again slightly lower than the mean (i.e. 15), due to the slightly positively skewed distribution (i.e. .037). The spread of the scores ranges from 1(Q1) to 29(Q3) of which 50% has a range of 12 (IQR), which would make the semi-interquartile range 6 for the CBT, which is very similar to PBT. Furthermore, both the skewness and kurtosis values of the CBT are between -2 and +2 likewise indicating a reasonably normal distribution.

4.2.4 Test of Normality

Because the distribution of the test scores dictates the type of further analyses to carry out (i.e. parametric or non-parametric) and how to interpret subsequent descriptive and inferential data, further investigation of the normality of distribution in both modes was carried out through utilizing the Shapiro-Wilk’s test of normality of which the results are shown in table 32 below.

Table 32. Shapiro-Wilk’s Normality Test PBT & CBT

	Shapiro-Wilk		
	Statistic	df	Sig.
TotalPB	.947	102	.000
TotalCB	.971	102	.026

Table 32 above shows the PB and CB scores are not normally distributed despite earlier descriptive indications of the data suggesting otherwise. The results for the PBT are significant at the .001 level whereas for the CBT they are significant at the .05 level. This is further illustrated in the histograms below visualizing a clear deviation from the normal bell curve in both modes.

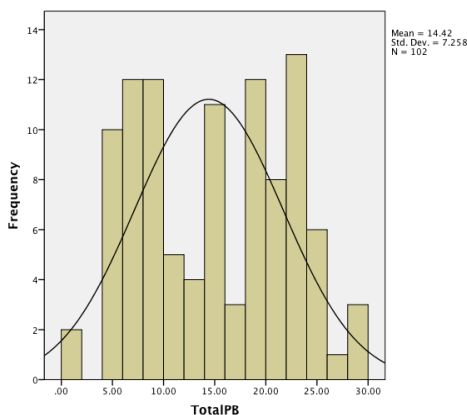


Figure 24. Score Distribution in PBT

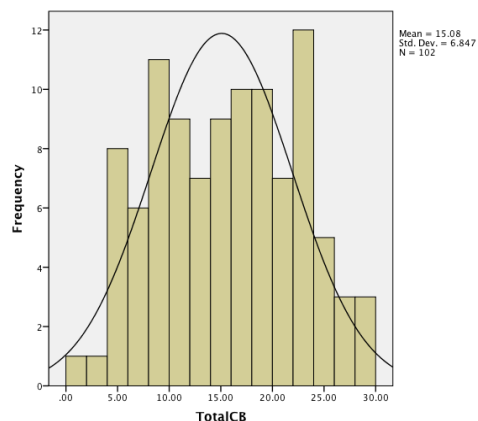


Figure 25. Score Distribution in CBT

Both figure 15 and figure 16 show a clear deviation in score distribution from the indicated

normal bell curve, which appears to be more extreme in the PBT compared to the CBT. Nevertheless, in both cases the divergence is significant and it was therefore required to treat the data as non-parametric when further describing the results and investigating differences and relationships between the two modes.

4.2.5 Score Comparisons PBT and CBT

The next step was to investigate the magnitude of the difference in spread of the scores between both modes through comparing the medians in PBT with CBT. This was done two-fold: First, the score distributions signified by the median and 25% and 75% quartiles are illustrated through a boxplot (figure 20 below) including both PBT and CBT. Next, the Wilcoxon signed rank test was utilized to show the magnitude of the difference between PBT and CBT.

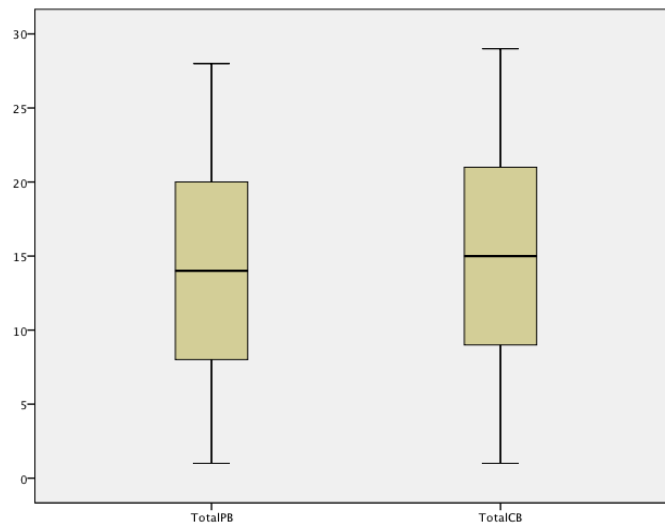


Figure 26. Boxplot Median CBT and PBT Scores

Figure 20 above further illustrates the 25%-75% percentiles that were numerically given in table 30 for PBT, which were between 8 (Q1) and 20.25 (Q3), and for CBT, numerically given

in table 31, were between 9 (Q1) and 21 (Q3). The median (i.e. second quartile) in PBT was 14 whereas in CBT it was slightly higher at 15. This further indicates the similarity of grouping of scores in PBT and CBT.

In order to compare the significance of overall score differences between the two modes for non-parametric data, the Wilcoxon signed ranks test was used to achieve this. Instead of comparing means, which is normally done using a t-test, the Wilcoxon signed rank test examines whether two variables' medians are the same. The results are shown in table 33 below.

Table 33. Wilcoxon Test Results of Difference Total PBT & CBT Scores

Test Statistics^a

	TotalCB - TotalPB
Z	-1.442 ^b
Asymp. Sig. (2-tailed)	.149

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks

As table 33 above shows, no significant difference between the medians of the PBT and CBT was found (i.e. $p = .149$) signifying similar central tendency of the scores in both PBT and CBT modes further supporting that test-takers' performance was not significantly affected by the newly introduced testing mode, i.e. CBT. A further check was to correlate the PBT and CBT scores with each other to substantiate the results above by visualizing the strength of the relationship between the two, which is presented in the following section.

4.2.6 Correlational Analyses PBT and CBT

By investigating the covariance between the two tests (i.e. PBT and CBT) more data is generated on the individual response patterns of the test-takers in both modes; the greater the covariance, the similar the tests would appear to be. Correlational analyses were run on the two tests in order to achieve this. The appropriate test to obtain correlational values for non-parametric data was the Spearman rho correlation test whose results are shown in table 34 below.

Table 34. Correlations CBT and PBT Main Study

Spearman's rho Correlations

		TotalPB	TotalCB
TotalPB	Correlation Coefficient	1.000	.785**
	Sig. (2-tailed)	.	.000
	N	102	102
TotalCB	Correlation Coefficient	.785**	1.000
	Sig. (2-tailed)	.000	.
	N	102	102

** . Correlation is significant at the 0.01 level (2-tailed).

Table 34 shows a moderately high and significant correlation between the PBT and CBT (i.e. 0.785) at the 0.01 level (i.e. $p < 0.01$). This provides further evidence in support of the validity of the newly introduced testing mode (i.e. CBT) as it indicates the degree of parallel behavior for each student on both modes. These results, in addition to the reliability figures and descriptive statistics for each mode, and comparison of overall performance on both modes further confirmed the suitability of the study tests used for this study's purpose. These results also partially support the interface design discussion in the second pilot study whose conclusions would be interpretable provided the test used showed acceptable reliability, and preliminary

evidence indicating that similar constructs were measured between both modes (i.e. construct validity).

Following the results obtained from the descriptive statistics, median comparisons, and correlational data on the total scores for PBT and CBT, the next step in further investigating the reliability and validity of the test scores was to investigate whether there existed any significant differences in scores for each item individually (i.e. the item level), which is discussed in the following section.

4.2.7 Item Performance Comparison PBT and CBT

To examine whether there are any discrepancies on overall performance between the two modes at the item level, the percentage of students answering the items correctly in PBT and CBT was compared for each of the thirty items of the reading test. Because the data are non-parametric, the Wilcoxon's signed rank test was the method of choice to examine the significance of differences on item performance between the two modes as an alternative to the paired samples t-test, which is applicable to parametric data.

Table 35. Comparison of Item Level Performance between PBT and CBT

Item	PBT Total %	CBT Total %	Wilcoxon		Item	CBT	PBT	Wilcoxon	
			p	z				p	z
1	88.2	91.1	.366	-.905	16	67.6	60.7	.209	-1.257
2	16.6	25.4	.039	-2.065	17	24.5	26.4	.617	-.500
3	61.7	58.8	.602	-.522	18	48.0	50.0	.715	-.365
4	50.0	48.0	.715	-.365	19	36.2	33.3	.602	-.522
5	61.7	57.8	.346	-.943	20	41.1	34.3	.127	-1.528
6	28.4	36.2	.157	-1.414	21	88.2	91.1	.467	-.728
7	77.4	85.2	.088	-1.706	22	39.2	35.2	.394	-.853
8	38.2	48.0	.059	-1.890	23	64.7	61.7	.564	-.577
9	80.3	87.2	.144	-1.460	24	35.2	26.4	.139	-1.480
10	55.8	51.9	.414	-.816	25	13.7	18.6	.225	-1.213
11	42.1	41.1	.853	-.186	26	32.3	41.1	.139	-1.480
12	21.5	22.5	.782	-.277	27	32.3	34.3	.732	-.343
13	39.2	40.1	.857	-.180	28	48.0	47.0	.857	-.180
14	80.3	69.6	.016	-2.400	29	54.9	50.9	.414	-.816
15	61.7	68.6	.209	-1.257	30	45.0	30.3	.131	-1.512

As shown in table 35 above, the significance of the differences found between the two modes is reflected through the calculated p-value of the scores for each item pair (i.e. item 1 PBT & item 1 CBT, item 2 PBT & item 2 CBT etc.) No significant differences were found for the majority of the test items at the $p < .05$ level. Only performance differences on item 2 (i.e. $p = .039$) and item 14 (i.e. $p = .016$) turned out to be statistically significant, which could have been an indication of either a case of construct irrelevant variance or construct underrepresentation

(Messick, 1998). Item 2 was part of passage 1, which included items 1-10. Item 14 was part of passage 2, which included items 11-20. As mentioned in chapter 3, the time allocated for the think-aloud sessions was 35 minutes for each test-taker in each mode. As including both passages in the think-aloud would have resulted in too long of a verbalization from the test-takers, a decision had to be made on which of the two passage containing the item that was significantly differently performed on would be included. The justification for choosing the passage containing item 14 over passage 1 containing item 2 was twofold:

1. Performance in CBT for item 14 was significantly lower than in PBT as opposed to item 2 where the PBT scores were lower than the CBT scores
2. Magnitude of significance (0.16 compared to 0.39)

Both think-aloud verbalizations and retrospective interviews are anticipated to further shed light on the cause of the observed significance on this particular item, which is further discussed in chapter 5.

4.2.8 Post-Test Questionnaire

Test takers' views or viewpoints on how they experienced the test-taking process or certain features of the test could be a useful addition in test validity studies (e.g. Bachman & Palmer, 1996; American Educational Research Association, 1999; McNamara, 2000). Urquhart & Weir (1998) mentioned that structured feedback from test-takers on study instruments, the test tasks, and texts involved could give an important broad view on how the study's sample responds to the test and the possible impact of certain test features on test-takers in general. Finding this out was particularly important in this study's context (i.e. Saudi Arabia) as CBT is still in its infancy there and, due to the rapid advancements in technology in addition to the

ample advantages CBT has over PBT (e.g. Funke; 1998, Butcher, 1987; Butcher et al., 2000; Mason, 2001; Roever, 2001) particularly in large-scale assessments, it is likely to eventually become an integral part of the assessment system in public and private educational establishments in the future (e.g. Al-Amri, 2008). In language testing research, test-takers' views have been gauged ample times for these purposes, i.e. to corroborate or triangulate statistical data analyses (e.g. Boo, 1997; Kirsch et al., 1998; Taylor et al., 1998; Gorsuch, 2004; Al-Amri, 2008; Flowers et al., 2011). This study likewise administered a post-test questionnaire to the test-takers, which was given to gauge their perceived ease of use of the CBT by asking questions about certain features of the interface compared to the PBT. The questions were divided into two parts; the first part elicited students' general perception of the ease of use of a number of key features of the CBT such as scrolling, navigating, readability of text, and answering items whereas the second part required students to express their preference for one mode over the other in relation to these aspects. The purpose of this questionnaire was threefold:

1. To initially gauge an overall impression on the potential feasibility of using CBT in the target context from a test-takers' perspective.
2. To substantiate both performance analyses and processes analyses provided equivalence was established in both by a substantial number of test-takers.
3. To guide further possible interview questions for the think-aloud sample. For example, if a substantial number of students would have indicated that they had had significant issues with key features of the interface despite the amendments that were made, the think-aloud reports themselves could then have given further insights into how this could possibly have affected cognitive behavior. Table 36 below shows the results of the post-test questionnaire.

Table 36. Descriptive Summary of Participants' Ease of Use Questionnaire Results Part I

Question	Response Options	Percentage
1. The text and questions on screen were easy for me to read.	1.Strongly Disagree	10.78%
	2.Disagree	9.80%
	3. Neither	21.56%
	4.Agree	19.60%
	5.Strongly Agree	38.23%
2. The text and questions on paper were easy for me to read.	1.Strongly Disagree	15.68%
	2.Disagree	15.68%
	3. Neither	24.50%
	4.Agree	15.68%
	5.Strongly Agree	28.43%
3. The size of the computer screen was big enough.	1.Strongly Disagree	5.88%
	2.Disagree	4.90%
	3.Neither	11.76%
	4.Agree	26.47%
	5.Strongly Agree	50.98%
4. It was easy to navigate through the test using the navigation buttons.	1.Strongly Disagree	5.88%
	2.Disagree	5.88%
	3.Neither	18.62%
	4.Agree	14.70%
	5.Strongly Agree	54.90%
5. Using the scrolling feature was not problematic for me.	1.Strongly Disagree	8.82%
	2.Disagree	8.82%
	3.Neither	17.65%
	4.Agree	21.56%
	5.Strongly Agree	43.14%
6. I think I did better on the computer-based test than on the paper-based test. Why?	1.Strongly Disagree	3.92%
	2.Disagree	8.82%
	3.Neither	17.65%
	4.Agree	20.58%
	5.Strongly Agree	49.02%

As shown in table 36 above, around 60 % of the test-takers found it easy to read the text and test questions on screen whereas only 20% indicated difficulties with reading on screen. Around 20% reported that it was neither difficult nor easy to read text and questions on screen indicating neither a positive nor a negative impact for this group. This would mean that at least

80% of the total test-takers perceived that reading on screen did not have a negative impact on them when taking the reading test on computer. As for question 2, the paper-based experience was slightly different with a total of around 70% of the test-takers not experiencing a negative impact of reading on paper and the remaining 30% perceiving that reading on paper was not easy indicating some difficulty in this regard. Around 90% of the test-takers were satisfied with the screen size as opposed to 10% who indicated that the screen should preferably have been larger than the 17" screen used in this study's CBT. Around 88% of the test-takers did not experience any problems navigating on screen of which 70% found it easy as opposed to around 12% who indicated some difficulty with navigating on screen. 18% of the test-takers reported some difficulty with scrolling whereas 82% did not experience this difficulty. Around 70% of the test-takers felt they performed better on the CBT as opposed to only around 13% who reported they thought they did better on the PBT. Around 17% reported they did not perform better on either one. Table 37 below further illustrates how the test-takers in this study experienced the features of both modes in relation to each other.

Table 37. Summary of Participants' Ease of Use Questionnaire Results Part 2

Questions	Options %		
	Computer	Paper	No Diff.
7. In which test was the text easier to read?	47%	29%	24%
8. Which test did you prefer taking? (Why)?	51%	37%	12%
9. In which test was it easier to write down answers? (Why)?	65%	30%	5%
10. In Which test was it easier to change answers? (Why)?	83%	6%	11%
11. In which test were the reading passages easier to navigate through?	72%	21%	7%

As shown in table 37 above, about half of the total participants felt that the text on CBT was easier to read than on the PBT compared to around 30% who felt PBT was easier to read. About a quarter of the participants reported it was the same on both modes. As for test mode preference, about half of the test-takers preferred taking the CBT over the PBT whereas 37 % reported the opposite. Slightly over 10% of the total participants did not prefer either mode to the other. A large difference was found between test-takers' experience when writing down answers, as 65% reported that it was easier to write down answers on computer than on paper. Only 5% reported no preference for either of the two modes. The difference found was even greater with regards to changing answers, as 83% found this to be easier on computer than on paper as opposed to only 6% who found changing answers on paper was the easier of the two. 11% did not prefer either mode to the other in this regard. As for navigating through the passage, 72%

found this to be easier in CBT compared to 21% that found this to be the case in PBT. Only 7 % of the test-takers did not experience a difference in difficulty here.

4.3 Discussion of Results Part 1: Test-Takers' Performance in PBT and CBT

4.3.1 Comparability of Scores in PBT and CBT

Results obtained from the analyses of this study's test (i.e. reliability, mean comparisons, correlational analysis, and item analysis) appear to be in agreement with a large number of studies that examined the effect of testing mode on reading test results in terms of the non-significance of the differences between the two modes (for compatibility reasons with regards to computer devices only studies from late 20th and 21st century are discussed).

4.3.1.1 Reliability in PBT and CBT

Various studies that investigated score comparability employed reliability analyses as an initial indication of item variance/invariance (e.g. Boo, 1997; Hagler et al., 2005; Coles et al., 2007; Al-Amri, 2008). This study's internal consistency measures were .911 for the PBT and .879 for the CBT, which were both well above the .8 recommended in a language testing context (e.g. Bachman, 2004). A significant increase in reliability figures was achieved through:

1. Increasing the number of test items (i.e. n=30)
2. Increasing the number of participants (i.e. n=102)
3. Amending test-items that appeared not to work sufficiently well (i.e. item 3, 9, and 15)
4. Amending features of the Interface through piloting and usability testing

The reliability results in combination with SEM measures in both modes supported the reliability of the items representing the overall construct of (expeditious) reading in this case. In

addition, they provided initial supporting evidence for the absence of an effect of the newly introduced testing mode (i.e. CBT) on test-item internal consistency on the CBT. Item correlational analysis, which produced a moderately high and significant correlation of .785 corroborated these findings and further implied the similarity of the constructs measured, provided the TA protocols showed equivalent, construct relevant processing in both modes to support this. Related research that included effect of testing mode on its reliability in PBT and CBT was that of Boo (1997). In his study (that included a reading section), which also involved other variables such as computer attitudes, familiarity, and anxiety, no significant effect of the newly introduced testing mode (i.e. CBT) was found on the test's reliability, which was the same for this study. No interactions of computer familiarity and attitudes towards using a computer were detected either, which provided further supporting evidence towards the construct validity of the test used in his study. Other studies such as Choi et al. 's (2003) showed relatively low internal consistency figures on the reading comprehension section of their proficiency test of .755 for the PBT and .668 for the CBT, in particular for the CBT. Choi et al (2003) argued that the lower figures for the CBT reading section of their test could be attributed to the relatively small number of items in the test. However, internal consistency increased to 9.3 after they had corrected for measurement error. Interestingly, they achieved an initial internal consistency of over .8 for both PBT and CBT on the listening component, which they argued was due to the advantages of that same CBT (e.g. through using visual cues). Likewise, Al-Amri's (2008) three tests in PBT and CBT had an internal consistency of .57 on PBT 1, .65 on PBT 2, .70 on PBT 3 and .58 on CBT 1, .64 on CBT 2 and .65 on CBT 3. These figures were even lower than Choi et al.'s (2003) initial reliability measure for the most part, which Al-Amri (2008) argued was due to the nature of the tests used in his study i.e. institutional tests and/or the low number of test items

(15, 14, and 15 for test 1, 2, and 3 respectively). He then recalculated the internal consistency combining the three tests and found .535 on PBT and .707 on CBT. Although this is a slight improvement (which was more prominent in CBT), it is still on the lower end, which therefore could be due to the nature of the test itself, since theoretically, the number of items increased by doing so. This study likewise used an institutional test in the same target context involving similar participants (i.e. Saudi Arabian preparatory year students) but showed much higher reliability figures, which possibly indicates a different main cause for the lower reliability of his study's tests and perhaps other contributing factors but not necessarily the tests themselves.

4.3.1.2 Test-Takers' Performance in PBT and CBT

The descriptive statistics further shed light on score distributions/ shapes of test-takers' scores in PBT and CBT. Score medians for both modes were similar with 14 in PBT and 15 in CBT. Although the overall median for the CBT was higher than the PBT by about 4%, the Wilcoxon signed ranks test for related samples showed that this difference was not significant between the two modes (i.e. $p = .149$) further evidencing the absence of an effect of the new testing mode on test-takers' performance. Performance comparison analyses at the item level in PBT and CBT showed a significant difference between the two modes on two items, i.e. item 2, and item 14. Item 14 is further investigated as in this case the CBT performance was significantly lower than PBT whereas for item 2 it was the opposite. Think-aloud reports are expected to further reveal whether possible underlying cause(s) are related to the interface or others through investigating test-takers' cognitive processing in the two modes.

The non-significant overall performance difference found in this study was in accordance with several of the more recent studies that investigated reading performance in PBT and CBT.

Fitzpatrick & Triscari's (2005) study that included a comparable number of high school students (n=2205) yielded similar results, as they did not find a significant difference in scores between the national reading test they administered in its PB and CB form. Higgins et al. (2005) who also used a national reading test for the 219 4th grade students in their study revealed similar results (i.e. non-significant difference between PBT and CBT). Green & Maycock (2004) who compared the PBT and CBT version of the IELTS (which included a reading section) concluded that both forms for their population sample were equivalent. Blackhurst, (2005) who later investigated part B of the IELTS test Green & Maycock reported on, likewise confirmed comparability between the two modes.

On the other hand, others did report a negative effect of CBT on test-takers' reading performance, which was statistically significant. For example, Choi et al. (2003) who used a proficiency test assessing the four language skills (which included reading) investigating test construct validity found a significant effect of CBT on the listening, vocabulary, and reading section of their test. They attributed this difference to eye fatigue, which appeared to have been a commonly drawn conclusion when significant (negative) effects of CBT were found around that time (e.g. Boo, 1997; Larson, 1999; Choi, 2000). The issue that exists is that there is little empirical research that provides data to support this claim, which, although eye-fatigue (also referred to as eyestrain) is a reality in CBT or computer use in general, makes it somewhat subjective as an argument for being the cause for inferior performance in CBT compared to PBT. This can be seen in a number of the studies where participants reported eye fatigue but at the same time did not find any significant effect on overall performance (e.g. Blackhurst, 2005; Darroch et al., 2005; Al-Amri, 2008).

Furthermore, eye fatigue, which in reality is fatigue of the iris muscle (e.g. Campbell and Durden, 1983; Taptagaporn and Saito, 1990) is in the field of science commonly thought to be relieved by simply looking away from the screen at a distant object on regular intervals i.e. every 25-30 minutes (Cheu, 1998). Nevertheless, based on the typographical elements review in chapter 2 (section 2.6.7.1), the high contrasting of text and background, as indicated by Galitz (2007), would be a more likely alternative explanation as especially the studies carried out in the 21st century that reported eye fatigue had used computers with higher resolutions. This could have resulted in eye fatigue due to this combination of higher resolution and (too) sharp contrast (i.e. black text on white background), which was avoided in this study's interface design by using less contrastive colour combinations.

The only study that, like this study, found no overall performance differences yet some at the item level was Pommerich's (2004) who made amendments to the interface in-between the two testing cycles because of this. Pommerich (2004) mentioned that one of the possible underlying causes was impediment of spatial organization in the passage by having to scroll as opposed to PBT where because of the absence of this it was easier for test-takers to locate relevant information in the passage. This is unlikely to be the underlying cause for the observed difference with item 14 in this study because the text passage is significantly shorter (i.e. 303 words) which minimized scrolling therefore limiting effect on spatial memory. However, the possible underlying cognitive aspect will be further discussed in the chapter that follows to shed more light on whether the possible cause for this difference is to be attributed to the testing mode. The section that follows summarizes and concludes this chapter.

4.4 Summary

Overall, the presented findings related to RQ1 combined suggest an acceptable reliability and validity of the tests and the test items used in this study for this particular purpose in this context. Item reliability was preliminarily established through item-total correlations in both modes and descriptive statistics were given for both modes to initially describe the score distribution characteristics. Although the initial impression based on the descriptive statistics indicated reasonably normally distributed data, further analyses revealed that this was not the case. Descriptive statistics on score grouping (i.e. median) and variability/ dispersion (interquartile range and semi-interquartile range) suggested similar distribution between the two modes. Non-parametric comparison analyses confirmed that there was no statistically significant difference between the PBT and CBT score distributions. Further correlational analyses supported these findings by indicating a strong relationship between PBT and CBT. These results provide supporting evidence towards the absence of test-mode effect on test-takers' performance and would suggest mode comparability. However, two significant differences were found at the item level of which one favoured PBT (item 14) whereas the other favoured CBT (item 2). For the importance to this study, i.e. investigating the suitability of the developed interface for this study, item 14 will be further investigated in the think-aloud study in chapter 5, as this item *negatively* affected CBT performance as opposed to item 2, which favoured CBT. Investigating test-takers' processes could reveal whether this difference can be attributed to the interface itself or possibly to a different source.

Chapter 5 Results & Discussion, Part 2a: Comparing Processes

5.1 Introduction

This chapter discusses the cognitive processes test-takers utilized when taking this study's L2 reading test in both PBT and CBT. In order to investigate the effect of interface design on test-takers cognitive processes, the following research question and accompanying hypothesis were formulated:

RQ2. *Is there any effect of administration mode on test-takers' cognitive processes when taking a lower-level L2 reading test?*

H₀: *There is no effect of administration mode on test-takers' cognitive processes when taking a lower-level L2 reading test.*

Test-takers' verbalizations in this study in both PBT and CBT were recorded, transcribed, segmented, coded, and analyzed in order to look in-depth into test-takers' processing when taking the reading test. This part of the study was largely exploratory, as the cognitive processes of lower-level students involved in performing expeditious reading tasks has not been investigated in this manner as of yet to the researcher's knowledge other than by Bax (2013), however, he used a different instrument to identify processes involved when taking a L2 reading test (i.e. eye tracking technology). In presenting the results, a category-by-category arrangement is maintained as mentioned in chapter 3 informed by previous studies such as Cohen and Upton (2007), Kobrin (2000), and Al-Amri (2008). Each strategy category with its included strategies is described in a table defining each strategy to which the researcher assigned a unique code. The test-taking strategies that were found (i.e. TS) were further categorized post-hoc according to their occurrence in the think aloud protocol. This was done for discussion purposes only and

does not carry further significance to the study's outcomes. The total strategy categories were eleven and were divided as follows:

Category 1: Overall Test-Level Strategies (OS)

This category includes the overall strategies used by a test-taker when reading the text. For example, when a test-taker read the whole passage and then started reading the questions and answering them, this strategy was categorized as an OS-strategy (i.e. OS1).

Category 2: Initial reading of the passage (IR)

When a test-taker started reading the passage initially, i.e. before answering the questions, the strategies identified through the think-aloud verbalizations during reading of the passage initially were categorized as IR-strategies.

Category 3: Test-Taking Strategies Related to Reading of Questions (TS)

Any verbalizations during the reading of the questions by the test-takers were placed in this category. For example, when the think-aloud verbalization revealed that a test-taker read the question and then read it again, this strategy was included in this category.

Category 4: Test-Taking Strategies Related to Reading of Passage (TS)

Logically following the reading of the question, test-takers read the passage to search for the answer. Strategies that were utilized such as scanning and search reading were included in this category. Think-aloud verbalizations indicating these strategies were allocated to this category in among others.

Category 5: Test-Taking Strategies Related to Aiding in Answering Questions (TS)

Chapter 5: Results & Discussion 2

Verbalizations when answering the test items were allocated to this category. For example, when a test-taker guessed the answer, used his background knowledge or answered the question from memory, these strategies were included here.

Category 6: Test-Taking Strategies related to Items after having answered them (TS)

When test-takers' verbalizations indicated that they went back to the question after having answered it, the strategies that were utilized were allocated to this category. For example, there were several instances where test-takers discovered an answer to a previous question later on and then returned back to that question to correct it, which fell into this category.

Category 7: Supporting Strategies (SUP)

This category included strategies such as taking notes or underlining information in the text to aid them when having to return to it for clarification purposes for example. These strategies were not verbalized but observed by the researcher during the test administration.

Category 8: Executive Strategies (EX)

Verbalizations indicating the target of search were included in this category. When a test-taker monitored his location in the passage, i.e. using his overall knowledge of the location of certain information within the text, it was also assigned to the executive strategies category.

Category 9: Evaluative Strategies (EV)

Verbalizations indicating the Evaluation of possible answers to questions within the text or successful/unsuccessful searches were included in this category. Evaluating the meaning of a word or phrase read was also one of the evaluative strategies assigned to this category.

Category 10: Inferencing Strategies (INF)

Verbalizations indicating pronoun referencing, inferring word meaning from context or background knowledge were assigned to this category.

Category 11: Affective Strategies (AFF)

Verbalizations indicating self-motivation or any others of similar essence were considered as affective strategies.

Excerpts from the think-aloud reports representing the strategies as they occurred in the think-aloud recordings are given for illustrational purposes. Additional comments are given by the researcher when deemed necessary. As part of RQ2 was to explore all strategies used by the test-takers, and, 48 out of 50 of the found strategies were found in PBT as well as in CBT, it is not specified in the examples in which of the two testing modes the strategy was used. Only when a particular strategy is unique to one of the two testing modes (i.e. PBT or CBT) it is specified in the example. Before discussing the strategy frequencies and differences in both modes, an overview of the total number of strategies (i.e. types) and occurrences (i.e. tokens) of these strategies in both modes are given.

5.2 Overview Overall Strategy Types and Tokens in PBT and CBT

Table 38 below shows the total number of strategy types found in PBT and CBT. The process of counting the strategies was based on the occurrence of the strategy type first, each receiving a unique strategy code as explained in section 3.9.4.3.5, chapter 3. After that, the number of occurrences for that particular strategy type was calculated for all participants (n=18), which lead to a total of x-number of strategy tokens for each participant. The strategy

occurrences might seem to be at the lower end compared to other studies, however, this study's focus was on expeditious reading operations only, which likely exclude higher-level text processes based on the nature of the test items. Section 5.7 will further validate this through a comparison of reading types elicited by the test items between PBT and CBT.

Table 38. Overview Strategy Types and Strategy Type Tokens PBT and CBT

Testing Mode	Total Identified Strategy Types	Total Strategy Type Frequencies	Sig. of Diff. Total PBT & CBT Frequencies
PBT	45	913	t= 1.762
CBT	42	805	p= .096

As shown in table 38 above, the total number of strategies found in the PBT was forty-five whereas the CBT yielded a total of forty-two test-taker strategies. The strategy types identified were the same for forty strategies in both modes. As for the frequency of the strategy types in the paper-based test, a total of 913 with a mean of 50.72 were identified compared to 814 in the computer-based test showing an average of 44.72 respectively. The paired sample t-test that was carried out showed a non-significant difference between the frequency totals of strategies used in both modes (i.e. p= .096). The total number of strategy tokens for each test-taker was calculated in PBT and CBT using the method explained in detail in the methodology chapter (i.e. based on episodic segmentation). An overview of the total strategy tokens identified for each test-taker in both modes is shown in table 39 below.

Table 39. Total Strategy Tokens each Participant in PBT and CBT

Participant No.	Total Strategy Tokens PBT	Total Strategy Tokens CBT
1	32	34
2	33	37
3	62	56
4	39	40
5	53	36
6	25	24
7	69	47
8	41	34
9	27	26
10	50	33
11	57	55
12	66	57
13	41	40
14	71	43
15	70	48
16	49	32
17	53	62
18	75	101
Totals	913	805

Table 39 above shows the total number of overall strategy tokens used by each participant on the PBT and CBT- test. The number of strategy tokens may seem on the lower side, however, the study's test-passage used for the think-aloud consisted of only ten items (i.e. questions). Furthermore, the items assessed local expeditious and local careful reading processes, which was expected to rule out usage of more global level processing from the test-takers. The lowest number of observed strategy tokens used by participants on the PBT was 25 (i.e. participant 6), which is 2.7% of the total strategy tokens in PBT whereas the lowest number on the CBT was 24 (i.e. participant 6), which is 3% of the total strategy tokens in CBT. The highest

number of observed strategy tokens used on the PBT was 75 (i.e. participant 18), which is 8% of the total strategy tokens in PBT, while on the 101 strategy tokens were the largest identified number in CBT (i.e. participant 18), which is 12.5% of the total strategy tokens in CBT. The paired samples t-test on the total strategy tokens identified in both modes revealed no significant differences between PBT and CBT. Furthermore, correlational analyses showed a reasonably high and significant correlation of .692 at the .01- level, which, taken together with the paired sample results, preliminarily suggests an absence of an effect of the CBT on overall cognitive processes used by test-takers. However, significant differences at the individual strategy-level could still exist between PBT and CBT, as exemplified in Pommerich's (2004) study, for example. Therefore, paired samples t-tests were utilized on every strategy in each mode. The results are presented below assorted by coding category as illustrated in chapter 3 (p. 223). The table for each category includes the mean frequency of strategy occurrence in PBT and CBT, its standard deviation, and paired-samples t-tests' results. For uniformity in discussing the strategies, the same key as detailed in chapter 3 (p. 324) should be adhered to in order to interpret test-takers' verbalizations in the examples given in this chapter, and, likewise, in subsequent discussions.

5.3 Overview Strategies by Category

5.3.1 Category 1: Overall Test-Level Strategies

This category includes the overall strategies utilized at the test level, i.e. how the test-taker approached the test in terms of reading of the text passage and answering associated test items, which was based on Kobrin's (2000) study (see Appendix J). As this strategy could only

be applied once in each mode, a total of eighteen occurrences in each mode were found parallel to the number of test-takers (i.e. n=18), which is 2 % of the total strategy tokens in PBT and 2.3 % of the total strategy tokens in CBT. The four overall test-level strategies identified were as follows:

OS1: Reads passage first then answers questions: e.g. (Appendix M).

OS2: Starts to read passage then skips to questions before finishing reading: e.g. Paul Newman was born (.) his future career (.) When did Newman first work in the theatre? (.) [OS2].

OS3: Reads all questions first, then reads passage, then answers questions: e.g. Where*=when did Newman first work in the...? (.) What is the name of the? (.) car racing start? (.) just taking a look at the questions (.) [OS3].

OS4: Reads and answers one question at a time: e.g. When did Newman first work in the theatre? (.) in 1925 (.) he start*=started working in the theatre and several TV (.) Next (.) What is the name of Newman's company? (.) [OS4].

5.3.1.1 Descriptive Statistics and Paired Samples T-test's Results

Table 40. Descriptive Statistics Strategy Category 1 PBT & CBT

Strategy (Code)	N	Mean PBT	S.D PBT	Mean CBT	S.D CBT	Paired Sample t-Test	
						t	p
OS1	18	.39	.502	.33	.485	.566	.58
OS2	18	.33	.485	.28	.461	.566	.58
OS3	18	.06	.236	.00	.000	1.00	.33
OS4	18	.22	.428	.39	.502	-1.84	.08
Valid N	18						
Total F.		18		18			

As shown in table 40, the highest frequency of participants that chose to read the complete passage first and answer the questions after having read the passage completely was found in PBT (i.e. OS1) whereas this strategy was used second most frequent in CBT. At a slightly lower but similar frequency, participants started to read the passage first, and then skipped to answering the questions before they had finished reading the whole passage in PBT (i.e. OS2), which was the third most frequently used strategy in CBT. The third most common overall strategy participants used in PBT was to read one question and answer one question at a time (OS4), but it was the most frequently used strategy in CBT. The strategy that was only used once (in PBT) was reading of all questions first, then the passage, and then answering the questions and appeared only once in the paper-based test (i.e. OS3). The participant who used this strategy happened to be of the higher achieving students assessed in the think-aloud study. He read out every question in the test first briefly and started reading the passage. When he had finished reading the passage, he answered the questions one by one, though he did refer back to the text whilst answering the questions.

The paired samples t-tests' results show that the overall strategies used by the participants did not differ significantly between the PBT and CBT testing modes (i.e. OS1 $p = .58$, OS2 $p = .58$, OS3 $p = .33$, OS4 $p = .08$). This indicates that, generally, when a participant employed an overall strategy in PBT he did so in CBT. The strategies were mostly divided over OS1, OS2, and OS4 with similar frequencies. These statistics further indicate that testing mode did not have a significant effect on the overall test approach by the participants (i.e. students did not alter the overall approach to completing the reading test significantly between the two modes).

5.3.2 Category Two: Strategies related to Initial Reading of the Passage

This strategy had a total of 32 tokens in PBT, which is 3.5% of the total strategy tokens in this mode. The total strategy tokens in CBT were 25, which is 3.1% of the total strategy tokens in this mode. Worth to note is that because some of the test-takers read the whole passage first in one mode and did not do this in the other mode, the strategy token data in this category could have possibly caused significant differences between the two modes at the strategy level as a result. However, only one student opted not to initially read the passage in one mode and did so in the other mode, which resulted in no significant differences at the strategy level. The reading strategies identified in this category were as follows:

IR1: Reads whole passage carefully (enclosed in appendix M) [IR2].

IR3: Reads a portion of the passage carefully (enclosed in Appendix N) [IR3].

IR5: Pauses and thinks about reading: e.g. All the money from Newman's (...) (In the beginning it is asking about his work in the theatre) (.) [IR5].

IR6: Repeats word(s)/phrase(s)/sentence(s) to aid in comprehension: e.g. they have co-starred in six films every*(= ever) since the film Winning (.) since the film winning, OK (.) [IR6].

IR7: Paraphrases/summarizes portion(s) of reading passage to aid in comprehension: e.g. It was what he called an uncomfortable start in the movies (.) (this is the film, good) (.) in the role of a Greek slave (.) [IR7].

IR8: Translates word(s)/phrase(s)/sentence(s) to aid in comprehension: e.g. However, after graduating (.) (graduate) (.) he started working in the theatre and on several TV shows in New York (.) [IR8].

IR9: Indicates that he doesn't understand a word/phrase meaning in passage: e.g. He was living in Los Angeles when he became engaged to Joanne Woodward an actress he had first known (.) Actress? (What is this word?) (.) in New York (.) [IR9].

5.3.2.1 Descriptive Statistics and Paired Samples T-test's Results

Table 41. Descriptive Statistics Strategy Category 2 PBT & CBT

Strategy (Code)	N	Mean PBT	S.D PBT	Mean CBT	S.D CBT	Paired Sample t-Test	
						t	p
IR1	18	.39	.502	.28	.461	1.458	.163
IR3	18	.22	.428	.22	.428	.000	1.000
IR5	18	.39	.502	.33	.485	1.000	.331
IR6	18	.28	.461	.17	.383	1.458	.163
IR7	18	.22	.428	.11	.323	1.458	.163
IR8	18	.28	.461	.22	.428	1.000	.331
IR9	18	.00	.000	.06	.236	-1.000	.331
Valid N Total Fr.	18	32		25			

A significant number of the participants read the whole passage carefully initially (when applicable) (i.e. IR1). This frequency is inevitably interrelated with the preceding overall reading strategies to some degree as logic would assume that students who employed OS2 or OS4 would not have read the whole passage carefully but rather employed IR3. The students that did employ IR1 would therefore have been more likely to use overall strategies OS1 or OS3. These frequency statistics further indicate that, when a participant did read the whole passage, he read it carefully as opposed to reading through the passage rapidly, which was a strategy that was not detected in the initial reading of the passage. This was further indicated in the post-test interviews, where test-takers mentioned on a number of occasions that the main reason for using this strategy was to get a good understanding of what the paragraph was about before attempting

to answer the questions. During the initial reading of the passage (whether the whole passage or merely a portion of it) what happened most frequently was that the student paused for a moment and thought about reading.

Both repeating words/phrases (IR6) and translating words/phrases (IR8) in order to aid in comprehension occurred in similar frequencies (i.e. IR6: PBT=.28, CBT=.17, IR8 PBT=.28, CBT=.22).

These strategies were used fairly regularly amongst the test-takers. However, at times a word was reread because the student felt that the pronunciation of the word was incorrect. These were not counted as IR6, as there was typically no indication from the student's side that it might or might not have helped him with the actual understanding of that particular word/phrase.

Many test-takers translated words or phrases in a sentence into Arabic (sometimes wrongly) in order to increase their comprehension whether they were on the higher end performance-wise or on the lower end.

Paraphrasing of sentences/phrases in order to aid in understanding occurred at a slightly lower frequency than translating and repeating words/phrases (IR7). Only one occurrence was identified during the initial reading of the passage when a student clearly verbalized that he did not understand the meaning of a word or phrase in the passage he was reading. Although some of the students managed to derive the meaning of unknown words from the context, it was not a requirement in order to answer the item(s) correctly in most cases, as the items mainly assessed processing skills at the local level including detecting and/or matching explicitly stated information in the text.

As for the difference between the initial reading strategies (IR) used by participants on the PBT and CBT, table 41 shows no significant differences between the two testing modes.

Three out of the seven strategies (i.e. IR1, IR6, and IR7) had a p-value of .163, three had a p-value of .331 (i.e. IR5, IR8, and IR9), and the remaining strategy had a p-value of 1 (i.e. IR3). These results indicate that when students read the passage initially, the testing mode did not alter the way they went about doing it.

5.3.3 Category Three: Strategies related to Reading of Questions

The total strategy tokens for this strategy category in PBT were 404, which is 44.2% of the total tokens in this mode. AS for the CBT, the total strategy tokens were 369, which is 45.8% of the total strategy tokens in this mode. This indicates that the strategies used in this category are among the more frequently used. The following eight strategies were found in this category:

TS2: Rereads/repeats question or word/phrase in question stem for clarification: e.g. Which film made Newman a star? (.) Which film Newman star? (.) Which film Newman star? (.) He went to Los Angeles and made his first film (.) [TS2].

TS3: Translates question or word/phrase in question stem to aid in comprehension: e.g. What is the name of Newman's company? (.) (This means that is the company that he worked for) [TS3].

TS4: Paraphrases question stem for clarification: e.g. When did Newman's interest in car racing start? (.) When did Newman begin to care for cars? (.) [TS4].

TS5: Guesses meaning of unknown word(s) in question: e.g. When did Newman first work in the theatre? (.) They are asking what he do*(=does) in the museum (.) [TS5].

TS6: Reads question stem and then scans passage for keyword(s): e.g. Which film made Newman a star? (.) We need the star <scans passage> (.) the star (.) hmm (.) [TS6].

TS7: Reads question stem and then search reads the passage/portion to look for clues to the answer (i.e. keywords): e.g. 'I read the question (i.e. Where did Newman first know Woodward from?) and when I started reading the last paragraph it mentioned that he was living in Los

Angeles when he became engaged to Jean Woodward. So the answer is when he was living in Los Angeles' (*interview S10*) [TS7].

TS8: Reads question and then uses spatial memory to locate keywords: e.g. What is a method actor? (.) A method actor (.) (I remember I found this in the second paragraph) <goes directly to location in passage> (.) A method actor who believes in the role before beginning the film (.) [TS8].

TS26: Goes back to question for clarification: e.g. When did Newman first work in the theatre? (.) Something about a theatre (.) many awards (.) but never won an Oscar (.) No it's not about that (.) so it must be in the beginning (.) so we are looking here we are looking (.) When did Newman first work in the theatre? (.) [TS26].

5.3.3.1 Descriptive Statistics and Paired Samples T-test's Results

Table 42. Descriptive Statistics Strategy Category 3 PBT & CBT

Strategy (Code)	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-Test	
						t	p
TS2	18	3.17	2.256	3.22	3.246	-.145	.886
TS3	18	3.06	2.796	2.56	2.281	.856	.404
TS4	18	3.83	2.813	3.33	3.049	1.584	.132
TS5	18	1.56	3.091	1.56	3.166	.00-	1.000
TS6	18	4.56	2.684	4.67	2.657	.416	.682
TS7	18	1.67	1.609	1.56	1.653	.809	.430
TS8	18	2.72	2.164	2.50	1.855	.747	.466
TS26	18	1.89	1.711	1.11	1.023	2.522	.022
Valid N	18						
Total F.		404		369			

By far the most frequently used strategy related to reading of the question is where the student reads the question stem and then starts to scan the passage for keyword(s) (i.e. TS6). This strategy had a mean frequency of 4.56 in PBT and a mean frequency of 4.67 in CBT, which

indicates that, because the passage consisted of ten items, the students used scanning to look for relevant information for about half of the time. This high frequency was expected due to the nature of the test items, which required expeditious reading operations to locate relevant information in the passage. Paraphrasing of the question stem occurred second most frequently in both PBT (i.e. 3.83) and CBT (i.e. 3.33). Rereading of a word/ phrase in the question stem to aid in comprehension was used third most frequently in both PBT (3.17) and CBT (3.22). Translating the question or part of it occurred fourth most frequently both in the PBT (i.e. TS3=3.06) and in the CBT (i.e. TS3=2.56). Using spatial memory to locate keywords after reading the question occurred fifth most frequently in both the PBT (i.e. TS8=2.72) and CBT (i.e. TS8=2.50). The sixth most frequently used strategy related to reading the question was returning back to the question for clarification in PBT (i.e. TS26=1.89) and was the least frequently occurring strategy in CBT (i.e. TS26=1.11). The seventh most frequently used strategy in PBT (i.e. TS7=1.67) and least most frequently occurring in CBT (TS7=1.56) was reading of the question stem and then reading the passage or a portion of it. The least frequently used strategy in PBT was guessing the meaning of unknown words in the question (i.e. TS5=1.56) and was least in CBT (i.e. TS5= 1.56) as was TS7.

The majority of the strategies related to the reading of the questions as shown above showed no significant differences at the p-level. Half of the strategies occurred between $p = .1$ and $.5$ (i.e. TS3, 4, 7, and 8), three out of the remaining four were between $p = .6$ and 1. Only one strategy is significant at the $p < 0.5$ levels at 0.22 (i.e. TS26). With regards to this item, participants resorted to using it significantly less in the CBT mode than in the PBT mode.

5.3.4 Category Four: Strategies related to Reading of Passage

The total strategy tokens found in this category were 187 in PBT, which is 20% of the total strategy tokens in this mode. As for the CBT, a total of 176 strategy tokens were identified, which accounts for about 22% of the total strategy tokens in that mode. The following seven strategies were identified in this category:

TS9: When found keyword(s)/clue(s), reads sentence containing clue(s)/keyword(s) carefully: e.g. When did Newman first work in the theatre? (.) When did Newman first work in the theatre? (.) (Come back here) (.) After graduating he started working in the theatre (.) [TS9].

TS10: Rereads sentence containing clue(s)/keyword(s) for clarification: e.g. What is a method actor? (.) Yes method actor (.) will go to line 3,4,5,6 <counts lines> (.) Method actor who believes in living the role before beginning the film (.) <TS9> who believes in living the role before beginning the film (.) Next (.) [TS10].

TS11: Paraphrases sentence/ part of sentence containing clue(s)/keyword(s) for clarification: e.g. A method actor is one who believes in living the role before beginning the film (.) That means he lives the actual role before starting acting (.) [TS11].

TS12: Translates word(s)/phrase(s) in sentence containing clue(s)/keyword(s) for clarification: e.g. He studied the boxer's speech and watched him box (.) speech? (.) (learned language) [TS12].

TS13: Reads sentence before/after sentence containing key information for contextual clarification: e.g.

R: 'What did you look for when you had read the question? Did you look for a specific word?'

S: 'I search for first Woodward in the paragraph, so I read the sentence before it and after' (*interview S6*) [TS13].

TS14: Guesses meaning of unknown words in passage: e.g. However, after graduating he started working in the theatre (.) What is theatre? Is it a museum? (.) [TS14].

TS27: Rereads part of passage for clarification: e.g. He studied the boxer’s speech and watched him box (.) the boxer (.) him box? (.) the picture brought Newman (.) hmm (.) method actor, no (.) he spent days (.) he studied the boxer’s speech and watched him box (.) him box (.) the boxer maybe (.) [TS27].

5.3.4.1 Descriptive Statistics and Paired Samples T-test’s Results

Table 43. Descriptive Statistics Strategy Category 4 PBT & CBT

Strategy (Code)	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-Test	
						t	p
TS9	18	4.44	2.749	4.44	2.770	.000	1.000
TS10	18	1.44	1.338	1.33	1.283	1.458	.163
TS11	18	1.17	2.007	1.11	2.026	1.000	.331
TS12	18	.56	.922	.83	1.823	-.792	.439
TS13	18	1.44	1.338	1.06	1.162	2.715	.015
TS14	18	.00	.000	.06	.236	-1.00	.331
TS27	18	1.33	1.495	.94	1.259	2.122	.049
Valid N	18						
Total F.		187		176			

With regards to the test-taking strategies employed when reading the passage, by far the most frequently used strategy was carefully reading the sentence containing the keyword(s) (when found) in PBT (i.e. TS9= 4.44) as well as in CBT (i.e. TS9=4.44). The frequency token was 88 for both modes, which means, as with TS6, that the test-takers read the sentence carefully about half of the time because the number of items in the passage was 10. The second most frequently used strategy was to reread that same sentence for clarification purposes in PBT (TS10= 1.44) as well as in CBT (i.e. TS10=1.33). The other second most frequently used strategy was reading the sentence before and after the sentence that contained the keyword(s) for contextual clarification in PBT (i.e. TS13=1.44), which was fourth most frequently used in CBT

(i.e. TS13= 1.06). The third most frequently used strategy was rereading part of the passage for clarification in PBT (i.e. TS27=1.33), which was fifth most frequently used in CBT (i.e. TS27=.94). The fourth most frequently used strategy was paraphrasing the sentence/part of the sentence containing keywords for clarification in PBT (i.e. TS11=1.17), but it was the third most frequently used strategy in CBT (i.e. TS11=1.11). The fifth most frequently used strategy was translating words/phrases in the sentence containing the keyword (s) for clarification in both PBT (i.e. TS12= .56), which was the sixth most frequently used in CBT (i.e. TS12= .83). Guessing meaning of unknown words only occurred in CBT and was the least frequent strategy used in that mode (i.e. TS14= .06).

As the paired samples T-test results in table 43 show, five out of the seven test-taking strategies related to the reading of the passage did not indicate any significant differences. However, TS13 showed a significant difference at the $p < .05$ level and so did TS27 with a p-value of .015 for the former and .049 for the latter.

5.3.5 Category Five: Strategies related to Aiding in Answering Questions

The total strategy tokens related to answering questions in PBT were 70, which accounted for 7.7% of the total strategy tokens found in this mode. The total strategy tokens in category in CBT were 49, which represent 6.1% of the total strategy tokens in this mode. The total number of strategies identified was five in this category, which are as follows:

TS17 Uses background knowledge to aid in answering question: e.g. He played the role of the boxer (.) This is the famous film with Sylvester Stallone (.) Rocky Graziano in the film Someone up there likes me (.) [TS17].

TS18 Provides answer to question from memory: e.g. Where did Newman first know Woodward from? (.) This was in New York (.) [TS18].

TS20 Guesses answer: e.g.

R: ‘Then Question 2; what is the name of Newman’s company?’

S: ‘I could not find the name of the company so I guessed it was *uncomfortable*.’ (interview S13) [TS20].

TS25 Moves to next question without answering item: e.g.

R: ‘The first question you answered from memory...hmm... ok...what about question 2? You did not find the answer is that right?’

S: ‘I did not find the answer, so I left it and moved on to the next question’ (interview S15) [TS25].

TS28 Uses knowledge of punctuation/capitalization rules to evaluate possible answer: e.g. What is the name of Newman’s company? (.) The picture brought Newman stardom overnight (.) (this cannot be the answer because if it were a name it would be in capital letter) [TS28].

5.3.5.1 Descriptive Statistics and Paired Samples T-test’s Results

Table 44. Descriptive Statistics for Strategy Category 5 PBT & CBT

Strategy (Code)	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-Test	
						t	p
TS17	18	.11	.471	.00	.000	1.000	.331
TS18	18	1.83	1.295	1.56	1.580	.814	.427
TS20	18	.28	.461	.17	.383	1.485	.163
TS25	18	1.11	1.451	.50	.514	1.826	.085
TS28	18	.06	.236	.00	.000	1.000	.331
Valid N	18						
Total F.		70		49			

As table 44 shows, the most frequently utilized strategy in the category related to answering questions is providing an answer from memory in both PBT (i.e. TS18=1.83) and CBT (i.e. TS18=1.56). The second most frequently used strategy was moving to the next question without answering the item in PBT (i.e. TS25=1.11) as well as in CBT (i.e. TS25= .50).

The third most frequently used strategy in this category in the PBT (i.e. TS20= .28) as well as in the CBT (i.e. TS20= .17) was guessing the answer. Using background knowledge to aid in answering the question was the fifth most frequently used strategy in PBT (i.e. TS17= .11) and was not used in CBT. The least frequently used strategy in PBT was the use of knowledge of punctuation/capitalization rules to aid in answering a question (i.e. TS28= .06), which was like strategy TS17 not used in CBT.

The paired samples T-tests in table 44 further show no significant differences for any of the strategies between the paper-based mode and the computer-based mode with p-values ranging from .427 (i.e. TS18) to .085 (i.e. TS25). These results indicate that the testing mode (i.e. PBT or CBT) did not significantly affect the cognitive approach students maintained in this category, i.e. answering the items in the test.

5.3.6 Category Six: Strategies related to Items after having answered them

The total observed strategy tokens in this category were 32 in PBT, which accounts for 3.5% of the total strategy tokens in this mode. The total strategy tokens found in this category in CBT were 17, which represent 2.11% of the total strategy tokens in this mode. The total number of strategies in this category was four, as shown below:

TS21: Reconsiders or double checks response: e.g. It was what he called an uncomfortable start (.) No (.) actually I do not know the answer [TS21].

TS22: Discovers answer to item later on and goes back to change previous answer: e.g. How many questions left? (.) I think 3(.) When did Newman make his first film? There is a mistake here (.) I will write the same answer (.) when he was thirty (.) The first question I wrote a wrong answer (.) I did not know until the 8th question (.) [TS22].

TS24: Changes incorrect answer into correct answer after rereading: e.g. What is a method actor? Woodward an actress he first known in New York (.) no no no not Los Angeles but New York <changes answer Q5> (.) [TS24].

TS30: Translates answer for clarification: e.g. He went to Los Angeles and made his first film (.) his first film (.) His first film <writes down answer> (it refers to the first film) (.) [TS30].

5.3.6.1 Descriptive Statistics and Paired Samples T-test's Results

Table 45. Descriptive Statistics Strategy Category 6 PBT & CBT

Strategy Code	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-Test	
						t	p
TS21	18	1.11	1.605	.72	1.227	1.941	.069
TS22	18	.28	.461	.17	.383	1.458	.163
TS24	18	.22	.428	.06	.236	1.844	.083
TS30	18	.17	.383	.00	.000	1.844	.083
Valid N	18						
Total F.		32		17			

The most frequently used strategy in this category was reconsidering or double checking a response in the PBT (i.e. TS21=1.11) as well as in the CBT (i.e. TS21= .72). The second most frequently used strategy was going back to change the answer after having discovered the correct answer later on in PBT (i.e. TS22= .28) as well as in CBT (i.e. TS22= .17). Changing an incorrect answer into the correct answer is the strategy that could follow from TS22 as it is what a number of students did after having gone back to that particular item he realized needed correction. In PBT this happened with a frequency of .22 whereas in CBT it occurred with a frequency of .06. The least frequently used strategy used in this category was translating the answer for clarification in PBT (i.e. TS30= .17), which did not occur in CBT.

The paired samples T-tests in table 45 show that the strategies used in the PBT and CBT within this category were similar. Reconsidering or double checking a response occurred relatively more often in PBT than in CBT, however, this did not lead to a significant difference between the two as the t-test shows ($p = .069$). This indicates that in addition to the cognitive behavior of students pertaining to strategies when answering an item, their cognitive behavior after having answered the item is likewise comparable between the two testing modes.

5.3.7 Category Seven: Supporting Strategies

The total strategy tokens in this category were 34 in PBT, which equals 3.7% of the total strategy tokens in this mode. The total strategy tokens in CBT were 26, which accounts for 3.2% of the total strategy tokens in this mode. The following three strategies were in this category:

SUP1: Taking notes while reading (observed by researcher).

SUP2: Underlining information in text (PBT)/ highlighting text while reading (CBT): e.g. The picture brought Newman stardom overnight. Newman went on to make films such as Cat on a Hot Tin Roof. The Hustler, Butch (.) I will underline this sentence here so we can come back to it later to translate (.) Cassidy and the Sundance Kid (.) [SUP2].

SUP3: Asking oneself questions: e.g. Uncomfortable start (.) role of a Greek slave? (.) isn't that the name of a company? (.) [SUP3].

5.3.7.1 Descriptive Statistics and Paired Samples T-test's Results

Table 46. Descriptive Statistics for Strategy Category 7 PBT & CBT

Strategy	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-Test	
						t	p
SUP1	18	.06	.236	.06	.236	n/a	.331
SUP2	18	.06	.236	.06	.236	n/a	
SUP3	18	1.78	1.896	1.33	1.749	1.000	
Valid N	18						
Total F.		34		26			

The most frequently used strategy in this category was asking oneself questions in PBT (i.e. SUP3=1.78) as well as in CBT (i.e. SUP3=1.33). The two remaining strategies were both used once in the PBT and the CBT; the first one pertained taking notes while reading (i.e. SUP 1). There is no verbalization for this strategy as the researcher observed one student using this strategy during the think aloud sessions and took note of that. The other strategy involved underlining information in the text when it concerned the PBT whereas in CBT highlighting the text was the strategy to achieve this (i.e. SUP2). Both SUP1 and SUP2 occurred once in both testing modes and therefore have come out as identical. For this reason, there was no need to provide figures for the t-test. The remaining supporting strategy used did not yield a significant difference in strategy usage between the PBT and the CBT although seven more tokens were counted in the PBT on the totals.

5.3.8 Category Eight: Executive Strategies

The total strategy tokens in PBT were 46, which accounts for 5% of the total strategy tokens identified in this mode. The total strategy tokens in CBT were 44, which represent 5.5% of the total strategy tokens in CBT. This category comprised of the following two strategies:

EX1: Verbalizing target of search (word/idea): e.g. OK, start now (.) beginning to read to understand the idea of the paragraph (.) [EX1].

EX2: Monitors location in passage/test: e.g. (his film) since the film winning (.) no it's not here (.) it must be in the beginning (.) I am looking here and it is in the beginning (.) maybe I will find it here (.) [EX2].

5.3.8.1 Descriptive Statistics and Paired Samples T-test's Results

Table 47. Descriptive Statistics for Strategy Category 8 PBT & CBT

	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-Test	
						t	p
EX1	18	1.83	2.176	1.94	2.437	-.148	.884
EX2	18	.72	1.127	.40	1.295	.622	.542
Valid N	18						
Total F.		46		44			

This strategy category consisted of two strategies of which verbalizing target of search (word/idea) was the most frequently used strategy in both the PBT (i.e. EX1=1.83) and in the CBT (i.e. EX1=1.94). Monitoring location in the passage/test was the other strategy in this category, which had a mean occurrence of .72 in PBT and .40 in CBT (i.e. EX2). The paired samples T-test in Table 47 above shows no significant differences in strategy usage between the PBT and CBT mode in this category with p-values around .5 and .9 respectively. This implies

that the change in testing mode did not significantly affect test-takers' executive strategies that were identified in this category.

5.3.9 Category Nine: Evaluative Strategies

The total strategy tokens identified in this category were 38 in PBT, which accounts for 4.2% of the overall total strategy tokens in PBT. The total strategy tokens in this category in CBT were 39, which equals 4.8% of the overall total strategy tokens in this mode. The following three strategies were found:

EV1: Considering /rejecting a possible word/phrase as possible answer to the question: e.g. (what did he do next?) (.) then he went to Los Angeles (.) (I do not think that is the answer) (.) [EV1].

EV2: Indicating whether a search for information is successful/unsuccessful: e.g. believes in living the role (.) he spent days (.) no (.) studied the boxer's speech (.) no (.) picture brought Newman (.) the picture brought Newman stardom overnight (.) can't be the answer (.) hmm (.) high school (.) no (.) pass this question we will come back later (.) [EV2].

EV3: indicates that he doesn't understand the meaning of a word/phrase read: e.g. the next film (.) morning to night (oh too long) Graziano (.) stardom overnight (.) Newman (.) Hustler (what does it mean, hustler? I don't know) (.) [EV3].

5.3.9.1 Descriptive Statistics and Paired Samples T-test's Results

Table 48. Descriptive Statistics for Strategy Category 9 PBT & CBT

	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-Test	
						t	p
EV1	18	.17	.707	.50	.985	-1.458	.163
EV2	18	1.67	2.000	1.50	1.654	.300	.768
EV3	18	.28	.575	.17	.383	.809	.430
Valid N	18						
Total F.		38		39			

The most frequently used strategy out of a total of three strategies identified in this category was the indication of a successful/unsuccessful search in PBT (i.e. EV2=1.67) as well as in CBT (i.e. EV2=1.50). Indication of not understanding the meaning of a word/phrase read in the passage occurred second most frequently in the PBT (i.e. EV3= .28) and least frequently in the CBT (i.e. EV3= .17). The least frequently used strategy in this category was considering a word/phrase in the passage as a possible answer to the question in PBT (i.e. EV1= .17) and second most frequently in CBT (i.e. EV1= .50).

The paired samples T-test in Table 48 shows no significant differences for any of the three strategies in the evaluative category with p-values ranging from .16 up to .76. This further indicates that for this particular category, the testing mode did not affect participants' cognitive behavior in this study.

5.3.10 Category Ten: Inferencing Strategies

The total strategy tokens identified in this category in PBT were 40, which accounts for 4.4% of the overall total strategy tokens in PBT. The total strategy tokens identified in this

category in CBT were 37, which represents 4.6% of the overall total strategy tokens in this mode. The following three strategies belong to this category:

INF1: verifies referent of a pronoun: e.g. When he was thirty he went to Los Angeles and made his first film. It was what he called an uncomfortable start in the movies, in the role of a Greek slave (.) (this is it!) (.) *It* refers to the first film that he made (.) [INF1].

INF2: infers meaning of new word by context e.g. Watched him box (.) box (.) what does it mean? (.) him box (.) Graziano studied the boxer’s speech and watched him box (.) so box is the boxing he does (.) [INF2].

INF3: infers meaning of new word through background knowledge: e.g. he studied the boxer’s speech (.) speech means talking (.) [INF3].

5.3.10.1 Descriptive Statistics and Paired Samples T-test’s Results

Table 49. Descriptive Statistics for Strategy Category 10 PBT & CBT

	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-Test	
						t	p
INF1	18	1.94	.236	1.89	.323	.566	.579
INF2	18	.17	.383	.17	.383	.000	1.000
INF3	18	.11	.471	.00	.000	1.000	.331
Valid N	18						
Total F.		40		37			

The most frequently used strategy out of the three strategies identified in this category was verifying the referent of a pronoun (i.e. INF1). This was assumed to be the case beforehand as two questions of the test assessed pronoun referencing (i.e. Q9 & Q10). The mean frequency measure for the PBT was 1.94 whereas the mean frequency measure for the CBT was 1.89.

The second most frequently used strategy in this category was inferring meaning of a new word through context in PBT (i.e. INF2= .17) as well as CBT (i.e. INF2= .17). The least

frequently used strategy in this category was inferring meaning of a new word through background knowledge in PBT (i.e. INF3= .11) and was not used in CBT. As the paired samples T-test shows in table 49 above, no significant differences in strategy usage between the two modes were found in this category with p-values ranging from .3 up to 1. This further indicates that as far as cognitive behavior is concerned, altering the testing mode did not significantly affect this process.

5.3.11 Category Eleven: Affective Strategies

The total strategy tokens in the PBT mode in this category were 7, which is 0.8% of the overall total strategy tokens in this mode. In CBT, this was only 1, which is only 0.1% of the overall total strategy tokens in this mode. This category consisted of the following strategy:

- AFF1: a.** Self-motivation: e.g. ‘stay focused’ or ‘you can do it’
b. ‘In the name of God’

5.3.11.1 Descriptive Statistics and Paired Samples T-test’s Results

Table 50. Descriptive Statistics for Strategy Category 11 PBT & CBT

	N	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-test	
						t	p
AFF1	18	.39	.698	.06	.236	1.844	.083
Valid N	18						
Total F.		7		1			

This strategy occurred more often in PBT than in CBT but not significantly more. Several times when the test-taker began the test, he mentioned “in the name of God”, which is what Muslims say before commencement of an action/activity (i.e. test) in this case. It was

referred to as Aff1b because of the scarcity of usage in the test and it carried no significance performance-wise between PBT and CBT.

5.4 Discussion of Results Part 2a: Processes in PBT and CBT

5.4.1 Differences Test-Takers' Processes in PBT and CBT

The second element of this study entailed investigating the impact of test-mode administration on test-takers' performance (i.e. RQ2) of which the results were presented in section 5.3 above. An interface was developed based on a synthesis of the literature related to interface design, language testing, and human-computer related factors, which led to a comprehensive model comprising optimal settings for a computer interface when assessing reading on computer (see p.105, chapter 2). The majority of the strategies applied in PBT were also applied in CBT with no significant differences between the frequencies of occurrence in either mode, which suggested no significant effect of test-mode alteration on test-takers' cognitive behaviour. However, some of the strategies were only used in one of the two modes and not in the other. For example, IR9 and TS14 were not used in PBT but only in CBT whereas strategies TS17, TS28, TS30, and INF3 were not used in CBT but only in PBT. This fact did not endanger the validity of the results as despite these strategies only occurring in one of the two modes, the differences between these strategies were not significant, which means that the frequency in the mode the strategy did occur in would have been inconsequential otherwise it would have resulted in a significant difference, which was not the case. As for the strategies that were found in both PBT and CBT, a total of 3 strategies showed significant differences in frequency between the two modes, which were TS26, TS13, and TS27. All 3 were related to answering the test item, i.e. TS26 was related to reading the item (Category 3) whereas TS13 and

TS27 involved strategies related to reading the passage in order to answer the test item (Category 4). The strategies, mean frequencies, and paired samples t-tests' results are summarized below in table 51.

Table 51. Descriptive Statistics of Significant Differences PBT & CBT

	Mean (PBT)	S.D (PBT)	Mean (CBT)	S.D (CBT)	Paired Sample t-test	
					t	p
TS26	1.89	1.711	1.11	1.023	2.522	0.22
TS13	1.44	1.338	1.06	1.162	2.715	0.15
TS27	1.33	1.495	0.94	1.259	2.122	0.49

Apart from the significance of the differences between the two modes for each strategy, it is noteworthy that in all 3 cases the strategies were significantly *less* used in CBT, and each of the strategies had either to do with difficulties understanding the question where test-takers needed to go back to the question for clarification (i.e. TS26), difficulties with understanding keywords where they needed to read around the sentence containing keywords for contextual clarification (i.e. TS13), or general difficulties with understanding parts of the passage (local level), which therefore needed rereading of parts/ a part of it (i.e. TS27). The significance of this is that the PBT showed relatively greater difficulties than the CBT in understanding the question and the passage containing the relevant information to answer the question, which could be a possible explanation as to why the median in CBT was higher (i.e. M=15) than in PBT (i.e. M=14) though this difference was statistically not significant (i.e. p=. 149). In case of possible practice effect, which could have been an argument for the less problems encountered in CBT reflected through the significantly less instances of these strategies, a cross-over design was adhered to in order to control for this, so this would unlikely have been a justified explanation for

this. The cross-over design would further have contradicted a possible argument that memory had played a role due to the fact that the same test was used on both occasions, as this would likely have cancelled out this difference, i.e. it would more likely have led to a non-significant difference. Furthermore, effect of memory (i.e. test-takers remembering the test contents from the previous session and therefore using different/ less strategies) would then most likely have become apparent through their cognitive behaviour by utilizing different strategies or the lack of using certain strategies on the second testing occasion as opposed to the first, which was also not the case. In addition, students were asked in their post-test interviews whether they remembered the contents of the previously taken test to which they replied in the negative. Only one student (S4) recognized the main character's name in the second session, but he neither remembered what the passage was about nor the content of the test-items, which was confirmed through the recordings of his cognitive processes, which did not indicate any behaviour suggesting memory effect on the second testing occasion. As for the other 17 test-takers, there was no indication from the think-aloud recordings that suggested any memory effect. The interviews proved to be of significant importance to crosscheck this possible issue and therefore aided in increasing the validity and reliability of the inferences to be drawn from this section's findings. The reason being that, had memory of the previously taken test played a part, this would have most likely shown by them relying significantly more on, for example, memory related strategies as opposed to using the expected operations to locate relevant information or careful reading related strategies such as word, phrase, and sentence-level understanding of the text. When test-takers did use memory related strategies, they generally did so in both modes on the same item, but these strategies were generally related to remembering what they had just read in the passage. In addition, the questionnaire administered to the test-takers of the main study's sample after they

had completed both tests (i.e. in PBT and CBT) showed that, overall, test-takers were more comfortable taking the CBT than the PBT and did not indicate any problems with features of the interface. On the contrary, there was a clear gravitation towards perceived usability superiority of the CBT over the PBT from the test-takers in the main study as well as from test-takers in the think-aloud group who were asked about their overall experience with the CBT and PBT. This could have been motivated by the fact that the students included in the main study sample and the think-aloud sample were at least moderately computer familiar and could therefore have skewed the responses in favour of CBT as indicated in other studies (e.g. Higgins et al., 2005). However, having all computer familiar participants was a prerequisite in order to investigate the effect of the newly introduced testing mode's interface design on processes and performance, as unfamiliarity could have introduced construct irrelevant variance by negatively affecting test-takers who were not familiar with computers through causing difficulties on, for example, the operational side. Nevertheless, in light of differences in cognitive behaviour, it can be argued based on the comparative study of strategy usage between the two modes that there is a significant effect of testing mode on test-takers' cognitive processes, which then answers research question 2 (RQ2).

The following section qualitatively describes the cognitive processes of test-takers when answering the 10 think-aloud test items in order to generate evidence for the test's cognitive validity. This is done through illustrating what processes-levels test-takers go through when answering the test items and whether these processes are appropriate to the processes the items were anticipated to elicit in advance.

5.5 Results & Discussion 2, Part 2b: Describing Cognitive Processes

5.5.1 Introduction

The results presented in chapter 2a showed no significant differences in strategies between the two modes for the most part and the 3 that did, although favoring CBT, did not lead to significant performance differences. Another point in support of strategy equivalence in the two modes is that test-takers for the most part used the same strategy order when answering test items, i.e. reading the question first, then utilizing expeditious reading operations or memory strategies to locate relevant information, and after that employing mainly careful reading related strategies and processes to ensure sufficient understanding leading to answering the item in question correctly in both modes. Because of this, it was possible to qualitatively describe the processes involved when answering the test's items in a unified way, i.e. without having to distinguish between the two testing modes in terms of processing levels involved when answering the test items. The underlying theoretical model presented in chapter 2 (section 2.5) described the expected sequence of cognitive processing when answering the test items in this study's test. Two stages were identified in the model consisting of reading operations to locate the relevant information (i.e. stage 1), and more careful reading operations and strategies as the test-taker is thought to try to construct a profounder meaning of the located information to ensure correctly answering of the test item (i.e. stage 2). This model would then further allow for distinguishing between successful and unsuccessful attempts of test-takers when answering the test items through identifying (possible) differences in processing levels between the two, which is expected to provide further insights into the cognitive validity of the CBT. Before describing the processes involved when answering the 10 items included in the think-aloud study, an overview is given of students' performance on the proficiency test and the PBT and CBT version

of the TA-study test.

5.5.2 Students' Performance on Think-Aloud Study Test in PBT & CBT

Student performance among the eighteen participants was similar between the two modes ranging from a hundred percent to as low as twenty percent. Important to mention is that only one passage was chosen for the think aloud as the time needed for a participant to complete one passage with accompanying items was estimated to be around 35 minutes. Furthermore, all of the 30 test items of this study's reading test were measured local expeditious reading related operations followed by local careful reading processes, therefore, the ten items accompanying the passage were expected to be sufficient to get a clear insight into the processes activated by test-takers when processing this study's overall reading test.

Table 52. Test-Takers' Scores Think-Aloud Test PBT and CBT

Test Taker	PBT Score %	CBT Score %	P-Test Score %	Test Taker	PBT Score %	CBT Score %	P-Test Score %
1	80	90	93	10	90	100	93
2	80	90	93	11	30	30	67
3	80	90	93	12	60	60	73
4	90	100	95	13	50	40	48
5	50	50	62	14	80	90	91
6	70	90	76	15	50	50	80
7	50	50	77	16	80	70	71
8	50	50	72	17	20	20	53
9	70	70	70	18	70	70	38

As table 52 above shows, test-taker performance in the think-aloud study is rather mixed ranging from twenty percent to a hundred percent among the eighteen participants in both modes. The possible value these varied scores could have for further discussion in this study is

that it enables discrimination of cognitive processing between higher achievers and lower achievers and could further identify strategy/processing patterns (in addition to overall patterns) for either one.

5.6 Overview Expeditious Reading Operations

The test items in this study's think-aloud study were purported to elicit local expeditious reading behaviour, that is, as the initial reading operation to locate information relevant to answering the test items (Urquhart & Weir's, 1998). The targeted sub-skills belonging to this expeditious reading type were scanning and/or search reading. Evidence found for eliciting these local expeditious reading operations on the test items would serve the following two purposes:

1. It would confirm construct relevancy of the test items when they activate the reading operations/ sub-skills they were meant to activate.
2. It would further validate the local expeditious reading type proposed by Urquhart & Weir (1998) as a sub-skill as it would provide corroborating evidence for the divisibility argument within the reading construct.

The 4 main operations/strategies utilized to locate relevant information to answer the test items in the passage were scanning (SC), search reading (SE), spatial memory (SP), and memory (ME). Out of 360 instances (i.e. 18 test-takers, 10 items x 2 modes) 94 utilized scanning to locate required information in the passage in both PBT and CBT. This is around half of the total instances, which confirms the earlier mentioned indication based on the mean frequencies for strategy TS6 (i.e. PBT= 4.56, CBT= 4.67), which involved scanning of the passage to locate relevant information directly after having read the question. The second most frequently used strategy was TS8 (i.e. PBT = 2.72, CBT= 2.50), which represents using spatial memory to locate

relevant information. The third most frequently utilized operation was TS18 in PBT (i.e. 1.83) and was least frequently used in CBT (i.e. 1.56), which was answering the item directly from memory. The least frequently used operation in both PBT and CBT was TS7 (i.e. PBT= 1.67, CBT= 1.56), which represents search reading to locate required information in the passage. On one occasion (i.e. S4, item 7, CBT) a test-taker skimmed through the passage to find the answer to a test-item (SK). On two separate occasions involving different test-takers (i.e. S6, item 2, CBT, and S8, item 5, CBT), other strategies were used (OT). In both of these cases the test-taker moved on to the next item without answering the item in question. This means that, for the most part, test-takers resorted to expeditious reading operations to locate relevant information in the text, and, when they did not, it was because they remembered either the relevant information's location in the passage or they remembered the right information to answer the question from having read it initially. One of the reasons for using memory instead of expeditious reading could have been that the passage was not very long (i.e. 323 words) consisting of only 2 paragraphs, which might have triggered this 'shortcut' to finding the answer.

The following section discusses think-aloud test items, which are discussed in the following 4 parts:

Part 1. Item description

Part 2. Expeditious reading operations utilized by test-takers

Part 3. Descriptive account of common cognitive processes utilized by test-takers

Part 4. Illustration and discussion of levels of processing described in part 3 in light of Khalifa & Weir's (2009) cognitive model of reading

5.7 Description Cognitive Processes/Strategies Utilized as per Test-Item

5.7.1 Item 11

Question: When did Newman **first work** in the theatre?

Sentence containing answer: However, after graduating, he **started working** in the theatre and on several TV shows in New York.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Process level: **Lexis** (word matching & synonym matching), *grammar/syntax*, propositional meaning

Item difficulty (mean): PBT= .42, CBT= .41

5.7.1.1 Operations/Strategies Item 11

The majority of test-takers used either scanning (i.e. S3, S5, S10, S16, and S17) or search reading (i.e. S1, S7, S9, S11, S12, S13, S15, S18), which amounted to a total of around 70% for this item. The remaining 30% utilized memory strategies were either using spatial memory to locate key information (i.e. S4, S6, S14) or answering the question directly from memory (S8, PBT). Test-taker S8, however, used a different memory strategy on the same item in CBT compared to PBT; in PBT he answered the question directly from memory whereas in CBT he located the keyword relevant to the answer from memory and subsequent strategies led to answering the item.

5.7.1.2 Descriptive Account Processes/Strategies Item 11

After the relevant information/keyword(s) had been located, a large number of test-takers started to carefully read the sentence that contained the identified information in order to

comprehend the information needed to answer the item correctly. This was reflected earlier in the descriptive statistics in sections 4.3.4 and 4.3.5 where TS6 and TS9 had the highest frequency level and were about the same for both (i.e. TS6= 4.56, TS9=4.44). As for answering this particular test item, those who answered it correctly in PBT did so in CBT, and those who answered the item incorrectly likewise did so in both modes. The keyword match that dictated the search was ‘theatre’ for many test-takers (interview S4, R=researcher, S=student):

R: *So what did you look for in order to answer question 11 after you had read it? Did you look for a specific word or phrase?*

S: *Yes, theatre.*

Others tried to match ‘work’ with the information in the passage to answer the test item (interview S15):

R: *Could you tell me how you answered question 1?*

S: *I translated question first. After that, I took the word ‘work’ and looked for it in the passage.*

Below is an example of a test-taker successfully answering item 11 (S10):

First question (.) When did Newman first work in the theatre? <scans passage> First, first, first (.) He start to work in New York (.) What date? (.) <scans passage> Date, date, date (.) after graduating (.) after graduating he start working, OK <writes down (correct) answer> Next one (.)

Some students, although using the required strategies to locate the information, were unsuccessful (initially) in answering item 1 (S3):

When did Newman first work in the theatre? <turns page> (.) I think I know the answer (.) In the first paragraph (.) <scans passage> there is no date (.) uh, oh, yes, in Ohio (.) no (.) I think <scans again> there is no date (.) <turns back to question, rereads it and turns back to passage> (.) ah (.) when he was thirty (.) found the question.” <writes (wrong) answer>.

Others who utilized an entirely different strategy in an attempt to answer item 1 were likewise unsuccessful (S2 CBT):

When did Newman first work in the theatre? When he was thirty (.) when he was thirty. <writes down (wrong) answer>

In the example above the test-taker directly verbalized the answer from memory after having read the question without going back to the passage.

5.7.1.3 Levels of Processing Item 11

Item 11 required lexical understanding as well as grammatical/syntactical structure understanding and propositional understanding of the relevant text in order to generate the correct answer to the question. As the first example shows, S10 scanned for the word ‘first’ but found the word ‘started’ for which synonym matching was required (i.e. start vs. first). Because of this, S10 started to search for a date, as he clearly verbalized, likely assuming that starting work would most likely be represented by a date in the text. Because of correct syntactic parsing (subject in a time clause), S10 knew that *after graduating* had to refer to the start of his work and not *when he was thirty*. This becomes clear in the following example where S3 answers this item incorrectly (answer given=when he was thirty) due to association of *when* through mere lexical matching and lack of appropriate syntactic parsing. The same happened with S2, who likewise utilized lexical matching but through memory, which resulted in the same error, i.e. answering ‘when he was thirty’ instead of ‘after graduating’. These examples show that what was required to answer this item correctly was both lexical and grammatical/syntactical understanding to enable the test-takers to discriminate between these two possible answers, as failure to do so led to an incorrect answer.

5.7.2 Item 12

Question: What's the **name** of **Newman's company**?

Sentence containing answer: All the money from 'Newman's Own' salad dressing, popcorn, and spaghetti sauce, now a multi-million dollar business, goes to charity.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Process level: **Lexis** (word-class matching), *grammar/syntax*, propositional meaning, inferencing

Item difficulty (mean): PBT= .22, CBT= .23

5.7.2.1 Operations/Strategies Item 12

Search reading (S7, S9, S12, S13, S18), scanning (S3, S5, S10, S11, S15, S16, S17), and either using spatial memory (S2, S14) or answering directly from memory (S1, S4, S8) were utilized by test-takers for item 12. There was one incident where the same test-taker opted for a strategy involving other than expeditious reading in one of the two modes (S6, on CBT test item 12). The different strategy used in this case was the test-taker not answering the item. In PBT, the test-taker read the question first, then reread the question for clarification and then used spatial memory to locate key information in the passage. When the utilized strategy proved unsuccessful, he moved to question 3 without answering the item. In case of item 12 in CBT, the test-taker read the question first, reread the question, and moved to the next item without having attempted to search for the information in the passage.

5.7.2.2 Descriptive Account Processes/Strategies Item 12

Item 12 was the most difficult item in the passage with a mean of .22 in PBT and .23 in CBT. None of the test-takers successfully answered this question in their first attempt and many left the question to be answered later (S3):

What's the name of Newman's company? Company <scans passage> I think ehh (.) I think ehh (.) the company (...) <turns page> I'll come back to this later (.) going to next question (.)

The majority of the attempts were very laborious on the test-takers' part and they spent by far the longest on this item. Many did not successfully answer this question in the think-aloud although some were successful but not certain about their answer. The example below shows a test-taker who was unsuccessful in answering the question (S15):

Look question 2: What is the name of Newman's company? (name of his company) <scans passage> (his company his company) he start working (.) <writes (wrong) answer 'several TV shows'>.

Some test-takers after having utilized the appropriate strategies locating the information answered the item without certainty about the correctness of the answer given. Below are two examples of this; the first example involves a test-taker who was unsuccessful in his attempt (S18 PBT):

(.) where is the company? (.) company company company (.) woodward (.) I think uncomfortable (.) What the name of Newman's company? Uncomfortable <writes answer>.

The second example shows an excerpt from the think-aloud report of a test-taker that coped with the same problem as the previously shown test-taker; however, unlike test-taker S18, he did get the answer right utilizing very similar strategies whilst being uncertain about its correctness (S3 PBT):

(.) No (.) I think I'll go with Newman's Own (.)<writes down correct answer> (.) maybe it's the name (.) I think it's wrong but (.) It's wrong or it's right (.)

The excerpt below shows an instance where a test-taker successfully answered the test item and was sure of the answer given, however, he discovered the answer while he was reading to answer the next question, i.e. Q13 (S14):

When did Newman's interest in car racing start? <search reads passage> The money from Newman's Own salad dressing (.) oohhh (.) from Newman's Own salad dressing (.) goes to charity, aha (.) Newman's Own! <writes answer to question 2> we found it! (.)

5.7.2.3 Levels of Processing Item 12

This item required higher-level processing such as contextual inferencing (i.e. only the company's name was given but not in relation the actual word 'company' as in the question, which therefore necessitated inferencing). This was likely the main reason that the majority of the test-takers got this item wrong, as lexical matching/ synonym matching and grammar/syntactical knowledge were not sufficient to achieve that in this case.

This shows in the example above where S15, due to unsuccessful contextual inferencing, (wrongly) guessed the meaning of unknown words in context, which led to the wrong answer (i.e. several TV-shows). The same happened with S18, however, he resorted to guessing 'uncomfortable' (equally a wrong guess) to be the right answer based on contextual inferencing flaws. Interestingly, this worked in favour of S3, who appeared to answer (guess) this item correctly based on that same shortcoming in inferencing ability but compensating this with punctuation knowledge i.e. proper name = (possibly) the name of the company. There was no discernible distinction between higher and lower proficiency test-takers for this item, although

the strategies for higher proficiency test-takers tended to be more global, as some appeared to try and (mainly unsuccessfully) integrate information across sentences to formulate an answer, which the lower-level students could not.

In the final example, S14, who did appear to have the appropriate inferencing skill, found the answer when he was search reading to answer the following question, i.e. item 13, as he skipped item 12 because he had not found the answer initially.

5.7.3 Item 13

Question: When did Newman's **interest in car racing** start?

Sentence containing answer: Ever since the film 'Winning', Newman has been **interested in car racing**, and in 1979 he came second in the twenty-four hour Le Mans race.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Processing level: **Lexis** (word matching), grammar/syntax, propositional meaning

Item difficulty (mean): PBT= .39, CBT= .40

5.7.3.1 Operations/Strategies Item 13

Scanning was the most frequently used reading operation for locating key information for this item (S4, S5, S7, S8, S9, S10, S11, S13, S15) and search reading was the second most frequently used (S1, S2, S12, S18), which amounts to a total of around 70%. The remaining 30% was divided between both spatial memory to locate information (S14, S16, S17) and providing the answer directly from memory (S3). As with item 12, test-taker S6 utilized a different strategy between the two modes to locate the information relevant to item 13. He used spatial memory to

locate information related to item 13 in PBT but scanned the passage to locate that same information in CBT.

5.7.3.2 Descriptive Account Processes/Strategies Item 13

Item 3 was of moderate difficulty with a mean of .39 for the PBT and .40 for CBT respectively. The target word(s) when searching for the relevant information by test-takers was *car*, *car racing*, or both *interest* and *car racing*. Below is an excerpt of the think-aloud protocol where a test-taker's strategies led to successfully answering item 13 (S18):

When did Newman's interest in car racing start? When did (.) interested in car (.) <search reads passage> car racing (.) ever since (.) 1,2,3,4 (.) yes it's here (.) ever since the film winning Newman has been interested in car and (.) ehh (.) in car racing <goes back to question> what's the question? When did Newman (.) in car (.)? ever since Newman has been interested in car racing (.) ehh (.) film winning (.) film winning (.) (no not film winning) hmmm (.) they six films (.) winning, since the film winning <writes down (correct) answer>.

Below is an excerpt of a test-taker's verbalization where he answered the item correctly despite clear indications that he lacked lexical knowledge of the words read as shown through mispronunciation on a number of occasions (S13):

When did Newman's interest in car racing start? Car racing start, start, start (.) when did Newman's interest in car racing start? In car start, when did Newman car start. When did Newman in car? Newman has been interested in car. Newman, new man, Newman has been interested in car rakin* (=racing) (.) Every sign* (=since) the film warning* (=winning) Newman has. Every science* (=since) the film weighing* (=winning) Newman has been incared* (=interested) raking* (=racing) < (correctly) answers item>.

The excerpt below shows an unsuccessful attempt by a test-taker trying to answer item 3 (S16):

When did Newman's interest in car racing start? When did Newman begin to care for cars? Here, his marriage was the strongest (.) yes this is it (.) His marriage was long and strong (.) Now he started (.) Yes, 1979 he started <writes down (wrong) answer>.

Some test-takers relied on memory strategies to answer test item 3. This excerpt below shows a test-taker successfully answering the item after having used spatial memory to locate the relevant information (S6):

When did Newman's interested* (=interest) in car racing start? He was interesting* (interested) in car racing. <goes straight to relevant location in passage> Newman has been interesting* (=interested) in car racing since the film winning<writes down (correct) answer>.

Here the test-taker, after having read the question directly, went to the location where the relevant information was present and then read the sentence containing the relevant information entirely for confirmation purposes following which he wrote the correct answer.

5.7.3.3 Levels of Processing Item 13

As in the example above, S18 arrived at the correct answer through initially search reading the passage through which he found the relevant information. He then assumed through syntactical knowledge of the sentence structure and propositional understanding of it, that the subordinate clause *ever since the film winning* preceding the subject *Newman*, related to what followed the subject, was the answer to *when* in the question. Most likely, because it is not common as in what he is probably used to finding in this situation (i.e. a specific date/time etc.), he further checked by reading the sentence that followed, which confirmed that his initially found answer was the correct one, which eventually was what he wrote down as the answer. In example 2, S13 clearly lacked sufficient lexical knowledge, which showed through instances of

graphophonic miscuing exemplified by mispronunciations on several occasions (i.e. *science*, *rakin*, etc.). However, his syntactical/grammatical knowledge of the sentence containing the answer to the item still led him to answer the item correctly. The example of S16 clearly shows that not possessing the required syntactical knowledge lead to an incorrect alternative, as the test-taker here incorrectly applied the synonym matching strategy associating *when* with the year 1979, which would have been more logical at a first glance but not correct in this case, as it followed the coordinating conjunction *and* implying a different time period from the preceding clause. The example of S6, who used spatial memory to locate the relevant information, showed that he likely possessed the required lexical and syntactical knowledge, as he directly connected the correct information to answer the item after having read the sentence only once. This further supports that syntactical knowledge of the sentence was required to correctly answer this item.

5.7.4 Item 14

Question: **How many films** did Newman and Woodward *make together*?

Sentence containing answer: They have co-starred in **six films**.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Processing level: **Lexis** (word matching), *lexis* (synonym matching), grammar/syntax, propositional meaning

Item difficulty (mean): PBT= .80, CBT= .70

5.7.4.1 Operations/Strategies Item 14

Scanning was used most frequently for this item (S6, S12, S13, S16, S17, S18), which, jointly with search reading (S1, S2, S4, S5) amounted for 55% of the test-takers. The memory-

induced strategies spatial memory (S7, S11, and S15) and memory (S3, S8, and S14) were utilized by 6 out of 18 participants, which is amounts to about 30%. Test-taker S9 utilized spatial memory on item 14 PBT to locate information related to the test item but answered that same item directly from memory in CBT.

5.7.4.2 Descriptive Account Processes/Strategies Item 14

This item was the easiest item with a mean of .80 in PBT and .70 in CBT respectively.

The main target keywords to find the relevant information were a combination of *how many* and *films*. The excerpt below shows an instance where a test-taker successfully answered item 14 (S1):

How many films did Newman and Woodward make together? (how many films did he and her make together?) <starts search reading passage> He has (.) first (.) New York. Newman and Miss Woodward were married in Las Vegas in 1958. His marriage to Woodward is one of the longest and strongest in Hollywood. They have co-starred in six films (.) Six films< writes down (correct) answer>.

The excerpt below shows an instance where the student answers the item incorrectly despite using several strategies to locate the item (S13):

How many films did Newman and Woodward make together? Wood (.) Newman to make films such as (.) He made 45 new films (.) When he was living and worked (.) Newman and Miss Woodward (.) film (.) Newman Newman Newman (.) strong (.) evren* (=Inferno) (.) he has made over 45 films and he has won many awards (.) and he first film he won (.) they have stared*(=starred) in 6 films <writes down 45 films=wrong answer>.

Other test-takers used memory related strategies to answer this item. The following excerpt shows a test-taker answering the item correctly from memory (S4):

How many films did Newman and Woodward make together? This is about the films they did together (.). It is easy I don't have to return to the text because I have memorized it <writes down correct answer>.

5.7.4.3 Levels of Processing Item 14

This item mainly required lexical matching (i.e. films), synonym matching (i.e. together vs. **co**-starred) syntactic, and propositional understanding of the clause/sentence, as there were two instances in the passage that mentioned a certain number of films, i.e. 6 films, and 45 films. However, key was here (through propositional understanding) to choose the number of films in connection with both Woodward *and* Newman (6 films) as opposed to only Newman (45 films), referring back to the question, which clearly refers to the films they had made together. S1 directly wrote the answer after having read *they have co-starred in six films*, which indicates that the test-taker was aware of the connection between Woodward and Newman and the 6 films through required lexical, syntactic, (and propositional) knowledge. This is further shown through S13's example, who arrived at the wrong answer (i.e. 45 films) clearly due to not applying synonym matching, in addition to insufficient syntactical knowledge leading to an insufficient propositional knowledge of the clause/sentence, which would have enabled him to at least distinguish between the number of films related to Newman alone as opposed to the number related to Woodward and Newman together. Other test-takers such as S4, who was a higher proficiency student, answered directly (correctly) from memory, further showing the relative effortlessness when answering this particular item as indicated through the mean scores in both modes.

5.7.5 Item 15

Question: Where did Newman **first know Woodward** from?

Sentence containing answer: He was living in Los Angeles when he became engaged to Joanne Woodward, an actress whom he had first known in New York.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Processing level: **Lexis** (word matching), *grammar/syntax*, propositional meaning, anaphoric inferencing

Item difficulty (mean): PBT= .62, CBT= .69

5.7.5.1 Operations/Strategies Item 15

Unlike the previous items (11-14), item 15 was answered mostly by using memory-induced strategies. Out of 18 test-takers, 8 answered this item directly from memory (S1, S2, S4, S5, S9, S13, S16, S17), and 3 used spatial memory to locate key information (S10, S11, S15), which amounts to around 60% of the test-takers. Only 4 participants used scanning to locate the relevant information (S3, S6, S12, and S18). Test-taker S7 used spatial memory to locate the needed information to answer item 15 in PBT but search read the passage in order to locate that same information in CBT. Test-taker S8 paraphrased the question in his L1 after having read the item but then moved on to item 16 without returning back to this item later and therefore left it unanswered. Test-taker S14 utilized memory strategies in both modes but used spatial memory in PBT as opposed to directly answering from memory to achieve the same goal in CBT.

5.7.5.2 Descriptive Account Processes/Strategies Item 15

Item 15 was one of the relatively easier items in this passage with a mean of .62 in PBT and .69 in CBT respectively. The typical keyword(s) search was trying to answer *when* and *know*

Chapter 5: Results & Discussion 2

in relation to Newman and Woodward. Below is an example of a test-taker successfully answering this item (S14):

Where did Newman first know Woodward from? Engaged to an actress (.) he has made over fifty-five* (=forty-five) films but has never won an Oscar (.) he was living in Los Angeles when he became engaged to (.) Joanne Woodward an actress whom he had first known in New York (.) Newman and Miss Woodward were married (.) he had first known in New York (.) oh New York (.) Where did Newman **first** know Woodward from? So they met first time in New York, yes <writes down (correct) answer>

The two excerpts below show a test-taker answering the item incorrectly (S18):

Where did Newman first know Woodward from? Where did? <scans passage> Woodward (.) Woodward (.) Los Angeles (.) heyyy! (.) Los Angeles <writes down (wrong) answer>.

Example 2 (S3):

Where did Newman first know Woodward from? Ehh, from a movie I think (.) They were married (.) <search reads passage> hmmm (.) Las Vegas (.) He has made over 45 (.) became engaged (.) In Los Angeles he became engaged (.) to the actress (.) so the answer is when he was living in Los Angeles <writes down (wrong) answer> the question was where did, so Los Angeles.

A number of test-takers used memory related strategies and answered the question directly from memory (S4):

Where did Newman first know Woodward from? This was in New York <writes down (correct) answer> this goes with a capital because it is a city.

Here a test-taker resorted to memory initially but then double-checked as he had seen two names of cities in the same sentence. So he went back to the sentence containing the information (spatial memory), read it, and concluded it was New York, not Los Angeles (S10):

When did Newman first known (.) where did Newman first know Woodward from? In New York (.) In Los Angeles (.) wait <goes to location in passage> he had first known in New York, OK (.) <writes down correct answer>.

5.7.5.3 Levels of Processing Item 15

One of the essential requirements to answering this item correctly was test-takers' correct assignment of *first* in the question. As shown in the first example, S14 read the complete sentence containing the answer and then started reading the sentence following it. He then reread part of the sentence and reread the question where he identified *first* as being the key to the correct answer, as he then paraphrased the question to confirm understanding, following which he answered the item correctly. The contrast is clear in the second example, where S18 apparently missed this essential link as he merely matched *where* with the location *Los Angeles*, most likely because it occurred first in the sentence (see minimal attachment principle by Frazier, 1978; 1987). The same was the case for S3, who chose *Los Angeles* due to the same flaw as he recalled *where did?* and then wrote down the answer. S4 clearly had the required lexical, syntactical, and propositional understanding exemplified by directly answering the question from memory, as, most likely, had he not had the proper understanding, he would have chosen *Los Angeles* too based on the same principle. This is further exemplified by S10, who was not sure about the location as he had read two in the sentence. He then revisited the location in the

sentence and confirmed that it had to be *New York*, based on correct lexical, syntactical, and propositional understanding.

5.7.6 Item 16

Question: What is a **method actor**?

Sentence containing answer: Newman is a **method actor** *who* believes in living the role before beginning the film.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Processing level: **Lexis** (word matching), *grammar/syntax*, propositional understanding (anaphoric inferencing)

Item difficulty (mean): PBT= .61, CBT= .68

5.7.6.1 Operations/Strategies Item 16

Scanning was the most frequently used reading operation for this item for each test-taker in both modes (S1, S3, S5, S8, S11, S12, S13, S14, S16, S17, and S18). Search reading was utilized by one test-taker (S2), and spatial memory was used by five test-takers (S4, S6, S7, S9, and S10). Test-taker S15 scanned the passage to locate key information in PBT but search read in CBT to achieve that same goal.

5.7.6.2 Descriptive Account Processes/Strategies Item 16

Like item 15, item 16 was one of the easier items in this test passage with a mean of .61 in PBT and .68 in CBT respectively. The typical keyword(s) test-takers searched for were either *method*, *actor* but generally both together. The excerpt from a post-test interview below shows a

Chapter 5: Results & Discussion 2

successful attempt in answering this test item using scanning to locate the relevant information (S6):

R: *Alright, 6.*

S: *What is a method actor?*

R: *Yes.*

S: *A method actor I find it in the second paragraph when he said: Newman is method actor. Then he write: who believes in living the role before beginning the film, so I write: it's an actor who believes in living the role before beginning the film.*

Here the student explained that he had found the word *method actor* in paragraph two and subsequently answered the question with the information that followed the keyword (i.e. definition).

The excerpt from a test-taker's think-aloud verbalization that used spatial memory to locate the information and subsequent strategies led to successfully answering the item is shown below (S10):

What is a method actor? I read this (.) < goes directly to relevant part in passage> Believe in living (.) wait (.) believe in living the role before begin* (=beginning) the film (.) wait (.) in the film someone up there likes me. Newman is a method actor (.) what is a method actor? (.) A method actor believes in living the role before beginning the film (.) Yes (.) a method actor who believes in living the role before beginning the film <writes down correct answer> I don't understand what is meaning of this, OK (.)

The example below shows an unsuccessful attempt at answering this item correctly (i.e. only part of the answer was written) even though the relevant information was identified by the test-taker (S18):

What is method actor? Method where is the method...<scans passage> (.) ehh method method actor (.) has strong (.) the (.) environment (.) popcorn (.) method (.) and did some acting in high

school (.) not in paragraph 1(.) Newman is a method actor who believes in living the role (.) it's here...hmmm (.) who believes <writes down the answer><stops and reads passage> yes (.)<writes down answer> who believes (.) who believes the role before <incomplete answer>.

5.7.6.3 Levels of Processing Item 16

Some students answered the question through the assumption that *method actor* was followed by its definition. The example from the post-test interview with S6 shows that he had most likely answered the question based on this assumption. However, it could also have been the case that he answered based on syntactical knowledge, which introduces the adjective clause by the relative pronoun *who* indicating that what follows would modify *method actor*. The second example further illustrates this, as S10 clearly stated he did not know the meaning of the answer given, but based on his syntax/grammar, managed to answer the item correctly. S18's example further shows this, as he did not write the complete answer, which indicates insufficient knowledge of the grammar/syntax of the clause and subsequently, an insufficient understanding of the proposition, which would have necessitated inclusion of the whole clause.

5.7.7 Item 17

Which film made Newman a star?

Sentence containing answer: The next **film** he chose was his big break. He played the role of the boxer, Rocky Graziano in the film 'Someone Up There Likes Me'.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Processing level: **Lexis** (word matching), grammar/syntax, propositional meaning, inferencing

Item difficulty (mean): PBT= .26, CBT= .25

5.7.7.1 Operations/Strategies Item 17

As with the previous item, scanning was the most frequently used reading operation by the test-takers in both modes (S1, S3, S5, S8, S11, S12, S13, S16, S17, S18), which is around 55% of the total number of test-takers. Spatial memory to locate key words was used by 4 test-takers (S7, S9, S14, and S15) and 2 answered the item directly from memory (S2, S6). Test-taker S4 used his spatial memory to locate key information related to item 17 in PBT but skimmed through the passage to achieve that same goal in CBT. Test-taker S10 search read the passage to find key information in PBT whereas scanning served that same purpose in CBT on the same item.

5.7.7.2 Descriptive Account Processes/Strategies Item 17

This item proved to be one of the more difficult items in this test's passage with a mean of .26 in PBT and .25 in CBT respectively. The test-takers had to lexically match 'a star' in question with 'big break' in the text. Below is an example of a test-taker successfully answering this item (S10):

Which film made Newman a star? Which film? Yes, I remember it (.) biggest (.) I read it (.) I read it (.) <scans passage> Yes, break break break (.) Yes (.) no no no (.) Yes, the next film was his big break (Sweet!) Yes, yes, yes, yes (.) big break, yes (.) someone up there likes me <writes down (correct) answer> yes, yes (.)

The following excerpt is an example of an unsuccessful attempt to answer this test item (S13):

Which film made Newman a star? Which film Newman star? Which film Newman star? He went to Los Angeles and made his first film. It was what he called an uncomfortable in the movies.

Newman is a method actor who believing (.) believes in living (.) living the role before (.) before beginning the film (.) in the film before beginning in the before beginning the film (.) which first made Newman? He went to Los Angeles and made his first (.) Los Angeles (.) Los An <writes down (incorrect) answer> (.)

5.7.7.3 Levels of Processing Item 17

Test-takers generally had difficulties in synonym/word class matching of *big break* and *a star*, which was necessary to identify the correct movie name, as there were several movies mentioned in the test's passage. S10 correctly matched *big break* with *a star* in the question and further, through syntactical/grammatical knowledge and propositional meaning of the sentence that followed, identified that the name of the movie related back to 'big break'. The unsuccessful attempt of S13 confirms these requirements, as he formulated his answer based on firstly incorrect lexical matching of *first film* with *a star*, and, subsequently assigning a place name *Los Angeles* to it as the corresponding antecedent to *first film*, which is clearly incorrect. This could very well be because of the lack of syntactical/grammatical understanding of the clause involved, which, subsequently, led to insufficiently correctly establishing of the propositional meaning of the sentence.

5.7.8 Item 18

When did Newman make **his first film**?

*Sentence containing the answer: When he was thirty, he made **his first film**.*

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

*Processing level: **Lexis** (word matching), *grammar/syntax*, (anaphoric inferencing)*

Item difficulty (mean): PBT= .50, CBT= .48

5.7.8.1 Operations/Strategies Item 18

Scanning was the most frequently used reading operation to locate information in the passage to answer this item (S3, S5, S12, S13, S16, S17, and S18). Using spatial memory to locate information to answer this item was used by 6 test-takers (S6, S7, S9, S10, S14, S15) whereas directly answering the item from memory was done by 4 test-takers (S1, S2, S4, S11). The reason there is an asterisk beside S11's CBT strategy is that although he utilized the same strategy in both modes, it led to an incorrect answer in CBT as opposed to PBT. Test-taker S8 used spatial memory to locate key information in PBT whereas scanning was the reading operation utilized by this test-taker in CBT to achieve the same goal.

5.7.8.2 Descriptive Account Processes/Strategies Item 18

Item 18 was of moderate difficulty with a mean of .50 in PBT and .48 in CBT respectively. Lexical word matching of 'first film' in the question with the same phrase in the passage was required to answer the item correctly in addition to the ability to connect the time clause *when he was thirty* referring to the event of the first film made. Spatial memory was commonly utilized to locate the relevant information in the passage. Below is an excerpt of a test-taker's think-aloud verbalization where the item was successfully answered utilizing this strategy (S16):

When did Newman make his first film? The answer is present in the first paragraph (.) <goes directly to location in passage> When he was thirty, he went to Los Angeles and made his first film (.) His age was 30 (.) How should I write this? (.) < writes (correct) answer> Newman made his first film, when he was 30, good (.)

The think-aloud excerpt below shows a test-taker who was unsuccessful in answering this item (S13):

When did Newman make his first film? A next film (.) He played the role of the boxer (.) Newman (.) Method actor (.) He spend from morning till night (.) He studied (.) Newman wasn't to make film (.) Hot in raw in New York <writes (wrong) answer>.

This example illustrates how a test-taker answered this item directly from memory correctly although he double-checked for the correct way to formulate his answer (S4):

When did Newman make his first film? There is a mistake here (.) I will write the same answer (.) When he was thirty (.) In his thirties or thirty? <turns page> (.) No he was thirty <writes down (correct) answer> (.)

5.7.8.3 Levels of Processing Item 18

As the example of S16 above shows, in addition to lexical matching of *first* and *film* in the question and passage, grammatical knowledge played a key part in successfully answering this item, i.e. subject time clause = subject main clause, and, relating subject complement of the time clause to the object of the main clause. S16 did this correctly, which is illustrated by *his age was 30* after having read the sentence. S13 clearly did not have the correct grammatical foundation to enable him to answer this item correctly, as he even seemed to have been unable to assign time to the word *when* in the question, which is illustrated by him answering the question including a place name (i.e. New York). S4 did have the required grammatical knowledge, which could be inferred through him correctly answering the item directly from memory.

5.7.9 Item 19

“It was what he called an ‘uncomfortable’ start.” What does “it” refer to in line 5?

Sentence containing the answer: When he was thirty, he went to Los Angeles and made his first film.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Processing level: Lexis (word matching), anaphoric inferencing

Item difficulty (mean): PBT= .33, CBT= .36

5.7.9.1 Operations/Strategies Item 19

Both item 9 and 10 required pronoun referencing from the test-takers. Therefore, for both items, 17 out of the 18 used scanning as the reading operation to locate the key information (specific word to look for was given in question including the line it was to be found). Test-taker S10 used spatial memory to locate the keyword required for item 9 whereas test-taker S9 answered the question directly from memory. The reason for the asterisk beside the CBT strategy is, like with test-taker S11 when answering item 8, he provided an incorrect answer in CBT.

5.7.9.2 Descriptive Account Processes/Strategies Item 19

Item 9 was relatively moderately difficult with a mean of .33 in PBT and .36 in CBT respectively. As item 10, item 9 required successful pronoun referencing from the test-takers in order to generate the correct answer. An example of a successful attempt to answering this item is given below (S3):

“It was what he called an uncomfortable start.” What does it refer to in line 5? It was he called an uncomfortable start (.) <starts scanning> ehm (.) where is it? (.) uncomfortable start (.) ah, yes

(.) <reads sentence before sentence with keyword> I think it ehh, yeah, it refers to the film (.)
<writes (correct) answer>.

Below the test-taker was unsuccessful in answering this item and although he did not verbalize it during the think-aloud, the retrospective interview revealed the following (interview, S6):

R: *And then, question 19.*

S: *(reads question) It was what he called an uncomfortable start. What's it refer to in line 4?*

It refers to the company's name I think.

R: *Why do you think that?*

S: *Because in the second question, he asked for his company's name, so it was what he called uncomfortable, so it refers to his company if the answer in the second question is company.*

In the excerpt of the post-test interview below the test-taker answered the question correctly utilizing memory (interview, S15):

R: *OK, nine, you didn't read the passage but directly wrote the answer?*

S: *Yes.*

R: *How did you know it?*

S: *Because I remember in secondary school...the teacher told me...'it' refer to...it was in the thing...*

Here the student revealed that he had learned this strategy in secondary school, which helped him answer item 19 correctly.

5.7.10 Item 20

“He studied the boxer’s speech and watched **him** box.” What does “him” refer to in line 10?

Sentence containing answer: He spent days – from morning till night – with Graziano.

Anticipated reading type: Local Expeditious (i.e. scanning, search reading)

Processing level: Lexis (word-class matching), grammar/syntax, (anaphoric) inferencing

Item difficulty (mean): PBT= .34, CBT= .41

5.7.10.1 Operations/Strategies Item 20

See item 19.

5.7.10.2 Descriptive Account Processes/Strategies Item 20

Item 10 was, as expected, of similar difficulty as item 9, both assessing the same skill (i.e. pronoun referencing). Item 10 had a mean of .34 in PBT and .41 in CBT. As with item 9, successful pronoun referencing was key in generating the correct answer. The excerpt below shows a test-taker who was successful in answering this item (S16):

“He studied the boxer’s speech and watched him box.” What does **him** refer to in line 9? <counts lines and starts reading sentence> He spent days from morning till night with Graziano. He studied the boxer’s speech and watched him box (.) Yes, him refers to Graziano (.) How should I write this? <writes down (correct) answer>.

In the example below the test-taker utilized the similar strategies to the example above; however, he was unsuccessful in answering the item (S18):

“He studied the boxer’s speech and watched him box.” What does **him** refer to in line 9? Line 9, line 9 (.) 1,2,3,4,5,6,7,8,9 <starts reading> he studied the boxer’s speech and watched him box (.) him box, him box (.) hmm (.) Paul Newman, Paul Newman <writes (incorrect) answer>.

5.7.10.3 Levels of Processing Item 19&20

The examples of S3 for item 19, and S16 for item 20 both show that the test-takers correctly inferred from the pronoun denoted in the question that *the film* was referred to in item 19 and *Graziano* in item 20. The example of S6, who answered item 19 incorrectly, shows incorrectly assigning of the pronoun *it* to Newman’s company’s name instead of the required *film*. A similar miscue example is that of S18, who assigned the wrong antecedent to the pronoun

him, i.e. *Newman*, instead of *Graziano*. In both cases, the test-taker did not seem to follow the pronoun referencing strategy correctly. The underlying assumption for this is that in both cases there is no clear evidence of the test-taker analyzing the sentence preceding the sentence that contained the pronoun, which is generally what is required in pronoun referencing. This most likely led to the incorrect answers in both cases. These examples are in support of anaphoric inferencing as the underlying process required to answer both item 19 and 20, which is therefore a qualitative corroboration of these items' construct relevancy and validity.

5.8 Summary

This chapter presented and discussed the results contributing to answering research question 2, which was the effect of interface design on test-takers' cognitive processes in PBT and CBT. The first part (i.e. part 2a) compared the strategies applied by test-takers in both modes through frequency measures and paired-samples t-tests' results to investigate whether any significant differences were present between PBT and CBT. Three strategies were found, which indicated that test-takers had more difficulties understanding test items and the text passage as they used these strategies more often in **PBT**, which mainly included rereading the question and/or part(s) of the passage. However, despite the significance of the differences between the two modes for these three strategies, they did not significantly affect overall performance, which further supported an absence of effect on performance between the two modes. Furthermore, strategy order was not affected by testing mode either, which further substantiated the absence of mode effect on test-takers cognitive behaviour. These results combined (i.e. through answering RQ1: absence of effect on overall performance and RQ2: equivalent cognitive processing) enabled further qualitative analyses to be performed in order to investigate the process levels

utilized by test-takers when answering the test's items, which would contribute to establishing supporting evidence for the cognitive validity of the L2 reading test used in this study.

The second part of this chapter focused on this by qualitatively describing the processes employed when answering the items in the think-aloud study on an item-by-item basis. This was done in four stages starting with an overview of the item in stage 1, which included the test item itself, the sentence containing the answer, the expected reading operations to locate relevant information for this item, the expected process levels for the item, and the mean difficulty of the item in PBT and CBT. The overview was followed by an illustration and discussion of expeditious reading operations employed by test-takers. After that, common processes utilized when answering the item were described, and examples of other operations/strategies utilized than the ones anticipated were given, when found. This was followed by a discussion of the processing levels highlighting successful attempts and unsuccessful attempts through illustration in order to demonstrate possible differences in processing between the two. The next section discusses the steps taken to establish supporting evidence towards the cognitive validity of this study's test as in Weir's (2005) socio-cognitive framework for language test validation through Khalifa & Weir's (2009) cognitive model of reading.

5.9 Establishing Cognitive Validity

Investigating whether the processes elicited by the test items were comparable to the processes employed by the test-takers when answering the items was thought to provide evidence for the cognitive validity of this study's test. This was divided into two stages based on the multidivisible view of reading, which assumed (in this case) expeditious reading operations to locate relevant information followed by more careful reading behaviour in order to ensure

correctly answering the item. Before beginning this process, item 14 was further investigated and is discussed in the following section.

5.9.1 Expeditious Reading Operations

Results of the expeditious reading operations employed when locating relevant information in the passage showed that the 10 items included in the think-aloud study elicited for the most part either scanning or search reading. It further appeared that search reading was often chosen as an alternative when scanning did not deliver the required results, which might have difficulty related implications. Furthermore, when test-takers did not use expeditious reading operations they mainly chose memory related strategies to either locate relevant information in the passage or to answer the item directly (apart from one instance where skimming was used, and 2 instances where the test-taker did not answer the item). This does not take away from the validity of the items in terms of eliciting expeditious reading operations because there was no pattern identifiable to a particular item, test-taker, or testing mode when opting for different strategies. In addition, the fact that memory related strategies were chosen as the alternative to expeditious reading operations further strengthens the validity of the reading items eliciting expeditious reading operations as it indicates that test-takers in these instances had remembered the answer or its location from initially reading the passage, which, again was not relatable to either testing mode and therefore most likely had more to do with working memory capacity and/or L2 proficiency of the particular test-taker in that instance than with the items themselves. Correlational analyses on test-takers placement tests' results and memory strategies utilized showed a significant correlation at the .01 level of .654 (i.e. $p=.004$) indicating that the higher L2

proficiency, the more frequently memory related strategies were utilized further strengthening this notion.

Based on the results discussed above, it could then be argued that the items that were thought to elicit expeditious reading operations in order to locate relevant information in the text, effectively did so, which contributes to the view of reading as being a multidivisible construct including local expeditious reading as one of its reading types by providing qualitative evidence for this. This was further supported through careful reading following when the relevant information in the text had been located through aforementioned reading operations creating a clear distinction between the two reading types (i.e. TS6=scanning or TS7=search reading followed by TS9=careful reading). This confirms Urquhart & Weir's (1998), and Khalifa & Weir's (2009) indication that careful reading likely follows from expeditious reading operations, in this case, both at the local level.

5.9.2 Levels of Processing

The second step in providing supporting evidence for this test's cognitive validity was to see whether the appropriate process-levels would be elicited by the test-items in light of the processing levels in Khalifa and Weir's (2009) cognitive model of reading related to a language testing context as illustrated by Bax (2013) in order to further substantiate appropriateness of processes elicited from the test-takers. The results showed that for each item, when a test-taker did not utilize or incorrectly utilized the process level(s) required to answer the item, it led to an incorrect answer.

Item 11 required lexical and grammatical understanding to generate a correct answer, whereas item 12 required higher level processing and was therefore the most difficult item as

most test-takers lacked sufficient ability in their L2. For item 13 sufficient grammatical knowledge of the clause containing the answer was required to generate the correct answer. Item 14 was one of the easiest items and required mainly lexical and grammatical knowledge to answer it correctly. To successfully answer item 15, anaphoric inferencing in addition to lexical and syntactical processes was required. To answer item 16 correctly, adequate syntactical knowledge of the clause, in addition to sufficient lexical knowledge of the keywords in the question and passage was necessary. Item 17 required lexical matching/synonym matching, grammatical knowledge and a propositional understanding of the sentence to generate a correct answer. Item 18 mainly required application of lexical knowledge and syntactical knowledge to produce a correct answer. Both item 19 and 20 required lexical matching and pronoun referencing skills to successfully answer these two items. The abovementioned processing levels can all be traced back to Khalifa and Weir's (2009) reading model, which, in addition to the appropriately utilized expeditious reading operations, corroborates the relevancy of these operations and processes elicited by the test's items.

Chapter 6: Overview, Conclusions, Implications, and Recommendations

6.1 Introduction

This study was carried out in Saudi Arabia and included a total of 120 Saudi Arabian university students enrolled in the English Language Centre (ELC) of the Preparatory Year Program (PYP). A total of 102 students participated in the quantitative part of the study whereas 18 were part of the think-aloud study conducted.

The overall aim of this study was to contribute to the field of reading and language testing by investigating the effect of interface design on test-takers' performance and cognitive processes whilst taking an L2 reading test in PBT and CBT. A further contribution was to illustrate the processes test-takers employ when answering test-items aimed to elicit local expeditious reading operations in relation to careful reading, which has been identified as a relatively unexplored area in L2 reading research (Urquhart & Weir, 1998). Eliciting the appropriate reading processes to locate the relevant information in the text to answer the test items would then be a first step towards providing supporting evidence for the test's cognitive validity, which is one of the validity elements of Weir's (2005) socio-cognitive framework for language test validity. The second step in this process was to determine whether the process-levels elicited by the test task (i.e. after relevant information had been located) were the same processes test-takers employed when answering the test's items. The theoretical framework of reference used to investigate this was Khalifa and Weir's (2009) cognitive model of reading, which included Urquhart & Weir's (1998) four-level reading matrix of which expeditious reading is an element. This chapter reviews the steps taken, their results/findings, and implications for future research.

6.2 Overview of Research Findings

This section summarizes this study's findings and mentions conclusions drawn from these findings according to the research questions posed in this study.

6.2.1 Overview and Conclusions Performance in PBT and CBT

6.2.1.1 RQ1. *What is the effect of administration mode on test-takers' performance when taking a lower-level L2 reading test?*

As this study's aim was to investigate the effect of interface design following several indications from the field of language testing (e.g. Choi et al., 2003; Fulcher, 2003; Pommerich, 2004), a review of the literature on interface design was conducted according to a devised interface evaluation model, whose elements would embody a 'good interface' (Fulcher, 2003). The interface that was developed based on the literature review was then used in the CBT-version of the L2 reading test, which was administered to the same test-takers in both PBT and CBT mode on separate occasions.

Statistical analyses revealed that, although test-takers appeared to perform better in CBT overall, the difference between the two modes was statistically non-significant. Results of the post-test questionnaire suggested that from the test-takers' point of view, they were more comfortable with taking the CBT, which can be seen as supporting the quantitative findings (i.e. median CBT-score one point higher than PBT). Both PBT and CBT had a high internal consistency, which was very similar between the two modes (around .9 for both). Although the data were not normally distributed, further examination of the spread/distribution of the scores between the two modes revealed no significant differences. Furthermore, correlational analyses

showed a moderately high and significant correlation between PBT and CBT, which further supported an absence of mode effect on overall performance between the two modes.

Item analyses on the thirty items of this study's test were encouraging showing no significant differences between the two modes apart from item 2, and item 14, of which the former was in favour of CBT, because of which the latter was subjected to further qualitative investigation in order to reveal more about the possible underlying cause for this difference. However, no specific computer interface related cause could be found for the significant difference between the two modes on this item (further discussed in section 6.3.1). On the contrary, the overall score on CBT was one point higher than on PBT as mentioned above, suggesting that CBT would be favorable over PBT as far as test-takers' performance is concerned.

6.2.1.2 Conclusions RQ1

The results pertaining **RQ1** confirm the absence of an effect of the newly introduced administration mode overall, as no significant difference was found on overall performance. Although significances were found at the item level, i.e. item 2 (favouring CBT) and item 14 (favouring PBT), it did not affect overall performance and therefore the answer to RQ1 would be in the negative, i.e. no significant effect was detected on overall performance. For this reason, the null-hypothesis accompanying RQ1 was not rejected.

Of further interest was then whether the item performance effect on these two items (i.e. item 14 due to its statistically indicated negative effect) could be attributed to the computer interface (or (an) element(s) of it) of the newly introduced testing mode (i.e. CBT). This was further qualitatively investigated in RQ2, of which the results are reviewed below.

6.2.2 Processes in PBT and CBT

6.2.2.1 RQ2. *Is there any effect of administration mode on test-takers' cognitive processes when taking a lower-level L2 reading test?*

The think-aloud study carried out to answer this research question revealed similar cognitive processing between the two modes, and no significant differences in frequency counts between the two were found for the majority of the strategies.

However, 3 strategies showed significantly greater frequency instances in PBT than in CBT. Further examination revealed that all 3 of these strategies had to do with difficulties with either understanding the question or the text in the test's passage in **PBT**, which essentially would have favoured CBT over PBT, as reflected through performance differences between the two in RQ1 (i.e. PBT M=14 and CBT M=15). Nevertheless, these frequency differences would not have led to significant performance differences despite this (i.e. one point median difference).

These results were in agreement with one of the comparability studies that also looked at test-takers' cognitive processing in PBT vs. CBT. Al-Amri's (2008) study found a significant effect of testing mode on eight of the total of 60 test-takers' strategies in CBT and 66 in PBT. However, further examination revealed that these differences did not affect performance in any way, as was the case in this study. He concluded that process-wise and performance-wise the two testing modes could be considered to be equivalent despite these significant frequency differences found between these strategies. Furthermore, Al-Amri (2008) did not investigate item-level performance in his study, which could have overlooked further significances such as the ones found in this study's quantitative element (i.e. RQ1). The other comparability study that investigated test-takers' cognitive processes in PBT and CBT was Kobrin's (2000). However,

she did not find a significant effect of CBT on cognitive processes of the students in her study and stressed that the CBT did not appear to cause increased memory workload.

Overall, these results indicate that, despite the significant differences on the three strategies mentioned, the interface did not seem to have affected cognitive processes in any way as these strategies indicated difficulties in PBT. Nevertheless, one further step in investigating this was taken by looking at strategy order and performance in relation to the processes utilized on item 14, which is further discussed in section 6.2.3 below.

6.2.2.2 Conclusions RQ2

The results reviewed above show that for the most part cognitive processes were equivalent between the two modes, i.e. for the vast majority of utilized strategies no significant differences were found. Three strategies revealed significantly more frequency instances in PBT and were all three related to understanding the question item or part of the text passage. The fact that CBT was favoured in these cases of significance makes it even more remarkable that in the main study there was a significant difference found in favour of PBT. However, as no significant differences were found in any of the other strategies, as with RQ1, it failed to reject null-hypothesis accompanying RQ2 due to the evidence being unconvincing.

6.2.3 Interface Design

6.2.3.1 Scrutinizing Item 14 in PBT and CBT

As preliminarily indicated, the newly introduced testing mode did not significantly affect the main study's sample's test-takers' overall performance between both testing modes. At the item level however, item 14 revealed significantly lower performance in CBT and was further

investigated through comparing the cognitive processes between PBT and CBT to see if these would reveal any possible underlying causes to this difference and whether they could be attributed to the computer interface. As mentioned, no significant differences in overall strategy usage between PBT and CBT were found apart from TS13, TS26, and TS27, which after further scrutiny, appeared to favour CBT as opposed to PBT. Furthermore, test-takers utilized the same expeditious strategies on this item in both modes to locate the relevant information in order to answer it. One further step taken was to qualitatively examine whether any of the participants in the think-aloud sample had answered this item incorrectly in CBT yet correctly in PBT to see whether process-levels would reveal any significances leading up to this difference. S13 was the only test-taker out of the 18 participants for whom this was the case. Examination of the underlying processes showed that this test-taker did not use synonym matching in CBT, which resulted in an incorrect propositional understanding relating the *45 films* (which was his answer to item 14 in CBT) to both Newman and Woodward instead of the 6 films that would have been the correct answer to this item. There was no clear indication that the CBT was responsible for not executing this lexical process as, other than this, the test-taker behaved in exactly the same manner in both modes (i.e. same processes utilized and in the same order). Therefore, it appeared to have had more to do with the test-taker himself, as he was the only one that had answered this item correctly in PBT yet wrongly in CBT. Sixteen of the others answered this item correctly in both modes, and one test-taker answered the item incorrectly in both PBT and CBT (i.e. S11). Furthermore, only one of the 18 test-takers used a different strategy to locate the information to answer the test item between PBT and CBT (i.e. S10) but it did not affect the answer given (i.e. both items were answered correctly). In addition to the non-significant differences between strategies utilized, strategy order was neither affected on this item nor was it affected on the

other 9 items for all test-takers, which further strengthens the absence of mode effect on test-takers' cognitive behaviour when taking this L2 reading test. These results would suggest that the discrepancy found in the main study's sample have likely had a different underlying cause other than issues with the interface itself, which was of particular importance in this study.

6.2.3.2 Suitability of Computer Interface

The results indicated that although CBT performance was slightly higher than in PBT mode (i.e. PBT M=14, CBT M=15), this difference was not significant. Cognitive processes comparisons between the two modes further indicated that CBT did not affect test-takers' cognitive processes as the generally the same were found in both modes and no significant differences in frequencies was detected. The three strategies that did indicated more difficulties in understanding questions and text passage in PBT, which further suggests that the new testing mode did not affect test-takers' processes, at least not negatively. Although this is not arguable with regards to the RQ's as based on the significances found at the item level in RQ1 and in strategy frequencies for three items in RQ2, it does support that the computer interface, developed according to what has been indicated as optimal interface design in the literature, is suitable for the purpose it was developed for, i.e. not to interfere with the constructs measured. The fact that no clear cause related to the interface could be identified for the discrepancies on item 14, and non-significant differences in cognitive processes between the two modes (apart from the three discussed, which indicated more difficulties in PBT) supports this conclusion and, for this reason, the interface settings shown in the worked out template in chapter 2 on page 105, is a significant contribution to the field of reading and language testing and can be further developed/amended according to its set purpose.

6.3 Conclusions on Cognitive Validity CBT

The processes that contributed to investigating this study's test's cognitive validity reviewed above appeared to be in favour of validity. Firstly, RQ1 revealed no overall performance difference between the two modes followed by RQ2 which together were more in support of equivalence rather than discrepancy between PBT and CBT. Item difference did not affect overall performance nor was it possible to trace its origins back to the CBT and would more likely have different underlying causes. Similarly, strategy results revealed that the significant differences found were due to PBT being more difficult than CBT, which might explain the slightly better performance in CBT by 4%, which was, nonetheless, not significant. Think-aloud verbalizations showed that the test items, which were purported to elicit expeditious reading operations to locate relevant information to the test item in the text, were the processes test-takers employed when searching for relevant information in the test's passage. The alternative strategies chosen were memory related and correlational analyses indicated that this likely was linked to L2 proficiency, i.e. higher proficiency induced more frequent memory related strategies.

Levels of processing employed by test-takers further confirmed the construct relevance of the test's items as the process levels required to answer the items were employed by the test-takers, and, those who employed irrelevant strategies/processes or were either not able or did not employ the required strategies mostly answered the item incorrectly as a result (apart from the memory related strategies for the reason indicated earlier).

Based on the accumulated supporting evidence at the different stages of the investigation process, it can be concluded that there is a strong support in favour of the cognitive validity of this study's test contributing to its overall construct validity.

6.4 Overall Contributions of this Study

This study aimed to achieve a number of purposes as previously indicated in section 1.4 which are hoped to meaningfully contribute to the field of L2 reading language testing. These are further illustrated below specifying each of the different areas of contribution.

1. Optimal Computer Interface . One of the contributions of this study to the field of language testing is the development of a template comprising the optimal settings of a computer interface for a CBT of L2 reading through a synthesis of the literature on the different elements of the interface from various areas of knowledge including reading, language testing, and human computer interaction, which can be further developed by language testing organizations to aid in minimizing possible construct irrelevant variance in computer-based L2 reading tests.

2. Comparability Studies.

(Design). A further contribution of this study is that a within-subjects design was applied to comparing test-takers in both PBT and CBT. This is different to many studies that used between-subject designs, which did not control for test-takers' individual differences. Kobrin (2000), Choi et al., (2003) and Al-Amri (2008) are studies this study adds to as they likewise employed a within-subject design when comparing test-takers in two testing modes.

(Processes). Another contribution to the field of language testing is that, contrary to the majority of the comparability studies focusing solely on product comparisons (i.e. scores), this study added a cognitive dimension to it by examining test-takers' cognitive processes in PBT and

CBT, which enabled more comprehensive assessment of the reading construct and further investigations into the cognitive validity of an L2 reading test.

(Outcomes). The results of this study showed no effect of administration mode either on test-taker performance (RQ1) or the processes (RQ2), which is a further significant contribution to the field of language testing with regards to comparability studies in particular.

3. Assessment Format. The fact that this study is the first to the researcher's knowledge that involved open-ended questions (i.e. SAQ's) when investigating test-takers' cognitive behaviour in both modes, it further contributes significantly to the field of language testing supporting earlier theories of researchers that carefully formulated SAQ's could be a suitable alternative to MCQ's in language testing provided they were appropriately devised (e.g. Weir, 1990; Alderson et al., 1995; Alderson, 2000; Bachman & Palmer, 1996; Bachman, 2004; Magliano et al., 2007).

4. Local Expeditious Reading Behaviour in L2. This study investigated local expeditious reading behaviour in relation to lower-level processing, which has not been researched extensively as indicated by Urquhart and Weir (1998). Its contribution is significant to the field of L2 reading and language testing as it describes local expeditious reading in a language-testing context in both PBT and CBT providing a clearer insight into how this reading type is employed by L2 test-takers in this setting.

5. Multicomponentiality of the Reading Construct. Whether the reading construct is unitary or consists of divisible components has been an element of debate for reading and language testing researchers who have proposed various elements from which various views of reading emerged, i.e. a *unitary view*, a *bidivisible view*, and a *multidivisible view* (Weir and Porter, 1996). This study further contributes to the reading literature by providing empirical evidence for local expeditious reading being a separately identifiable component of the overall reading construct as

indicated in Urquhart & Weir's (1998) reading matrix, which was the theoretical framework upon which this study was grounded. This empirical evidence validated the local expeditious reading element of the framework and through the selection of test items that elicited this reading type in this study, provides further evidence for this by showing it is separately assessable. Although the multidivisible view of reading is assumed in this study, the evidence generated in support of this is that it consists of at least two elements (i.e. local expeditious reading and local careful reading), but this is due to this study's focus merely being on these two elements in order to address the gap in the current literature and therefore does not negate the existence of additional reading components.

6. Cognitive Validity of an L2 Reading Test.

Comparing the cognitive processes in PBT and CBT in this study's reading test was the first step towards investigating this study's test's cognitive validity. The two-stage process for establishing the cognitive validity of this study's CBT is a significant contribution to the field of language testing, as there is little published research that has done this. Furthermore, using Khalifa and Weir's (2009) cognitive model of reading as an anchor framework for establishing this did not only provide evidence in support of the cognitive validity of this study's test's but also empirically validated the framework itself, and provided supporting evidence for the construct relevancy of the test items selected for this study's purpose.

7. Target Context. This study's context is a relatively unexplored one in terms of CBT and English language assessment in general. This study contributes significantly to the target context, as it is the first study of its kind investigating expeditious reading operations using open-ended question format, and the second comparability study that investigated cognitive processes in PBT and CBT.

6.5 Study Limitations and Future Research

Despite the clear set aims and objectives to be achieved in this study beforehand, there have been a number of limitations in certain aspects of this research. These limitations did not significantly influence the overall validity of the generated results but could be improved on in further studies. Furthermore, this study provided a platform for a significant number of areas for further investigation for which suggestions are given in this section.

1. Study's Participants. The participants in this study were from the province of Hail in Saudi Arabia and were enrolled in the preparatory year program of one particular university. This limits the generalisability of the findings to the complete preparatory year population throughout the country involving other universities. Furthermore, other profession specific disciplines such as medicine have their own unique student population, which could result in different findings compared to students from different disciplines, even when from the same region. Therefore, future research should include preparatory year students from other universities from the various regions in the country in order to get a more complete insight into whether the results obtained in this study are region related or allow for interregional interpretations.

1. Reading Test.

Reading Types. This study's focus was on local expeditious reading operations in relation to mainly careful reading dictated by the test's items, which is only part of the academic reading construct, as this involves global reading and higher-level text processes in addition to local level reading and lower-level processes. In order to see whether the results of this study with regards to eliciting appropriate reading operations and process-levels can be related to academic reading where higher level processes are required, it would be recommended to involve reading tests that assess these global reading operations and higher process-levels to see whether it affects

cognitive behaviour and whether the cognitive validity of the reading test would still be warranted.

Passage Length. The reading passage used in this study's test was 303 words in length, which is relatively short compared to the newer versions of the TOEFL, for example, which might have induced the observed usage of more memory related strategies by this study's test-takers who had higher L2 proficiency levels when locating relevant information in the passage or answering some of the test items. Therefore, future research should include longer reading passages in order to see whether the frequency of these strategies reduces because of this change in passage length. Furthermore, it would be interesting to see what the effect of increased scrolling range due to this would have on test-taker behaviour and performance.

Assessment Format. Although one of the innovations this study introduced was open-ended CBT assessment through SAQ's, it would be interesting to investigate whether the results obtained from a performance perspective and from a cognitive processes perspective would be comparable when altering the assessment format as the likely introduction of, for example, more test-wisness strategies could alter the way test-takers interact with the CBT and could therefore introduce construct irrelevant variance affecting either processes or performance.

Test-Retest Reliability: Parallel Tests. As briefly discussed in section 3.8, concerns were expressed with employing the test-retest reliability method, i.e. assessing the same test-taker on two separate occasions on the same test and parallel tests were suggested as a better alternative to this (e.g. Anastasi, 1988; Alderson, 1991a; Weir, 2005). This study employed the test-retest method to control for test-takers individual differences, as using parallel tests when comparing cognitive processes in PBT and CBT had been reported as problematic earlier (e.g. Kobrin, 2000). Replicating this study using parallel tests in a within-subject design instead of using the

same test on two occasions would be recommended to see whether it affects student performance differently in any way in comparison to this study.

Instrumentation. Results from the PTQ indicated that test-takers favoured CBT over PBT. However, they completed this questionnaire after they had taken both modes of the reading test, which means that a 5-week gap was observed in between and therefore accuracy of recalling features of the firstly taken testing mode were most likely not optimally accurately comparable to the features of the second session's testing mode. A parallel test-retest reliability design would have been the method of choice in controlling for this, as in that case test-takers could have taken the PBT and CBT in a single session, or in two sessions one closely after the another, which would have provided more accurate results. Despite this, the fact that the students were all at least moderately familiar with computers might have influenced their perception as well but this was a prerequisite to investigate the effect of the independent variable in this study.

Language Skill. The skill of interest in this study's test was L2 reading commensurate to the identified gap in the literature underrepresenting L2 expeditious reading. However, other skills such as writing, listening, and speaking need to be investigated also in order to further contribute to the field of language testing.

4. High-Stakes Situations. This study's results showed one point difference on overall performance between PBT and CBT (i.e. M=14 in PBT and M=15 in CBT) favouring CBT. In a high-stakes situation, which is ultimately where it matters most, this might very well be the difference between passing an exam and failing one for some test-takers. Further investigation is therefore needed into the magnitude of this possible effect in these high-stakes contexts.

5. Institutionalized Test. This study's test was a reading test developed by the institution itself for achieving its internal objectives. Therefore, this study's results are not generalizable to internationally standardized tests such as the TOEFL or IELTS.

6. Gender. It would be of significant importance to include female participants in subsequent studies in order to see whether performance and behaviour are comparable between the two genders.

7. Validity Types. This study investigated the cognitive validity of the L2 reading test in both modes, which is only one element of Weir's socio-cognitive framework for language test validity. Therefore, other types of validity are encouraged to be examined in subsequent studies using the template developed for this study's purpose.

8. Software for Interface design. The hotpotatoes software used for developing this study's interface has a number of limitations, one of them being the lack of automated scoring features. This would be essential to stakeholders in the field of language testing, as this is one of the main administrative advantages computer-based testing has over traditional paper-based testing. Therefore, it is recommended to implement the proposed optimal interface settings in the model in this study into more advanced software programs that do contain this feature in order to benefit larger language testing projects.

6.6 Concluding Remarks

A number of important issues indicated in the field of reading and language testing have been investigated in this study. An interface design evaluation model was proposed leading up to a model reflecting optimal settings for an interface to be used by various stakeholders for reading

assessment purposes. The model proved to be suitable for this purpose supported by quantitative and qualitative evidence generated in this study.

Urquhart and Weir's (1998) local expeditious reading and careful reading types have been validated through this study by evidencing construct relevance through test-takers cognitive processes. The appropriateness of the processes is further validated through Khalifa and Weir's (2009) cognitive model of reading, which further validated the latter's model with regards to lower-level processes when reading a text in L2.

The aforementioned contributions further provide evidence for the cognitive validity of the CBT (and PBT) through illustration of test-takers' cognitive behaviour in the two modes.

This study therefore contributes significantly to the field of reading and language testing by providing stakeholders with a template comprising optimal settings for a computer interface as a basis for lower-level L2 reading assessment. It further contributes to the field of L2 reading by validating the aforementioned two reading types and the lower-level processes involved when carefully reading a text in addition to proposing a cognitively valid test of L2 reading, which therefore provides solid supporting evidence towards its construct validity. Due to these contributions in addition to this study's limitations, it further created further opportunities for further, more elaborate research in this area.

Bibliography

Abanomey, A. (2002). *The effect of texts' authenticity on reading comprehension test-taking strategies used by adult Saudi learners of English as a foreign language*. Unpublished PhD dissertation: Arizona State University, USA.

Adams, N. J. & Anctil, T. M. (2003). Computer-based Testing in Vocational Assessment and Evaluation: A Primer for Rehabilitation Professionals. *Vocational Evaluation and Work Adjustment Journal*, Vol. 35, pp. 5-16.

Afflerbach, P. & Johnson, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behaviour*, Vol. 16, pp. 307-22.

Al-Amri, S. (2008). Computer-based testing vs. paper-based testing: A comprehensive approach to examining the comparability of testing modes. *Essex Graduate Student Papers in Language & Linguistics*, 10, 22-44.

Alderson, J. C. (1981c). 'Report of the discussion on communicative language testing' in Alderson and Hughes 1981: 55-65.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a foreign language*, 5(2), 253-270.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Ernst Klett Sprachen.

Bibliography

Almazova, N., & Kogan, M. (2014). Computer Assisted Individual Approach to Acquiring Foreign Vocabulary of Students Major. In *Learning and Collaboration Technologies. Technology-Rich Environments for Learning and Collaboration*(pp. 248-257). Springer International Publishing.

Almond J. R., & Mislevy R. J., (1999). Graphical Models and Computerized Adaptive Testing. *Applied Psychological Measurement*, Vol. 23 No. 3, pp. 223–237.

Almond, R., Steinberg, L. and Mislevy, R. (2002). Enhancing the Design and Delivery of Assessment Systems: A Four-Process Architecture. *The Journal of Technology, Learning, and Assessment*, 1(5), available online at: <http://www.jtla.org>

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999), *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (1986). *Guidelines for Computer-based Tests and Interpretations*. Washington, DC: Author.

Anastasi, A., (1988). *Psychological testing*. London: Macmillan.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8(1), 41-66.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*, 13(2), 111-127.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bibliography

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Baker, J. R. (2003). The impact of paging vs. scrolling on reading online text passages. *Usability News*, 5(1). Online at: <http://psychology.wichita.edu/surl/usabilitynews/5I/pagingscrolling.htm>

Baker, J. R. (2005). Is multiple-column online text better? it depends. *Usability News*, 7, 2, Online at [http://www.surl.org/usabilitynews/72/pdf/Usability News 72 - Baker.pdf](http://www.surl.org/usabilitynews/72/pdf/Usability%20News%2072%20-%20Baker.pdf)

Baker, T. L., & Risley, A. J. (1994). *Doing social research*. New York: McGraw-Hill.

Banerjee, J., Majumdar, D., Pal, M. S., & Majumdar, D. (2011). Readability, subjective preference and mental workload studies on young Indian adults for selection of optimum font type and size during onscreen reading. *Al Ameen Journal of Medical Sciences*, 4(2).

Baudisch, P., Cutrell, E., Robbins, D., Czerwinski, M., Tandler, P., Bederson, B., & Zierlinger, A. (2003, August). Drag-and-pop and drag-and-pick: Techniques for accessing remote screen content on touch-and pen-operated systems. In *Proceedings of INTERACT* (Vol. 3, pp. 57-64).

Baudisch, P., & Rosenholtz, R. (2003). Halo: a technique for visualizing off-screen objects. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 481-488). ACM.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing* 30 (4), 441-465.

Bibliography

Beidel, D. C., Turner, S. M., & Taylor-Ferreira, J. C. (1999). Teaching study skills and test-taking strategies to elementary school students The Testbusters Program. *Behavior Modification*, 23(4), 630-646.

Belmore, S.M. (1985). Reading computer-presented text. *Bulletin of the Psychonomic Society*, 23, (12-14).

Bennett, R. E. (2002) Inexorable and Inevitable: The Continuing Story of Technology and Assessment. *The Journal of Technology, Learning, and Assessment*, 1(1), available online at: <http://www.jtla.org>

Bernard, M., Chaparro, B., & Thomasson, R. (2000). Finding information on the Web: Does the amount of whitespace really matter. *Usability News*, 2(1), 1.

Bernard, M., Fernandez, M. and Hull, S. (2002a). The effects of line length on children and adults' online reading performance. *Usability News*, 4, Online at: http://psychology.wichita.edu/surl/usabilitynews/42/text_length.htm

Bernard, M., Lida, B., Riley, S., Hackler, T. and Janzen, K. (2002b). A comparison of popular online fonts: which size and type is best? *Usability News*, 4, Online at: <http://psychology.wichita.edu/surl/usabilitynews/41/onlinetext.htm>

Bernard, M., Mills, M., Peterson, M. and Storrer, K. 2001, A comparison of popular online fonts: which is best and when? *Usability News*, 3, Online at: http://psychology.wichita.edu/surl/usabilitynews/3S/usability_news.html

Bernhardt, E. (1991). *Reading development in a second language: Theoretical research and classroom perspectives*. Norwood, NJ: Ablex.

Beymer, D., Russell, D. M., & Orton, P. Z. (2007). An eye tracking study of how font size, font type, and pictures influence online reading. *Proceedings INTERACT 2007*, 456-460.

Bibliography

Blackhurst A. (2005) Listening, reading and writing on computer-based IELTS and paper-based versions of IELTS UCLES Research Notes 21: 14-17

Blomberg, J.L. and Henderson, A. (1990). Reflections on Participatory Design: Lessons from the Trillium Experience, in Proceedings of CHI'90, ACM Press, 353-359.

Boo, J. (1997) *Computerized versus paper-and-pencil assessment of educational development: Score comparability and examinee preferences*. Unpublished PhD dissertation, University of Iowa, USA.

Boyarski, D., Neuwirth, C., Forlizzi, J., & Regli, S. H. (1998, January). A study of fonts designed for screen display. *In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 87-94)*. ACM Press/Addison-Wesley Publishing Co.

Bradshaw, A.C. (1998, February). Measuring the learning cost of presentation interference. *Paper presented at the 20th National Convention of the Association for Educational Communications and Technology (AECT)*. St. Louis, MO.

Bridgeman, B., Bejar, I. I., & Friedman, D. (1999). Fairness issues in a computer-based architectural licensure examination. *Computers in Human Behavior, 15*, 419–440.

Bridgeman, B., Lennon, M.L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS RR-01-23). Princeton, NJ: Educational Testing Service.

Brown, C. M. (1989). *Human-computer interface design guidelines*. Norwood, NJ: Ablex.

Brown, H. (1994). *Principles of Language Learning and Teaching*. Upper saddle River, NJ: Prentice Hall.

Bibliography

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Bugbee, A.C. (1992). Examination *on demand*: Findings in ten years of testing by computer 1982-1991. Edina, MN: TRO Learning.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, 28, 282-299.

Butcher, J. N. (1987). The use of computers in psychological assessment: An overview of practices and issues. In J. N. Butcher (Ed.), *Computerized Psychological Assessment: A practitioner's guide* (pp. 3-14). New York: Basic Books.

Butcher, J. N., Perry, J. N., & Atlis, M. M. (2000). Validity and utility of computer-based test interpretation. *Psychological Assessment*, 12(1), 6.

Campbell, F. W., & Durden, K. (1983). The visual display terminal issue: a consideration of its physiological, psychological and clinical background. *Ophthalmic and Physiological Optics*, 3(2), 175-192.

Carson CF, Ashton L, Dry L, Smith DW, Riley TV (2001) *Melaleuca alternifolia* (tea tree) oil gel (6%) for the treatment of recurrent herpes labialis. *J Antimicrob Chemother* 48: 450–451.

Cassady, J. C. & Gridley, B. E. (2005). The effects of online formative and summative assessment on test anxiety and performance. *Journal of Technology, Learning, and Assessment*, 4(1). Available from <http://jtla.org>

Chae, M. and Kim, J. (2004), Size and Structure Matter to Mobile Users: An Empirical Study of the Effects of Screen Size, Information Structure, and Task Complexity on User Activities with Standard Web Phones, *Behaviour & Information Technology*, forthcoming.

Bibliography

Chalhoub-Deville, M. (1990). *Issues in computer-adaptive testing of reading proficiency*. Cambridge: Cambridge University Press.

Chaparro, B., Baker, J., Shaikh, A., Hull, S., and Brady, L. (2004) Reading online text: a comparison of four white space layouts. *Usability News*, 6(2), Online at <http://psychology.wichita.edu/surl/usabilitynews/62/whitespace.htm>.

Chaparro, B. S., Shaikh, A. D., & Chaparro, A. (2006). Examining the Legibility of Two New ClearType Fonts. *Usability News*, 8(1).

Chapelle, C. A. (2001). *Computer Applications in Second Language Acquisition: Foundations for teaching, testing and research*. Cambridge: Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press.

Cheu, R. A. (1998). Good vision at work. *Occup Health Saf* 67 20-4.

Chisholm, W., Vanderheiden, G. and Jacobs, I., editors, 1999: Web content accessibility guidelines 1.0, <http://www.w3.org/TR/WCAG10/WAI-WEBCONTENT-19990505/> (accessed January, 2013).

Choi, I.-C. (2000). Language testing in the 21st century: prospects of computer adaptive testing and performance testing. *Language Research* 36, 205–41.

Choi, S. W., & Tinkler, T. (2002, April). Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. In *annual meeting of the National Council on Measurement in Education, New Orleans, LA*.

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295-320.

Bibliography

Clariana, R. B. & Wallace, P. E. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, 33(5), 593-602.

Clough, S. J., (2008). *Computerized versus Paper-and-Pencil Assessment of Socially Desirable Responding: Score Congruence, Completion Time, and Respondent Preferences*. Unpublished PhD-Thesis. University of Iowa, USA.

Cohen, A. D. (1984). On taking language tests what the students report. *Language testing*, 1(1), 70-81.

Cohen, A. D. (1986). Concurrent Verbal Reports in Second Language Acquisition Research. *English for Specific Purposes*, 5, 131-145.

Cohen, A. D. (1986). Mentalistic Measures in Reading Strategy Research: Some Recent Findings. *English for Specific Purposes Journal*, 5(2), pp.131-45.

Cohen, A. D., & M. C. Cavalcanti (1987). Concurrent Verbal Reports in Second Language Acquisition Research. *ESpecialist*, 16, 13-28.

Cohen, A. D. (1998). Strategies and Processes in test-taking and SLA. In Bachman, L. F. & Cohen, A.D. (Eds) *Interfaces between SLA and Language Testing Research*. Cambridge: Cambridge University Press.

Cohen, A. D. (1998). *Strategies in Learning and Using a Second Language*. London: Longman Press.

Cohen, A. D., & Pinilla-Herrera, A. (2009). Communicating grammatically: Constructing a learner strategies website for Spanish. *A new look at language teaching and testing: English as subject and vehicle*, 63-83.

Bibliography

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307-331.

Cohen, A. D. & Pinilla-Herrera, A. (2009). Communicating grammatically: Constructing a learner strategies website for Spanish. In 2009 LTTC International Conference on English Language Teaching and Testing. A new look at language teaching and testing: English as subject and vehicle (pp. 62-74). Taipei, Taiwan: Language Teaching and Testing Centre.

Cohen, A. D., & Upton, T. A. (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.

Cook, V. (2001). *Second Language Learning and Language Teaching*. Oxford: OUP.

Cook, L. K., & Mayer, R. E. (1983). Reading strategies training for meaningful learning from prose. In *Cognitive strategy research* (pp. 87-131). Springer New York.

Creed, A., Dennis, I. and Newstead, S. 1987, Proof-reading on VDUs, *Behaviour & Information Technology*, 6, 3 ± 13.

Creswell, J. & Clark, V. (2007). *Designing and Conducting Mixed Methods Research*. London: Sage Publications.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281.

Bibliography

Cunningham, G. K. (1998). *Assessment in the Classroom*. London: Falmer Press.

Darroch, I. and Goodman, J. and Brewster, S.A. and Gray, P.D. (2005) The effect of age and font size on reading text on handheld computers. *Lecture Notes in Computer Science* 3585:pp. 253-266.

Davies, A. (1977). The construction of language tests. In J. P. B. Allen & A. Davies (Eds.), *Testing and experimental methods. The Edinburgh Course in Applied Linguistics* (Vol. 4, pp. 38–104). Oxford: Oxford University Press.

Davies, A. (1990). *Principles of Language Testing*. Oxford: Basil Blackwell Ltd.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., McNamara, T. (1999). *Dictionary of Language Testing*. CUP.

Davis, J. H., Carey, M. H., Foxman, P. N., & Tarr, D. B. (1968). Concurrent Verbal Reports in Second Language Acquisition Research. *Journal of Personality and Social Psychology*, 8, 299-302.

De Beer, M. & Visser, D. (1998). Comparability of the paper-and-pencil and computerized adaptive versions of the General Scholastic Aptitude Test (GSAT) Senior. *South African Journal of Psychology*, 28(1), 37-42.

De Bruijn, D., De Mul, S., & Van Oostendorp, H. (1992). The influence of screen size and text layout on the study of text. *Behaviour & Information Technology*, 11(2), 71-78.

Del Galdo, E. (1990). Internationalisation and translation: some guidelines for the design of human -computer interfaces. In J. Nielsen (Ed), *Designing User Interfaces for International Use*, 1-10. New York: Elsevier.

Bibliography

Del Gado, E. & Nielsen, J. (Eds.). (1996). *International user interfaces*. New York: John Wiley & Sons.

Denscombe, M. (2000) *The Good Research Guide: for small-scale social research (second edition)*. Maidenhead: Open University Press.

Devriendt, Y.A. Computer-based testing. In M. Born, C.D. Foxcroft & R. Butter (Eds.) (2008), *Online Readings in Testing and Assessment*, International Test Commission.

Dillon, A., Richardson, J., McKnight, C. (1990b). The effect of display size and text splitting on reading lengthy text from the screen. *Behaviour and Information Technology* 9 (3) 215–227.

Dillon, A. (1992) Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics*, 35(10), 1297-1326.

Dillon, A. (1994) *Designing Usable Electronic Text: Ergonomics Aspects of Human Information Usage*. London: Taylor and Francis.

Dillon, A. (1996) *TIMS: A framework for the design of usable electronic text*. In: H. van Oostendorp and S. de Mul (eds.) *Cognitive Aspects of Electronic Text Processing*. Norwood NJ: Ablex, 99-120.

Dillon, J. T. (1988). *Questioning and teaching: A manual of practice*. New York: Teachers College Press.

Dimock, P. H. & Cormier, P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement & Evaluation in Counseling & Development*, 24, 119–126.

Bibliography

DiPierro, C., & Nachman, G. Raderman (2000). "Screen Size and Web Browsing." *UI Design Update Newsletter*.

Dix, A. (2005). Chapter 3: Human-Computer Interaction and Web Design. In *Handbook of Human Factors in Web Design*. Robert W. Proctor and Kim-Phuong L. Vu (Eds.). Lawrence Erlbaum. pp. 28-47

Dragow, F (2002). The work ahead: A psychometric infrastructure for computerized adaptive tests. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 67–88). Hillsdale, NJ: Lawrence Erlbaum.

Dörnyei, Z. (2003). Attitudes, orientations, and motivations in language learning: Advances in theory, research, and applications. *Language Learning*, 53(S1), 3-32.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Duchnicky, R. L., & Kolars, P. A. (1983). Readability of text scrolled on visual display terminals as a function of window size. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 25(6), 683-692.

Duncker, K. (1945). On problem-solving. In Dashiell, J. F. (Ed.) *Psychological Monographs* (pp.1–114). Washington, DC: American Psychological Association.

Dyson, M.C. (2004). How physical text layout affects reading from screen. *Behaviour & Information Technology*, 23(6), 377-393.

Dyson, M.C. (2005). How do we read text on screen. In Oostendorp, H. van, Breure, L., & Dillon, A. *Creation, Use, and Deployment of Digital Information*. Mahwah (N.J.), London: Lawrence Erlbaum Associates.

Bibliography

Dyson, M. C., & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, 54(4), 585-612.

Dyson, M.C. and Kipping, G.J. 1997. The legibility of screen formats: are three columns better than one? *Computers & Graphics*, 21, 703 – 712.

Dyson, M.C. and Kipping, G.J. 1998a. The effects of line length and method of movement on patterns of reading from screen. *Visible Language*, 32, 150 – 181.

Dyson, M.C. and Kipping, G.J. (1998b). Exploring the effect of layout on reading from screen. In R. D. Hersch, J. Andre Å.L and H. Brown (eds.) *Electronic Documents, Artistic Imaging and Digital Typography* (Berlin: Springer-Verlag), pp. 294 –304.

Earthman, E. A. (1992). Concurrent Verbal Reports in Second Language Acquisition Research. *Research in the Teaching of English*, 26, 351-384.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, 23, 249-261.

Ehara, K. (2008). *The effects of types of question on EFL learners' reading comprehension scores*. ProQuest.

Elkerton, J., & Williges, R. C. (1983, October). An Evaluation of Expertise in a File Search Environment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 27, No. 6, pp. 521-525). SAGE Publications.

Ericsson, K. & Simon, H. (1993). *Protocol Analysis*. Cambridge, Mass: MIT Press.

Evers, V. (1997). *Human-computer interfaces: Designing for culture* (Doctoral dissertation, Universiteit van Amsterdam).

Bibliography

Faiola, T. (1989). Principles and guidelines for screen display interface. *The Videodisc Monitor*, 8(2), 27-29.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27(3), 209-226.

Fernandes, T. (1995). *Global Interface Design*. London: Academic Press.

Fertig, M. (2003). Who's to blame? The determinants of German students' achievement in the PISA 2000 study.

Feng, X., & Mokhtari, K. (1998). Strategy use by native speakers of Chinese reading easy and difficult texts in English and Chinese. *Asian Journal of English Language Teaching*, 8, 19-40.

Field, J (2012) *Cognitive Validity* seminar talk at the CRELLA Spring Seminar, University of Bedfordshire, 15 March 2012.

Field, J. (2012). 1 The cognitive validity of the lecture-based question in the IELTS Listening paper. *IELTS Collected Papers 2: Research in Reading and Listening Assessment*, 2, 391.

Fitzgerald, A. S. C. O. R. E. (2012). Openness in English Language Teaching. *Proceedings of Cambridge 2012: Innovation and Im-pact-Openly Collaborating to Enhance Education*, 294.

Flowers, C., Kim, D. H., Lewis, P., & Davis, V. C. (2011). A Comparison of Computer-Based Testing and Pencil-and-Paper Testing for Students with a Read-Aloud Accommodation. *Journal of Special Education Technology*, 26(1), 1-12.

Fuhrer, S. (1973). A comparison of a computer-assisted testing procedure and standardized testing as predictors of success in community college technical mathematics (Doctoral dissertation, New York University, 1973). *Dissertation Abstracts International*, 34 (6), 3086.

Bibliography

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal*, 53(4), 289-299.

Fulcher, G. (2000). Computers in Language Testing. In Brett, P. & Motteram, G. (Ed) *A Special Interests in Computers: Learning and Teaching with Information and Communications Technologies*. Kent: IATEFL.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, 20(4), 384-408.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London, NY: Routledge.

Funke, J. (1998). Computer-based Testing and Training with Scenarios from Complex Problem-solving Research: Advantages and Disadvantages. *International Journal of Selection and Assessment*, 6(2), 90-96.

Gagné, R. H., & Smith, E. C. (1962). Concurrent Verbal Reports in Second Language Acquisition Research. *Journal of Experimental Psychology*, 63, 12-18.

Galitz, W. O. (2007). *The essential guide to user interface design: an introduction to GUI design principles and techniques*. John Wiley & Sons.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement*, 39(2), 133-147.

Galliers, R. D. (1991). Strategic information systems planning: myths, reality and guidelines for successful implementation. *European Journal of Information Systems*, 1(1), 55-64.

Gardiner, M. M. & Christie, B. Eds. (1987). *Applying cognitive psychology to user interface design*. Chichester: John Wiley & Sons.

Bibliography

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, 32(3), 301-319.

Geraci, M. G. (2002). *Designing Web-based instruction: A research review on color, typography, layout, and screen density* (Doctoral dissertation, Pacific University).

Geske, J. (2000). Readability of body text in computer mediated communications: Effects of type family, size and face. Retrieved November 23, 2012, from <http://www.public.iastate.edu/~geske/scholarship.html>

Glaser, B. G., & Strauss, A. L. (2009). *The discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers.

Goldberg, A. L. & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice.

Goldberg, A., Russell, M. and Cook, A. (2003) The Effect of Computers on Student Writing: A Meta-analysis of Studies from 1992 to 2002. *The Journal of Technology, Learning, and Assessment*, 2(1), available online at: <http://www.jtla.org>

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Literacy Research and Instruction*, 6(4), 126-135.

Gordon, C. J. (1990). Concurrent Verbal Reports in Second Language Acquisition Research. *Reading Horizons*, 31, 149-167.

Gorsuch, G. (2004). Test takers' experiences with computer-administered listening comprehension tests: Interviewing for qualitative explorations of test validity. *CALICO Journal*, 21(2), 339-371.

Bibliography

Gough, P. B. (1972). One second of reading. In J. F. Kavanagh and I. G. Mattingly (Eds.), *Language by ear and by eye*. Cambridge, Mass.: MIT Press.

Gould, J. D. and Grischowsky, N. 1984, Doing the same work with hard copy and cathode ray tube (CRT) computer terminals, *Human Factors* , 26, 323 ± 337.

Gould, J.D., Alfaro, L., Barnes, V., Finn, R., Grischowsky, N. and Minuto, A. (1987a) Reading is slower from CRT displays than from paper: attempts to isolate a single variable explanation. *Human Factors*, 29(3) 269-299.

Gould, J.D., Alfaro, L., Finn, R., Haupt, B. and Minuto, A. (1987b). Reading from CRT displays can be as fast as reading from paper. *Human Factors*, 29(5), 497-517.

Grabe, W. & Stoller, F.L. (2011). *Teaching and Researching Reading*, 2/E. London: Longman.

Grabinger, R. S. (1993). Computer screen designs: Viewer judgments. *Educational Technology Research and Development*, 41(2), 35-73.

Graesser, A. C. & Person, N. K. (1994). Question Asking During Tutoring. *American Educational Research Journal*, 31 (1), 104-137.

Greud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement*, 10(1), 23-34.

Green, B., Kingsbury, G., Lloyd, B., Mills, C., Plake, B., Skaggs, G., Stevenson, J., Zara, T., & Schwartz, J. (1995). *Guidelines for computerized-adaptive test development and use in education*. Washington, DC: American Council on Education Credit by Examination Program.

Green, A. 1998: *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Bibliography

Green, A. and Maycock, L. (2004). Computer based IELTS and paper based versions of IELTS *Research Notes* 18: 3-6.

Greene, S., & Higgins, L. (1994). "Once upon a time": The use of retrospective accounts in building theory in composition. In P. Smagorinsky (Ed.), *Speaking about writing* (pp. 115-140). Thousand Oaks, CA: Sage.

Grenfell, M. & Macaro, E. (2007). Language learner strategies – claims and critiques, in Cohen, A.D., Macaro, E. (Eds.), *Language Learner Strategies: 30 Years of Research and Practice*. Oxford University Press, Oxford, pp. 9-28.

Griffing, H. and Franz, S. I. (1896). On the Conditions of Fatigue in Reading. *Psychol. Rev.*, III, 513-520.

Haas, C. (1992). *Writing technology: Studies on the materiality of literacy*. Mahwah, NJ: Erlbaum.

Half-baked. (2004). Hot Potatoes. Version 6.0.3. Half-baked Software, inc. <http://web.uvic.ca/hrd/halfbaked/> (Accessed 9 November 2008).

Harrell, W. (1999). Effective monitor display design. *International Journal of Instructional Media*, v26 n4, 447-458.

Haynie, W. J. (1983). Student evaluation: The teacher's most difficult job. *Monograph Series of the Virginia Industrial Arts Teacher Education Council*, Monograph Number 11.

Hadadi, A., Luecht, R. M., Swanson, D. B., & Case, S. M. (1998). Study 1: Effects of modular subtest structure and item review on examinee performance, perceptions and pacing. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.

Bibliography

Hedgcock, J., & Ferris, D. R. (2009). *Teaching readers of English: Students, texts, and contexts*. Routledge.

Hetter, R., Segall, D. O., & Bloxom, B. M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W.A. Sands, B.K. Waters, and J.R. McBride (Eds.) *Computerized Adaptive Testing: From Inquiry to Operation*. Washington D.C.: American Psychological Association.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the Effect of Computer -Based Passage Presentation on Reading Test Performance. *The Journal of Technology, Learning, and Assessment*, 3(4), available online at: <http://www.jtla.org>

Higgins, J., Patterson, M.B., Bozman, M., & Katz, M. (2010). Examining the Feasibility and Effect of Transitioning GED Tests to Computer. *Journal of Technology, Learning, and Assessment*, 10(2), available online at: <http://www.jtla.org>.

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology*, 53(6), 826.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B. & Yan, F. (2006) Does it Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *The journal of Technology, Learning, and Assessment*, 5(2), available online at: <http://www.jtla.org>

Horton, S. (2000). *Web Teaching Guide: A practical guide to creating course web sites*. New Haven; Yale University Press.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2), 219-230.

Hudson, L. A., & Ozanne, J. L. (1988). Alternative ways of seeking knowledge in consumer research. *Journal of consumer research*, 508-521.

Bibliography

Huey, E. (1908). *The Psychology and Pedagogy of Reading*. New York: Macmillan.

International Telecommunication Union (2010). World Telecommunication/ICT Development Report 2010: Monitoring the WSIS targets. Geneva: ITU. Available online at: www.itu.int/dms_pub/itu-d/opb/ind/D-IND-WDTR-2010-PDF-E.pdf

Johnson, R. B., Onwuegbuzie, A.J. and Turner, L.A. (2007) Toward a definition of Mixed Methods Research. *Journal of Mixed Methods Research*, (1) 2, 112-133.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4), 329.

Kahn, P., & Lenk, K. (1998). Principles of typography for user interface design. *Interactions-New York-*, 15-29.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational measurement*, 38(4), 319-342.

Kapes, J. T., & Vansickle, T. R. (1992). Comparing paper-pencil and computer-based versions of the Harrington-O'Shea Career Decision Making System. *Measurement and Evaluation in Counseling and Development*, 25, 5-13.

Karoulis, A. & Pombortsis, A. (2004). The Heuristic Evaluation of Web-Sites Concerning the Evaluators' Expertise and the Appropriate Criteria List. *Informatics in Education, Vol. 3, (1)*, 55-74.

Kasper, G. (1998). Analysing verbal protocols. *Tesol Quarterly*, 32(2), 358-362.

Bibliography

Keng, L., McClarty, K. L., & Davis, L. L., (2008). Item-Level Comparative Analysis of Online and Paper Administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education, 21*, 207–226.

Kerry, T. (1987). Classroom questions in England. *Questioning Exchange, 1*(1), pp. 32-33.

Kesselman-Turkel, J., & Peterson, F. (2004). *Test-taking strategies*. University of Wisconsin Press.

Keyes, E. (1993). Typography, color, and information structure. *Technical Communication: Journal of the Society for Technical Communication, 4*, 638-654.

Khalifa, H. & Weir, C.J. (2009). *Examining Reading: Research and Practice in assessing second language reading*. CUP.

Kim, D. H., & Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement, 68*(4), 554-570.

Kim, J.-P. (1999, October). Meta-analysis of equivalence of computerized and P&P tests on ability measures. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review, 85*(5), 363.

Kirsh, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees (TOEFL Research Report 59)*. Princeton, NJ: Educational Testing Service.

Knezek, G. and Christensen, R. (1995). *A Comparison of Two Computer Curricular Programs at a Texas Junior High School Using the Computer Attitude Questionnaire (CAQ)*. Denton, TX: Texas Center for Educational Technology.

Bibliography

Klein, C., & Bederson, B. B. (2005, April). Benefits of animated scrolling. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems* (pp. 1965-1968). ACM.

Kline, P. (2000). *The handbook of Psychological Testing (second edition)*. London: Routledge.

Knezek, G. and Christensen, R. (1997). *Attitudes Toward Information Technology at Two Parochial Schools in North Texas*. Denton, TX: Texas Center for Educational Technology.

Kobrin, J. L. (2000). The Equivalence of Reading Comprehension Test Items Via Computerised and Paper-and-Pencil Administration. Unpublished PhD dissertation, The State University of New Jersey, USA.

Kobrin, J. L. & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education* 16(2),115-140.

Kobayashi, M. (1991). *On Validity of Reading Comprehension Tests*. Unpublished MEd thesis, University of Chiba, Japan.

Kobayashi, M. (1995). *Effects of Text Organisation and Test Format on Reading Comprehension Test Performance*. Unpublished PhD thesis, University of Thames Valley, UK.

Kobayashi, M. (2002). Methods Effects on Reading Comprehension Test Performance: Text organization and Response format. *Language Testing*, 19(2), pp.193-220.

Kolers, P. A., Duchnicky, R. L., & Ferguson, D. C. (1981). Eye movement measurement of readability of CRT displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 23(5), 517-527.

Korevaar, S. A.A. (2008) Equivalence of a Paper-Based and a Computer-Based Language Test of Reading as a Second Language. Unpublished MA-Dissertation. University of Luton, UK.

Bibliography

Kruk, R.S. and Muter, P. 1984, Reading of continuous text on video screens. *Human Factors*, 26, 339 – 345.

Kuehni, R. G., 2005. *Color: an Introduction to Practice and Principles* (2nd Ed). Wiley-Interscience, New Jersey.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, 6(2), 293-323.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longmans Green & Co.

Larson, J. 1999: Considerations for testing reading proficiency via computer-adaptive testing. In Chalhoub-Deville, M., editor, *Studies in language testing, Vol. 10. Issues in computer-adaptive testing of reading proficiency*. Cambridge: University of Cambridge Press, 71–90.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Lin, A. C. (1998). Bridging positivist and interpretivist approaches to qualitative methods. *Policy studies journal*, 26(1), 162-180.

Du Livre, L. C. (1912). *Lisibilite des Affiches en Couleurs*. Sheldons Limited House, Cosmos, 255.

Lee, J. (1986). The effect of mode of past computer experience on computerized aptitude performance. *Educational and Psychological Measurement*, 46, pp.727-733.

Lee, M. J., & Tedder, M. C. (2003). The effects of three different computer texts on readers' recall: based on working memory capacity. *Computers in Human Behavior*, 19(6), 767-783.

Bibliography

Lee, S. H., & Boling, E. (1999). Screen design guidelines for motivation in interactive multimedia instruction: A survey and framework for designers. *Educational Technology*, 39(3), 19-26.

Leeson, H. (2006). The Mode Effect: A Literature Review of Human and Technological Issues in Computerised Testing. *International Journal of Testing*, 6(1), pp.1-24.

Legge, G. E., Parish, D. H., Luebker, A., & Wurm, L. H. (1990). Psychophysics of reading XI. Comparing color contrast and luminance contrast. *Journal of the Optical Society of America A*, 7(10), 2002–2010.

Lewis, C. and Rieman, J. 1994: Task centered user interface design: a practical introduction. Available online at: <ftp://ftp.cs.colorado.edu/pub/distrib/clewis/HCI-Design-Book/>

Li, M. and Pu, H. (2010). Comparison between CBT and PBT: Assessment of Gap-filling and Multiple-choice Cloze in Reading Comprehension. *Journal of Language Teaching and Research*, Vol. 1, No. 6, pp. 935-941, November 2010.

Lieberman, D. (1979). Behaviourism and the mind: A (limited) call or a return to introspection. *American Psychologist*, 34, pp.319-33.

Ling, J., & Van Schaik, P. (2002). The effect of text and background colour on visual search of Web pages. *Displays*, 23(5), 223-230.

Lord, F. M. (1970). Some test theory for tailored testing. *Computer-assisted instruction, testing, and guidance*, W. H. Holtzman, Ed. New York: Harper & Row, 139-183.

Lund, O. (1999). Knowledge Construction in Typography: the Case of Legibility Research and the Legibility of Sans Serif Typefaces, PhD thesis (Reading University, UK).

Bibliography

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*, 31(3), 251-263.

Lynch, P. J., & Horton, S. Web Style Guide, Yale University, 2002.

Lynch, P. J., & Horton, S. (2009). *Web style guide: Basic design principles for creating Web sites*. Yale University Press.

Maguire, T., Hattie, J., & Brian, H. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research*. XL(2), pp. 109-26.

Manovich, L. (2001). *The language of new media*. MIT press.

Marks, A. M., & Cronje, J. C. (2008). Randomised Items in Computer-based Tests: Russian Roulette in Assessment?. *Educational Technology & Society*, 11 (4), 41–50.

Mason, B. J., Patry, M., & Berstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research*, 24(1), 29-40.

Maycock, L. and Green, A. 2005 The effects on performance of computer familiarity and attitudes to CB IELTS *Research Notes* 20:3-8.

McDonald, A. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers and Education*, 39, 299-312.

McDonough, S. (1995). *Strategy and skill in learning a foreign language*. London: Edward Arnold.

McDonough, S. (1999). Learner strategies. *Language Teaching*, 32, pp.1-18.

Bibliography

McMullin, J., Varnhagen, C. K., Heng, P., & Apedoe, X. (2002). Effects of surrounding information and line length on text comprehension from the web. *Canadian Journal of Learning and Technology/La revue canadienne de l'apprentissage et de la technologie*, 28(1).

McNamara, T. & Roever, C. (2006). *Language Testing: The Social Dimension*. Michigan: Blackwell Publishing.

Messick, S. (1980). Test Validity and the Ethics of Assessment. *American Psychologist*, 35(2), pp.1012-27.

Messick, S. (1989). Meaning and Values in Test Validation: the Science and Ethics of Assessment. *Educational Researcher*, 18(2), pp.5-11.

Messick, S. (1989). Validity. In Linn, R.L., editor, *Educational measurement*. 3rd edition. New York: American Council on Education/Macmillan Publishing Company, pp. 13–103.

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance. *Educational Measurement: Issues and Practice*, 15(2), pp.6-8.

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, 50, pp.741-49.

Messick, S. (1996). Validity and Washback in Language Testing. *Language Testing*, 13(3), pp.241-256.

Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 3–20). Boston: Kluwer Academic Publishers.

Bibliography

Mills, C.N., Potenza, M.T., Framer J.J. & Ward, W.C. (Eds.) (2002). *Computer-based testing: building the foundation for future assessments*. Routledge.

Moller, A.D. (1982). A Study in the Validation of Proficiency Tests of English as a Foreign Language. Unpublished PhD thesis. University of Edinburgh.

Muter, P. (1996). *Interface Design and Optimization of Reading of Continuous Text*. In van Oostendorp, H., and de Mul, S. (Eds.) (1996), *Cognitive aspects of electronic text processing*. Norwood, N.J.: Ablex.

Muter, P., & Maurutto, P. (1991). Reading and skimming from computer screens and books: the paperless office revisited?. *Behaviour & Information Technology*, 10(4), 257-266.

Nation, I. S. P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [Computer software]. *Wellington, NZ: LALS, Victoria University of Wellington, New Zealand*. Downloadable from <http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx>.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*, 22(1), 71-83. Organisation for Economic Cooperation and Development. (1987). *Information technologies and basic learning*. Paris: OECD Publications.

Nevo, N. (1989). Test-Taking Strategies on a Multiple-Choice Test of Reading Comprehension. *Language Testing*, 6(2), 199-215.

Nielsen, J. 1994: *Usability engineering*. San Francisco, CA: Morgan Kaufmann.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. *Proc. ACM CHI'94 Conf.* (Boston, MA, April 24-28), 152-158.

Bibliography

Nielsen, J. (1994b). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), *Usability Inspection Methods*, John Wiley & Sons, New York, NY.

Nielsen, J., and Molich, R. (1990). Heuristic evaluation of user interfaces, *Proc. ACM CHI'90 Conf.* (Seattle, WA, 1-5 April), 249-256.

Noyes, J. M., & Garland, K. J. (2003). VDT versus paper-based text: reply to Mayes, Sims and Koonce. *International Journal of Industrial Ergonomics*, 31(6), 411-423.

Noyes, J. M., & Garland, K. J. (2008). Computer-vs. paper-based tasks: Are they equivalent?. *Ergonomics*, 51(9), 1352-1375.

O'Dwyer, L. M., Russell, M., Bebell, D., & Tucker-Seeley, K. R. (2005). Examining the relationship between home and school computer use and students' English/language arts test scores. *Journal of Technology, Learning, and Assessment*, 3(3). Available from <http://www.jtla.org>

O'Hara, K., & Sellen, A. (1997, March). A comparison of reading paper and on-line documents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 335-342). ACM.

Onibere, E.A., Morgan, S., Busang, E.M. and Mpoeleng, D. (2001). Human-computer interface design issues for a multi-cultural and multi-lingual English speaking country: Botswana. *Interacting with Computers*, 13, 497-512.

Oostendorp, H., & Nimwegen, C. (1998). Locating information in an online newspaper. *Journal of Computer-Mediated Communication*, 4(1), 0-0.

O'Sullivan, B. (2000). *Towards a Model of Performance in Oral Language Testing*, unpublished PhD thesis, University of Reading.

Bibliography

Paek, P. (2005). Recent Trends in Comparability Studies. *Pearson, Educational Measurement*. Available online at: <http://www.pearsonedmeasurement.com/research/research.htm>

Paris, S. G., Wasik, B. A., & van der Westhuizen, G. (1988). Meta- cognition: A review of research on metacognition and reading. In J. E. Readence, & R. S. Baldwin (Eds.), *Dialogues in literacy research (37th yearbook of the National Reading Conference* (pp. 143-166). Chicago, IL: The National Reading Conference.

Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher*, 20(5), 12-20.

Parshall, C.G., Spray, J.A., Kalohn, J.C. & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Patton, M.Q. (2002). *Qualitative Research and Evaluation Methods*. Thousand Oaks, CA: Sage Publications.

Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.

Peters, R. (1992). Designing for the Computer Screen. *Designing information: new roles for librarians*, 1992, 147.

Phakiti, A. (2003a). A closer look at gender differences in strategy use in L2 reading. *Language Learning*, 53, pp. 649-702.

425 Phakiti, A. (2003b). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading comprehension test performance. *Language Testing*, 20, pp. 26-56.

Phakiti, A. (2006). Modeling cognitive and metacognitive strategies and their relationships to EFL reading test performance. *Melbourne Papers in Language Testing*, 11(1), pp.53-95.

Bibliography

Piolat, A., Roussey, J. Y., & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, 47(4), 565-589.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web Based and Paper and Pencil Testing of Applicants in a Proctored Setting: Are Personality, Biodata, and Situational Judgment Tests Comparable? *Personnel Psychology*, 56(3), 733-752.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment*, 3(6). Available from <http://www.jtla.org>

Polit, D. F., & Beck, C. T., Hungler, BP (2001). *Essentials of nursing research. Methods, appraisal, and utilization*. Fifth edition. Philadelphia PA, Lippincott: Williams & Wilkins.

Pollock, J. C., & Sullivan, H. J. (1990). Practice mode and learner control in computer-based instruction. *Contemporary Educational Psychology*, 15(3), 251-260.

Pommerich, M. (2002, April). The effect of administration mode on test performance and score precision, and some factors contributing to mode differences. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effect for passage-based tests. *Journal of Technology, Learning and Assessment*, 2(6), pp. 1-44.

Pommerich, M. (2007). The Effect of Using Item Parameters Calibrated from Paper Administrations in Computer Adaptive Test Administrations. *The Journal of Technology, Learning, and Assessment*. 5 (7), available online at: <http://www.jtla.org>

Popham, W. J. (1997). Consequential validity: Right Concern-Wrong Concept. *Educational measurement: Issues and practice*, 16(2), 9-13.

Bibliography

Pressley, M. & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.

Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main ideas and do not realize it: Confidence in response to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly*, 25, pp. 232-249.

Pritchard, R. (1990). Concurrent Verbal Reports in Second Language Acquisition Research. *Reading Research Quarterly*, 25, 273-295.

Proctor, R. W., & Vu, K. P. L. (Eds.). (2005). *Handbook of human factors in Web design*. L. Erlbaum Associates.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *The Journal of Technology, Learning, and Assessment*. 6 (3), available online at: <http://www.jtla.org>

Rickets, C. & Wilks, S.J. (2002). Improving student performance Through Computer-based Assessment: Insights from recent research. *Assessment & Evaluation in Higher Education*, 27(5), 475-479.

Rivlin, C. Lewis, R. & Davies-Cooper, R. "Guidelines For Screen Design." Blackwell Scientific Publications. Oxford: 1990.

Roethlein, B. E. (1912). The relative legibility of different faces of printing types. *The American Journal of Psychology*, 1-36.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology*, 5(2), 84-94.

Bibliography

Roever, C., & McNamara, T. (2006). Language testing: the social dimension. *International Journal of Applied Linguistics*, 16(2), 242-258.

Rosenshine, B V (1980) Skill Hierarchies in Reading Comprehension, in Spiro, R J, Bruce, B C and Brewer, W F (Eds), *Theoretical issues in reading comprehension*, Hillsdale, NJ: Lawrence Erlbaum, 535–559.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, 14(1), 15-22.

Rowan, B. E., (2010). *Comparability of Paper-and-Pencil and Computer-Based Cognitive and Non Cognitive Measures in a Low-Stakes Testing Environment*. Unpublished PhD-Thesis. University of James Madison, USA.

Russell, M. (1999). Testing on computers: a follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, 7(20).

Russell, M. & Haney (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Educational Policy Analysis Archives*, 5(3).

Russell, M. (1999) 'Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper'. *Educational Policy Analysis Archives*, 7(20). Retrieved February 1, 2010, from <http://epaa.asu.edu/epaa/v7n20/>

Russell M. & Haney.W. (2000). Bridging the gap between testing and technology in schools. *Education Policy Analysis Archives*, 8(19). Retrieved February 1, 2010, from: <http://epaa.asu.edu/epaa/v8n19.html29>

Russell, M. & Plati, T. (2001). Effects of computer versus paper administration of a state mandated writing assessment. Teachers College. Retrieved February, 2009, from: <http://www.tcrecord.org>

Bibliography

Russo, P. & Boor, S. (1993). How fluent is your interface? designing for international users. *Human Factors in Computing Systems*, 24-29. Proceedings of INTERCHI '93, Amsterdam. New York: ACM.

Sandene, B., Horkay, N., Bennett, R.E., Allen, N., Braswell, J., Kaplan, B., Oranje, A., (2005) Reports From the NAEP Technology-Based Assessment Project, Research and Development Series. NCES: Jessup, USA. US. Department of Education. Available online at: <http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf>

Sarig, G. (1987). High-Level reading in the first and the foreign language: some comparative process data. In Carrell, P. Devine, J. and Eskey, D. (Eds). *Research in Reading English as a Second Language*, pp.105-20. Washington, D.C.: TESOL.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. In I. Thompson (Ed.), *Special issue on computer-assisted language testing. Language Learning and Technology*.

Schaeffer, R.S. & Bateman, W. (1996). So many colors, so many choices: The use of color in instructional multimedia products. In Proceedings of selected research and development presentations at the 1996 national convention of the Association for Educational Communications and Technology. Indianapolis, IN: Association for Educational Communication and Technology.

Schedl, M, Gordon, A, Carey, P A and Tang, K L (1996) *An Analysis of the Dimensionality of TOEFL Reading Comprehension Items*, (TOEFL Research Reports 53) Princeton, NJ: Educational Testing Services.

Scholfield, P. (1995). *Quantifying language: A researcher's and teacher's guide to gathering language data and reducing it to figures*. Multilingual matters.

Bibliography

Shaikh, A. D. (2005). The effects of line length on reading online news. *Usability News*, 7(2), 1-4.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education*, 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-24.

Shermis, M. and Lombard, D. (1998). Effect of computer-based test administration on test anxiety and performance. *Computer in Human Behaviours*, 14(1), pp.111-23.

Shiotsu, T. (2010). *Components of L2 reading: linguistics and processing factors in the reading test performances of Japanese EFL learners*. Cambridge University Press. Cambridge, UK.

Shrock, S. A. & Coscarelli, W. C. (2007). *Criterion-Referenced Test Development: Technical and Legal Guidelines for Corporate Training*. San Francisco: Pfeiffer.

Sim, G., Read, J., & Holifield, P. (2007). *Evaluating the user experience in CAA environments: what affects user satisfaction?* Paper presented at the 10th International Computer Assisted Assessment Conference, Loughborough, UK.

Simpson, A. J. (2010). SL BLOG Since March 2009–Sabancı University, School of Languages, Istanbul/Turkey. *EMU*, 5, 7.

Skaalid, B. (2001). Web design for instruction: research-based guidelines. *Canadian Journal of Educational Communication*, v27 n3, 139-155.

Smith, B. (2003). *Conventional versus Computer-Based Administration of Measures of Cognitive Ability: An Analysis of Psychometric, Behavioural, Experiential and Relativity of Equivalence*. Unpublished PhD Thesis. University of Wollongong, USA.

Bibliography

Standards for educational and psychological testing (1999). Washington, DC: American Educational Research Association.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading research quarterly*, 32-71.

Sternberg, R. J. (1991). Are I reading too much into reading comprehension tests? *Journal of Reading*, 34(7), pp.540-45.

Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Language Testing*, 14(2), 214–31.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The Relationship between Computer Familiarity and Performance on Computer-Based TOEFL Test Tasks (TOEFL Research Report 61)*. Princeton, NJ: Educational Testing Service.

Taptagaporn, S., & Saito, S. (1990). How display polarity and lighting conditions affect the pupil size of VDT operators. *Ergonomics*, 33(2), 201-208.

Thompson, N. A. (2008). 'A Proposed Framework of Test Administration Methods'. *Journal of Applied Testing Technology*, 2008, 9 (5), pp. 1-15.

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations* (Synthesis Report 78). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tinker, M. A. (1963). *Legibility of print*. Ames, IA: Iowa State University Press.

Tinker, M. A. (1966). *Bases for Effective Reading*. Minneapolis, Minn.: University of Minnesota Press.

Bibliography

Tinker, M. A., & Paterson, D. G. (1931). Studies of typographical factors influencing speed of reading. VII. Variations in color of print and background. *Journal of Applied Psychology*, 15(5), 471.

Tognazzini, B. (2000). *Ask Tog: First Principles*. Retrieved 18 Nov. 2010:
<http://www.asktog.com/basics/firstPrinciples.html>

Trickett, S. B., & Trafton, J. G. (2007). A primer on verbal protocol analysis. *Handbook of virtual environment training*. Westport, CT: Praeger Security International.

Trochim, W. (2000). *The Research Methods Knowledge Base, 2nd Edition*. Atomic Dog Publishing, Cincinnati, OH.

Urquhart, A, and Weir, C J, 1998, *Reading in a second language: process, product and practice*. London: Longman.

Van Blerkom, M. L. (2009). *Measurement and Statistics for Teachers*. NY: Routledge.

Van de Vijver, F. J. R., & Harsveld, M. (1994). 'The incomplete equivalence of the paper-and pencil and computer versions of the General Aptitude Test Battery'. *Journal of Applied Psychology*, 79, 852- 859.

van der Linden, W. J. (2005). *Linear models for optimal test design*. NY: Springer Science+Business Media, Inc.

van der Linden, W. J., & Glas, C. A. W., (Eds.)(2000). *Computerized adaptive testing: theory and practice*. Dordrecht: Kluwer Academic Publishers.

van Teijlingen, E., & Hundley, V. (2002). The importance of pilot studies. *Nursing Standard*, 16(40), 33-36.

Bibliography

Vann, R., & Abraham, R. (1990). Strategies of unsuccessful language learners. *TESOL Quarterly*, 24, 177-195.

Vendlinks, T. & Stevens, R. (2002) Assessing Student Problem-Solving Skills With Complex Computer-Based Tasks. *The Journal of Technology, Learning, and Assessment*, 1(3), available online at: <http://www.jtla.org>

Wainer, H. (1990). *Computer adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, 12, 15–20.

Walliman, N. (2006). *Social research methods*. Sage.

Walter, O. B., Becker, J. Blorner, J. B., Fliege, H., Klapp, B. F. And Rose, M. (2007). Development and evaluation of a computer adaptive test for ‘Anxiety’ (Anxiety-CAT). *Quality of life Research*, 16(1), pp. 143-155.

Waters S.D., Pommerich M. (2007) Context effects in internet testing: A literature review. 22nd Annual Conference of the Society for Industrial and Organizational Psychology.

WebCT Inc. (2004) WebCT: Learning without limits, retrieved November 30, 2009, from: <http://www.webct.com/>

Weir, C. J. (1988). *Communicative Language Testing*. University of Exeter, UK.

Weir, C. J. (1990). *Communicative language testing*. London: Prentice Hall.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. New York: Palgrave Macmillan.

Bibliography

Weir, C. J., Hawkey, R. Green, T., Devi, S. (2009a) *The cognitive processes underlying the academic reading construct as measured by IELTS*. British Council/IDP Australia, research Reports Volume 9: 157-189.

Weir, C.J., O'Sullivan, B. and Jin Yan. (2007) *Does the computer make a difference? Reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS writing component: effects and impact. IELTS Research Report No.7*, British Council & IDP Australia.

Weir, C. J. Yang, H. and Jin, Y. (2000). An empirical investigation of the componentiality of L2 reading in English for Academic Purposes. *Studies in Language Testing*, vol 12, UCLES Cambridge University Press, Cambridge.

Weisenmiller, E. M. (1999). A Study of the Readability of On-screen Text. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. *Psychological Review*, 87, pp. 105-112.

Wiberg, M. (2003). An optimal design approach to criterion-referenced computerized testing. *Journal of Educational and Behavioral Statistics*, 28, pp. 97-110.

Wiley, D. (1991). Test Validity and Invalidity Reconsidered. In R. E. Snow & D. E. Wiley (Eds). *Improving inquiry in the social sciences: A volume in honour of Lee J. Cronbach* (pp.75-107). Hillsdale, NJ: Erlbaum.

Williams, E. and Moran, C. (1989). Reading in a foreign language at intermediate and advanced levels with particular reference to English. *Language Teaching* 22(4), 217–28.

Wise, S. L. (1996). A critical analysis of the arguments for and against item review in computerized adaptive testing. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Bibliography

Wolfe, E., & Manalo, J. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning and Technology*, 8(1), 53-65.

Wright, P., & Lickorish, A. (1983). Proofreading texts *on* screen and paper. *Behavior and Information Technology*, 2, 227-235.

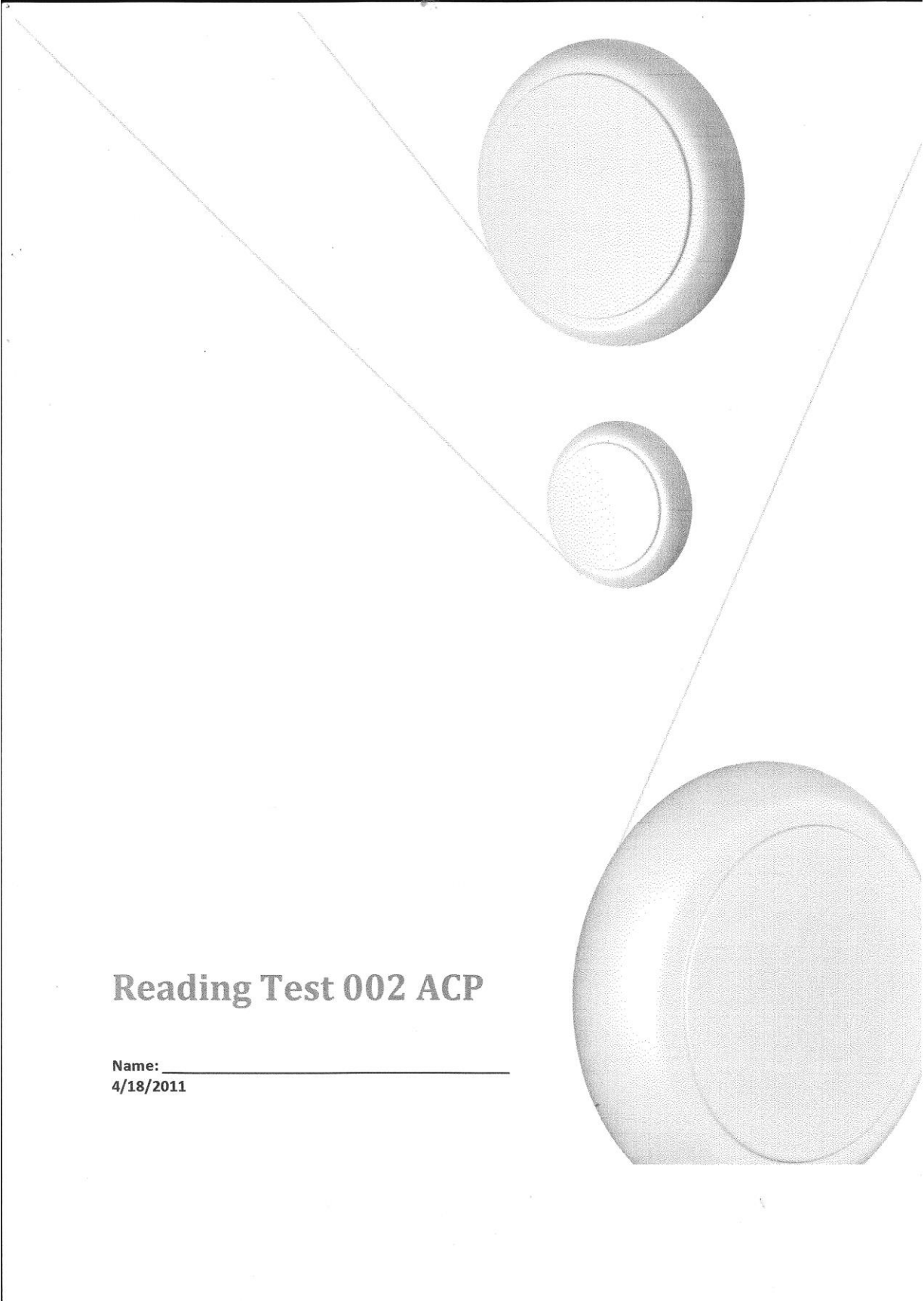
Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly*, 7(2), 119-136.

Zandvliet, D. & Farragher P. (1997). 'A comparison of computer-administered and written tests'. *Journal of Research on Computing in Education*, 29 (4), pp. 423-438.

Zheng, Y. & De Jong, J.H.A.L. (2011). Establishing Construct and Concurrent Validity of Pearson Test of English Academic. Research Note, Pearson Education Ltd.

Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(4), pp.554-568.

Appendix A: Study's Reading Test



Reading Test 002 ACP

Name: _____
4/18/2011

Read the passages below and answer questions 1 up to 30. You have 75 minutes.

Reading Test Part 1

Jonathan Wilson was born in Mexico City. Maria Gomez was born in New York. Jonathan grew up in a small village. When he was seventeen, he moved to the United States. He began to study English at night, and he worked in a factory during the day. He met Maria, who was a teacher, at the school. They got married and now they have a son and a daughter, Victor and Margaret. Mr. Wilson is a very active businessman and his wife is a singer. She is very famous. She sings on TV. When she isn't working, she always stays at home and does yoga. On Saturday, the Wilsons spend time together. They all work in the garden. On Sunday, they rest and have a big dinner together. Every weekend is special to them. It's their only time together as a family. The Wilsons have many friends in the city and they're all very interesting people. Last week, Mr. and Mrs. Wilson invited them to a party. Their guests arrived at about seven thirty and they all sat in the dining room. They ate cheese and crackers, drank lemonade and looked at the Wilsons' photographs. Mr. Wilson gets up every day at seven fifteen. He does his morning exercises for ten minutes and takes a shower. At seven forty-five he eats breakfast. Then, he talks to his wife before she goes to work. At eight thirty he leaves the house and drives to work. Mr. Wilson is the president of the Job Access Agency. He meets different people every day. He speaks English, French and a little Chinese. He has a staff of very energetic employees. His office assistant sorts the mail every morning and he talks on the telephone all day. His receptionist speaks only French. She types letters in French and talks to French people. This morning, Mr. Wilson got up late. He didn't do his exercises and he left home at eight forty-five. He arrived at work at nine fifteen and he had a very bad day at the office. He typed letters all day because the receptionist who usually types them is on vacation. Now he doesn't feel well. He has a headache.

Tonight, the Wilsons are going to celebrate New Year. They hope next year is going to be a very good year for the entire family. Mr. and Mrs. Wilson like Japan. They are going to take a long vacation there in summer. Their daughter is going to move to another city and their son is going to get his driver's license and buy a car.

1. How many children do Mr. and Mrs. Wilson have?

2. What is Mrs. Wilson's occupation?

3. What do Victor and Margaret do on Saturday?

4. What did the Wilsons' friends eat at the party?

5. What does Mr. Wilson usually do at quarter to eight?

6. How does Mr. Wilson usually go to work?

7. What language does Mr. Wilson's receptionist speak?

8. What time did Mr. Wilson leave for work this morning?

9. How did Mr. Wilson get a headache today?

10. Where are Mr. and Mrs. Wilson going to go next year?

Reading Test Part 2

Paul Newman was born in Cleveland, Ohio, in 1925, and did some acting in high school and college, but never seriously considered making it his future career. However, after graduating, he started working in the theatre and on several TV shows in New York. When he was thirty, he went to Los Angeles and made his first film. It was what he called an 'uncomfortable' start in the movies, in the role of a Greek slave.

The next film he chose was his big break. He played the role of the boxer, Rocky Graziano in the film 'Someone up There Likes Me'. Newman is a method actor who believes in living the role before beginning the film. He spent days – from morning till night – with Graziano. He studied the boxer's speech and watched **him** box. The picture brought Newman stardom overnight.

Newman went on to make films such as 'Cat on a Hot Tin Roof', 'The Hustler', 'Butch Cassidy and the Sundance Kid', 'The Sting' and 'Towering Inferno'. He has made over forty-five films and has won many awards, but he has never won an Oscar.

He was living in Los Angeles when he became engaged to Joanne Woodward, an actress whom he had first known in New York. Newman and Miss Woodward were married in Las Vegas in 1958. His marriage to Woodward is one of the longest and strongest in Hollywood. They have co-starred in six films. Ever since the film 'Winning', Newman has been interested in car racing, and in 1979 he came second in the twenty-four hour Le Mans race. He has a strong social conscience, and has supported causes such as the anti-nuclear movement, the environment, and driver education. All the money from 'Newman's Own' salad dressing, popcorn, and spaghetti sauce, now a multi-million-dollar business, goes to charity.

1. When did Newman first work in the theatre?

2. When did Newman make his first film?

3. Which film made Newman a star?

4. What is a method actor?

5. Where did Newman first meet Woodward?

6. How many films did Newman and Woodward make together?

7. When did Newman's interest in car racing start?

8. What's the name of Newman's company?

9. "It was what he called an 'uncomfortable' start." What does "it" refer to in line 4?

10. "He studied the boxer's speech and watched **him** box." What does "him" refer to in line 9?

Reading Test Part 3

When James Bond got back to his hotel room it was midnight. His windows were closed and the air-conditioning was on. Bond switched it off and opened the windows. His heart was still thumping in his chest. He breathed in the air with relief, then had a shower and went to bed.

At 3:30 he was dreaming, not very peacefully, of the three black-coated men with red eyes and angry white teeth, when suddenly he woke up. He listened. There was a noise. It was coming from the window. Someone was moving behind the curtains. James Bond took his gun from under his pillow, got quietly out of bed, and crept slowly along the wall towards the window. Someone was breathing behind the curtains. Bond pulled them back with one quick movement. Golden hair shone silver in the moonlight. 'Mary!' Bond asked, 'What are you doing here?'

'Quick, James! Help me in,' she whispered. Bond put down his gun and tried to pull her through the open window. At the last moment her foot got caught in the curtain and the window banged shut with a noise like a gunshot. Mary whispered, 'I'm terribly sorry, James!' 'Sh! Sh!' said Bond, and quickly led her across the room to the bathroom. He turned on the light and the shower. They sat down on the side of the bath. Bond asked again, 'what are you doing here? What's the matter?' 'James, I was so worried. A message came from HQ this evening. A top KGB man, using the name Hendriks, is staying at this hotel. I knew you were looking for him, but he knows you're here. He's looking for you!'

'I know,' said Bond. 'That man's here all right. So is a gunman called Scaramanga. Mary, did HQ say if Hendriks has got a description of me?'

'No, he hasn't. You were just described as secret agent James Bond.'

'Thanks, Mary. Now I must get you out of here.' Bond turned off the shower and opened the bathroom door. 'Now, come on.'

A voice came from the darkness of the bedroom. 'This is not your lucky day, Mr Bond. Come here both of **you**. Put your hands behind your necks.'

Scaramanga walked to the door and turned on the lights. His golden gun was pointing directly at James Bond.

1. When did James get back to his room?

2. Where did James keep his gun?

3. Where was Mary hiding before James woke up?

4. Why did the window bang shut?

5. Where were James and Mary sitting while they were talking?

6. Who was James looking for?

7. Who did Hendriks work for?

8. What's the gunman's name?

9. "Bond switched **it** off." What does "it" refer to in line 3?

10. "Come here both of **you**." What does "you" refer to in line 27?

Thank You

Appendix B: Computer Familiarity Questionnaire English Version

Name _____ Gender _____
 Nationality _____ Institution _____
 Please tick (✓) one box only on each line

Q 1 **How often is there a computer available to you to use at these places ?**

	Almost every day	A few times each week	Between once a week and once a month	Less than once a month	Never
a) At home	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
b) At university/college	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
c) In the library	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
d) At another place	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 2 **How comfortable:**

	Very comfortable	Comfortable	Somewhat comfortable	Not at all comfortable
a) are you with using a computer?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
b) are you with using a computer to write a paper?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
c) would you be taking a test on a computer?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 3 **How would you rate your ability to use a computer ?**

	Excellent	Good	Fair	Poor
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 4 **How often do you use a computer:**

	Almost every day	A few times each week	Between once a week and once a month	Less than once a month	Never
a) at home	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
b) at university/college	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
c) in the library	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
d) at another place	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 5 **How often do you use :**

	Almost every day	A few times each week	Between once a week and once a month	Less than once a month	Never
a) the Internet?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
b) a computer for e-mail/chat?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
c) the computer for school/studies?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
d) de computer for programming?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 6 How often do you use each of the following kinds of computer software ?

	Almost every day	A few times each week	Between once a week and once a month	Less than once a month	Never
a) Games	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
b) Word processing (e.g. Word ®)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
c) Spreadsheets (Excel ®, Lotus 123®)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
d) Drawing, painting, graphics	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
e) Data, text analysis(SPSS ®)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 7 How do you feel about using the keyboard (typing)

	Totally agree	agree	No feeling	Disagree	Totally disagree
a) I can type as fast as I can write	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
b) I do not think it is a problem for me	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
c) I find using the keyboard difficult	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 8 Have you ever taken a reading test on computer?

Yes No

Thank You

Appendix C: Computer Familiarity Questionnaire Arabic Version

الاسم _____ الجنس _____ الجنسية _____ المؤسسة _____ ضع إشارة (✓) في مربع واحد في كل سطر

Q 1 كم عدد المرات التي يتوفر لك جهاز حاسب آلي لاستخدامك في الامكنة التالية: ?

لا يتوفر أبدا	أقل من مرة شهريا	بين مرة أسبوعيا و مرة شهريا	عدة مرات في الأسبوع	تقريبا كل يوم
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

1 في البيت
2 في الجامعة / الكلية
3 في المكتبة
4 في مكان آخر

Q 2 ما هو مدى مهارتك في :
1 استخدام الحاسب الآلي
2 استخدام الحاسب الآلي لكتابة بحث
3 اخذ اختبار بواسطة جهاز الحاسب الآلي

غير قادر على استخدامه	متوسط المهارة	ماهر	ماهر جدا
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 3 كيف تقيم قدرتك على استخدام الحاسب الآلي ?

ممتاز	جيد	متوسط	ضعيف
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Q 4 كم عدد المرات التي تستخدم فيها الحاسب الآلي في الامكنة التالية:

لا يتوفر أبدا	أقل من مرة شهريا	بين مرة أسبوعيا و مرة شهريا	عدة مرات في الأسبوع	تقريبا كل يوم
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

1 في البيت
2 في الجامعة / الكلية
3 في المكتبة
4 في مكان آخر

Q 5 كم مرة تستخدم

:

لا يتوفر أبدا	أقل من مرة شهريا	بين مرة أسبوعيا و مرة شهريا	عدة مرات في الأسبوع	تقريبا كل يوم
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

1 الانترنت
2 البريد الإلكتروني أو الدردشة
3 الحاسب الآلي للدراسة
4 استخدام الحاسب للبرمجة

Q 6 كم عدد المرات التي تستخدم فيها تطبيقات الحاسب الآلي التالية ?

لا يتوفر أبدا	أقل من مرة شهريا	بين مرة أسبوعيا و مرة شهريا	عدة مرات في الأسبوع	تقريبا كل يوم
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

1 الألعاب
2 معالج الكلمات (الوورد)
3 برنامج الإكسل
3 الرسم و برامج الصور
4 تحليل البيانات

Q 7 ما هو تقييمك لاستخدام لوحة المفاتيح (الكي بورد)

غير موافق بشدة	غير موافق	محايد	موافق	موافق بشدة
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

1 أطيح بنفس سرعة كتابتي
2 لا تمثل مشكلة لي
3 اجد استخدامها صعبا

Q 8 هل سبق لك وان أجريت اختبار قراءة باستخدام الحاسب?

نعم لا

Thank You

Appendix D: Post-Test Questionnaire English Version

Name _____ Gender _____ Nationality _____
 Institution _____

Please tick (✓) one box only on each line

		1	2	3 No Diff	4	5	
1. The questions on screen were easy for me to read.	Strongly Disagree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Strongly Agree
2. The questions on paper were easy for me to read.	Strongly Disagree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Strongly Agree
3. The way I read the text was different on computer than on paper. Why? _____ _____	Strongly Disagree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Strongly Agree
4. The size of the computer screen was big enough.	Strongly Disagree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Strongly Agree
5. The way I answered the questions was different on computer than on paper. Why? _____ _____	Strongly Disagree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Strongly Agree
6. It was easy to navigate through the test using the navigation buttons.	Strongly Disagree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Strongly Agree
7. Using the scrolling feature was not problematic for me.	Strongly Disagree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Strongly Agree
8. I think I did better on the computer-based test than on the paper-based test. Why? _____ _____	Yes <input type="checkbox"/>			No <input type="checkbox"/>			

9a. Did you feel at any time that you did not know what to do?

Never <input type="checkbox"/>	Rarely <input type="checkbox"/>	Sometimes <input type="checkbox"/>	Very Often <input type="checkbox"/>	Always <input type="checkbox"/>
-----------------------------------	------------------------------------	---------------------------------------	--	------------------------------------

9b. Why?

10a. Was there any time that you wanted to ask for help?

Never <input type="checkbox"/>	Rarely <input type="checkbox"/>	Sometimes <input type="checkbox"/>	Very Often <input type="checkbox"/>	Always <input type="checkbox"/>
-----------------------------------	------------------------------------	---------------------------------------	--	------------------------------------

10b.

Why?

11. In which test was the text easier to read?	Computer <input type="checkbox"/>	Paper <input type="checkbox"/>	No Difference <input type="checkbox"/>
12. Which test did you prefer taking? Why? _____ _____	Computer <input type="checkbox"/>	Paper <input type="checkbox"/>	No Difference <input type="checkbox"/>
13. In which test was it easier to write down answers? Why? _____ _____	Computer <input type="checkbox"/>	Paper <input type="checkbox"/>	No Difference <input type="checkbox"/>
14. In Which test was it easier to change answers? Why? _____ _____	Computer <input type="checkbox"/>	Paper <input type="checkbox"/>	No Difference <input type="checkbox"/>
15. In which test were the reading passages easier to navigate through?	Computer <input type="checkbox"/>	Paper <input type="checkbox"/>	No Difference <input type="checkbox"/>

Thank You

Appendix E: Post-Test Questionnaire Arabic Version

الاسم _____	الجنس _____	الجنسية _____
المؤسسة العملية _____		
ضع إشارة (✓) في المربع لكل سطر		

قبل قراءة الأسئلة:

1... على الاختبار الورقي

1... على الاختبار بواسطة الحاسب

أقرأ النص أو جزء منه ببطء A	<input type="checkbox"/>	أقرأ النص أو جزء منه ببطء A	<input type="checkbox"/>
اقرأ النص بسرعة و بانتقاء لأخذ فكرة عامة عن الموضوع B.	<input type="checkbox"/>	اقرأ النص بسرعة و بانتقاء لأخذ فكرة عامة عن الموضوع B.	<input type="checkbox"/>
لا أقرأ النص وأتوجه مباشرة إلى الأسئلة C.	<input type="checkbox"/>	لا أقرأ النص وأتوجه مباشرة إلى الأسئلة C.	<input type="checkbox"/>

		1	2	3	4	5	
1. الأسئلة على الشاشة كانت سهلة بالنسبة لي للقراءة.	لا أوافق بشدة	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	أوافق بشدة
2. الأسئلة على الورقة كانت سهلة بالنسبة لي للقراءة.	لا أوافق بشدة	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	أوافق بشدة
3. طريقة قراءة الأسئلة من الشاشة تختلف عنها من الورقة. لماذا؟ _____	لا أوافق بشدة	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	أوافق بشدة
4. حجم شاشة الحاسب كان مناسباً.	لا أوافق بشدة	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	أوافق بشدة
5. طريقة إجابة الأسئلة على الحاسب كانت مختلفة عنها في الورقة. لماذا؟ _____	لا أوافق بشدة	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	أوافق بشدة
6. من السهل التنقل بين أجزاء الامتحان بواسطة زر الانتقال	لا أوافق بشدة	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	أوافق بشدة
7. استخدام خاصية الانتقال بين أجزاء الامتحان لم يكن مشكلة بالنسبة لي	لا أوافق بشدة	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	أوافق بشدة
8. الكلمات المستخدمة في أزرار الانتقال بين الأسئلة كانت واضحة.	لا أوافق بشدة	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	أوافق بشدة

9a. هل شعرت في أي لحظة أنك لا تعلم ما الذي يجب فعله.

مطلقا	نادرا	أحيانا	غاليا	دائما
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

?

9b. لماذا

10a. هل احتجت المساعدة في أي وقت بالامتحان.

مطلقا	نادرا	أحيانا	غاليا	دائما
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

10b.

لماذا

?

11. أي الاختبارين كان أسهل للقراءة?	الحاسب	الورقي	لا فرق
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. أي الاختبارين تفضل؟ لماذا _____ _____	الحاسب	الورقي	لا فرق
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. أي الاختبارين كان أسهل لكتابة الأجوبة. لماذا _____ _____	الحاسب	الورقي	لا فرق
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. في أي الاختبارين كان من السهل تغيير الإجابة. لماذا _____ _____	الحاسب	الورقي	لا فرق
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15. في أي الاختبارين كان من السهل التنقل خلال النص.	الحاسب	الورقي	لا فرق
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

16. اعتقد أنني أدبت الاختبار (الورقي/ الحاسوبي) بشكل أفضل من الاختبار (الورقي/ الحاسوبي).

لماذا _____

Thank You

Appendix F: University's Placement Test

Placement Exam



**Deanship of Preparatory Year
English Language Centre**

C

Student Name:		اسم الطالب:																				
Student ID#		الرقم الجامعي:																				
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20px; text-align: center;">2</td> <td style="width: 20px; text-align: center;">0</td> <td style="width: 20px; text-align: center;">1</td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> </tr> </table>	2	0	1								<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20px; text-align: center;">2</td> <td style="width: 20px; text-align: center;">0</td> <td style="width: 20px; text-align: center;">1</td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> </tr> </table>	2	0	1								
2	0	1																				
2	0	1																				
National ID #		رقم الهوية الوطنية:																				
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> </tr> </table>											<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> <td style="width: 20px;"></td> </tr> </table>											

Do not open this paper until you are told to do so.

Read the following instructions carefully:

- Write your name, student ID number and national ID number in the boxes above.
- Obey the exam rules: You must not talk to another student at any time.
 You must not look at another student's paper.
 All mobile phones must be switched off.
- If you break the rules, you will be reported and **you will get the appropriate disciplinary action.**
- If you need help at any time, raise your hand and **wait** for an invigilator to come to you.
- When you have finished, you may leave. Raise your hand and **wait** for an invigilator to take your paper **before** you leave your seat.
- **Before leaving** the examination room, you must **give** your paper to the invigilator. **If you do not, you may lose all the marks for this paper.**

Questions	Weighting	Final Mark	
1	4		
2	9		Subtotal
3	7		/20
4	6		
5	3		Subtotal
6	5		/20
7	6		
8	3		
9	6		Subtotal
10	5		/20
11	6		
12	4		Subtotal
13	8		/20
14	8		
GRAND TOTAL			
			/80

1. Fill in the gaps with capital or small letters. املا الفراغات التالية بالحروف الصحيحة

Small letters	f			g	q
Capital letters	F	B	H		

4

2. Fill in the spaces with the correct pronoun from the box. املا الفراغات التالية بالكلمات الصحيحة من الجدول

he	I	she	it	we	they	you
----	---	-----	----	----	------	-----



1. = he

2. a car = _____

9



3. = _____

4. Faisal = _____



5. = _____

6. Fatima, Yasmin and Noura = _____



7. = _____

8. My father = _____

9. cars = _____

10. the teacher and I = _____

3. Complete the sentences with *am, is or are*.

أكمل الجمل التالية ب *am, is or are*

- My friend _____ from Australia.
- We _____ tall.
- My name is Jane but I _____ not a teacher.
- They _____ my teachers.
- He _____ from Japan.
- _____ you American?
- This car _____ new.

7

4. Fill in the spaces with the correct words from the box. أكمل الجمل التالية بالكلمات الصحيحة من الجدول.

are	she	do	we	is	they	am	a	an	he	not
-----	-----	----	----	----	------	----	---	----	----	-----

1. It _____ apple.
2. No, the students _____ in the park.
3. Ahmad and Faisal are students. _____ from Jeddah.

6

5. Circle the correct option. ضع دائرة حول الكلمة الصحيحة.

1. My children (was / are / were / can) at home yesterday.
2. Mr. Smith (is / can / could / was) ride a horse when he was young.
3. I visited India when I (was / are / were / can) twenty.

3

6. Circle the correct option. ضع دائرة حول الكلمة الصحيحة.

1. My friends never (go / goes) swimming on weekdays.
2. My father is a teacher. He (teach / teaches) German.
3. Peter (live / lives) in a small house in London.
4. George and Linda are from Canada. They (speak / speaks) French and English.
5. My uncle is a pilot. He (fly / flies) all over the world.

5

7. Answer the questions in complete sentences. اجب على كل سؤال من الاسئلة التالية في جملة كاملة. الاجوبة القصيرة او المختصرة.

تحرك من بعض الدرجات

1. Where were you last Friday?

2. Where does your friend come from?

6

3. What do your friends study?

8. Circle the correct option.

1. (Has / Do / Have / Does) your teacher got a watch?
2. (Has / Did / Do / Have) you ever been to London before?
3. My children (have / do / can / does) dinner at 8:00 every day.

3

9. Fill in the table with the correct verb forms.

Infinitive	Past Simple	Past participle
write	_____	written
meet	met	_____
do	did	_____
make	_____	made
eat	_____	eaten
take	_____	taken

6

10. Complete the sentences with the comparative or superlative form of the adjective in brackets.

1. English is much _____ (easy) than Chinese.
2. Football isn't the _____ (popular) sport in my country.
3. My pen is _____ (expensive) than Ann's.
4. What is the _____ (big) city in Saudi Arabia?
5. This is the _____ (bad) day of my life.

5

11. Each sentence contains one mistake. Find the mistake, mark it and write the word correctly in the blank.

1. Everybody likes the shop assistant because he is a honest person. _____
2. My sister and I am cooking dinner for our family tonight. _____
3. The manager was'nt in his office when I called him. _____
4. Everyone is haveing a good time at the party. _____
5. Henry didn't knew how to cook until he moved out of his parents' house. _____
6. Jane has two cats. She likes their very much. _____

6

12. These sentences are in the active form. Write them in the passive form.

1. They make Toyota cars in Japan.



2. Alexander Fleming discovered penicillin in 1928.

13. Complete the sentences with relevant information.

1. Although we were stuck in a traffic jam, _____



2. I met the man who _____

3. I have an appointment with the dentist, so _____

4. My father was angry because _____

14. Write a paragraph that tells if you are with or against the idea of using mobile phones in schools.



Appendix G: University's Permission Letter

University of Hail
Hail
Saudi Arabia

Hail, 10th February, 2010

To whom it may concern,

The request has reached us to take part in a research project of Mr. S.A.A. Korevaar in partial fulfilment for the degree of PhD in Applied Linguistics.


The theme English as a Second Language has been indicated by the aforementioned.

We would be happy to facilitate this and will discuss further particulars about the concrete execution in due course.

I trust to have informed you sufficiently.

Kind regards,

Dr. Eid Al-Haisoni



Dean ELC
University of Hail
Saudi Arabia

Appendix H: Informed Consent English & Arabic

Hello,

I am currently researching Issues in Language Learning with a focus on Technology as part of my PhD studies in Applied Linguistics at the University of Bedfordshire in Luton, UK.

انا حاليا البحث في قضايا تعلم اللغات مع التركيز على التكنولوجيا كجزء من دراسات الدكتوراه في اللغويات التطبيقية بجامعة بيدفوردشير في لوتون، المملكة المتحدة.

In order for this research project to be successful, I would greatly appreciate it if you would be willing to participate in this study.

من أجل هذا المشروع البحثي لتكون ناجحة ، وسأكون ممتنا لو تفضلتم تكون على استعداد للمشاركة في هذه الدراسة.

I truly understand that you, as a professional, are extremely busy, and therefore the whole procedure will not take longer than 1 hour.

أنا أفهم أنك حقا، كمحترف، مشغولون للغاية ، وبالتالي فإن الإجراء بأكمله لن تأخذ وقتا أطول من 1 ساعة.

سياستنا يؤكد تقدير المشاركين فيه ما يلي؛

- Any information given by the participant will be processed as confidential
ستتم معالجة أية معلومات التي قدمها المشاركين على أنها سرية
- Any data received will be administrated in a securely manner
سوف يتم إدارتها أية بيانات وردت بطريقة آمنة
- Identity of participants will not be disclosed to supervisors, colleagues etc.
لن هوية المشاركين يتم الكشف عنها للمشرفين والزملاء الخ.
- All info received will serve for academic purposes only
سوف يخدم جميع المعلومات الواردة لأغراض علمية فقط

If you have any questions on the procedure, results, or if you have any other questions related to this study, please do not hesitate to contact me at:

إذا كان لديك أي أسئلة حول هذا الإجراء ، النتائج ، أو إذا كان لديك أي أسئلة أخرى ذات صلة بهذه الدراسة ، لا تترددوا في الاتصال بي على العنوان التالي :

0700500@beds.ac.uk

Participation in this study is not in any way compulsory. If you change your mind as you go along and therefore want to cancel your participation, you are free to do so at any time. Your data will then automatically be destroyed.

المشاركة في هذه الدراسة ليست بأي حال من الأحوال إلزاميا. إذا كنت غيرت رأيك كما تذهب على طول، وبالتالي تريد إلغاء مشاركتكم، وأنت حر في أن يفعل ذلك في أي وقت. سوف البيانات تلقائيا ثم يتم تدميرها.

If you are willing to participate in this research, please fill out the date below and tick off the box as an agreement to having read and understood the abovementioned:

إذا كنت على استعداد للمشاركة في هذا البحث ، يرجى ملء تاريخ أدناه والقراد قبالة مربع كما هو وجود اتفاق على قراءة وفهم ما تقدم :

Date

"I declare having read and understood the abovementioned and I am happy to participate in this research project"

"تعلن وجود قرأت وفهمت المشار إليها أعلاه ويسعدني أن أشارك في هذا المشروع البحثي"

Again, your co-operation is highly appreciated and of essential importance to this research, therefore I would be happy to send you an analysis of the outcome.

مرة أخرى ، هو في غاية الامتنان لتعاونكم والأهمية الأساسية لهذا البحث ، لذا سأكون سعيدا لنرسل لك تحليلا للنتائج.

Thank You

Appendix I: Samples of Textbooks Used in Target Context

11 Writing a biography

1 Complete the biography of Cher with *who*, *which*, or *where*.

Cher was born in the US on 20 May 1946 in El Centro, (1) _____ is on the California/ Mexico border. Her full name is Cherilyn Sarkisian and she is part-Cherokee and part-Armenian, Turkish, and French. She left high



school when she was 16 and went to Los Angeles, (2) _____ she planned to take acting lessons. There she met Salvatore Bono, (3) _____ was working at the Gold Star Studios (4) _____ Phil

Spector was recording many famous singers. He discovered that Cher could sing, and they became the singing duo Sonny and Cher. Their first hit song was 'I got you Babe', (5) _____ topped the charts in 1965. Cher was still only 19. They got married and had a daughter, (6) _____ they called Chastity. In 1975 Sonny and Cher were divorced, and later that



year Cher married Greg Allman, (7) _____ was another famous rock star. They had a son called Elijah Blue. But two years later Cher was divorced for the second time because of Allman's drink and drugs problems.

She decided to turn to acting again. In 1982 she appeared in her first major film, 'Come Back to the Five and Dime, Jimmy Dean, Jimmy Dean', (8) _____ was well received by the critics and



public. She went on to win Best Actress at the Cannes Film Festival in 1985 for her role in 'Mask', and finally she won an Oscar for 'Moonstruck' in 1987. However, in the 1990s she returned to pop music in a big way.

She has had three number one hits from her chart-topping album 'Believe', (9) _____ has reached a whole new audience. In her long career, Cher has been extremely successful both as a serious actress and as a pop star, (10) _____ is an extraordinary achievement.

2 Divide the text into five paragraphs according to these headings:

- introduction
- early career
- private life
- later career
- life now

3 Write a similar biography of somebody who you think is interesting.

6 Reported to direct speech

T 14.4 Read the report of an interview with Laurence Wilmot. Then write the actual words of the interview.

INTERVIEW WITH

Laurence Wilmot

actor and musician



I asked Laurence how he felt about winning the Best Television Actor award. He told me that he had been very pleased and surprised. He said that he had not expected to win, and he also wanted to thank all the other actors in the programme. I asked him what it had been like to play the part of Sherlock Holmes, and he said that it had been great fun.

I asked him if he had ever played a Shakespearian role, and he told me that he had. He'd played Othello off Broadway last year, and he'd enjoyed it very much.

I asked Laurence what sort of music he liked, and he told me that he had always liked jazz. In fact, he said he played in a jazz band called Saxophony. When I asked him where the band played, he told me they mainly played in small clubs.

Finally, I asked him if he ever wanted to direct a play, and he told me that he hoped to one day, but he didn't know when it could happen because he was so busy acting and playing jazz.

Interviewer How do you feel about winning the award, Laurence?

Laurence I'm (1) _____ . I didn't expect (2) _____ , and I (3) _____ all the other actors.

Interviewer What (4) _____ Sherlock Holmes?

Laurence It (5) _____ great fun.

Interviewer (6) _____ a Shakespearian role?

Laurence Yes, (7) _____ Othello off Broadway last year. (8) _____ very much.

Interviewer What sort (9) _____ , Laurence?

Laurence I have always liked (10) _____ . In fact, (11) _____ called Saxophony.

Interviewer (12) _____ direct a play?

Laurence (13) _____ one day, but (14) _____

Appendix J: Kobrin's (2000) List of Strategies

Table 18

Overall Test-Taking Strategies for Computerized and Paper-and-Pencil Tests

Overall Test-Taking Strategy	First Test		Second Test	
	CT	PP	CT	PP
Read passage first, then answered questions	9 (75%)	8 (67%)	7 (58%)	8 (67%)
Started to read passage, then skipped to questions before finishing reading	1 (8%)	1 (8%)	0	0
Read all questions first, then read entire passage, then answered questions	0	1 (8%)	0	1 (8%)
Read and answered one question at a time	1 (8%)	0	0	1 (8%)
Read one or more questions, then read passage, then answered questions	1 (8%)	2 (17%)	5 (42%)	1 (8%)
Other	0	0	0	1 (8%)

Appendix K: Al-Amri's (2008) Taxonomy

1- Affective (AFF)

- 1- Starting with the name of God.
Example: (in the name of Allah بسم الله الرحمن الرحيم)
- 2- God willing.
Example: (God willing ان شاء الله)
- 3- Self motivation.
Example: (Think Abdulkareem think فكر يا عبدالكريم فكر)
- 4- Thanking God.
Example: (Thank Allah الحمد لله)

2- Management (M) of the test as a whole

2.1 (Overall test) management (OTM) of reading and answering questions

- 1- Started reading the whole passage then answering questions in order.
Example: (Let's read the whole passage first then go to the questions خل نقرأ القطعة أول شي بعدين نروح للأسئلة)
- 2- Started reading paragraph by paragraph and answering questions relevant to each paragraph.
Example: (let's read paragraph by paragraph and answer their questions خل نقرأ مقطع مقطع ونجاوب أسئلة كل مقطع (لوحده)
- 3- Started answering questions in order and returning to relevant paragraphs without reading the whole passage first.
Example: (let's read the questions first and return to relevant paragraphs خل نقرأ الأسئلة بعدين نرجع للمقطع الخاص بكل (سؤال)
- 4- Switching between any of the above.
Example: (let's read the whole passage then the questions....after some reading the student changed his mind and said "it's better to read the questions and answer them in order" خل نقرأ القطعة وبعدين الأسئلة...بعد فترة من القراءة "أحسن خل نحل الأسئلة بالترتيب)

2.2 Time management strategies (TIM)

- 1- Making reference about time allocation.
Example: (How much time left? Oh only 15 minutes left (كم بقي من الوقت، الله ما بقي الا خمسة عشر دقيقة)
- 2- Budgeting the time on the test.
Example (I will spend 10 minutes on this passage (أخلي عشر دقائق لهذي القطعة)
- 3- Adjusting pace.
Example: (I shall be quick (يبغى لنا نسرع شوي)

2.3 Task management strategies (TAM)

2.3.1 Before task

- 1- Activating prior knowledge.
Example: (Have I taken it before هل أخذناها من قبل)
- 2- Checking the total number of questions in the test.
Example: (how many questions for passage one (كم سؤال على القطعة الأولى)

2.3.2 During task

- 3- Monitoring the sequence of letters of right answers.
Example: (OK, ABDA this is the sequence (هذا هو الترتيب ABDAايوه)

- 4- Searching for questions relevant to the paragraph just read or to be read.
Example: (where is the question about paragraph one **الوین السؤال الخاص بالمقطع الأول**)
- 5- Monitoring how much of the task is done (passage read / questions answered).
Example: (OK, now I finished paragraph one and two and their questions **زين الحين خلصنا المقطع الاول والثاني مع اسئلتهم**)
- 6- Counting paragraphs to find the relevant one.
Example: (one two three four, OK this is paragraph five **الواحد اثنين ثلاثه اربعة، ايوه هذا هو المقطع الخامس**)

2.3.3 After task

- 7- Searching for unanswered or doubtful items.
Example: (let's see where the unanswered question is **خل نشوف وين السؤال اللي ما حليناه**)
- 8- Checking that all questions have been answered.
Example: (let's make sure that I have answered all the questions **خل نتأكد أن جاوبنا على كل الأسئلة**)

3- (Re) Reading (R) related to individual questions

3.1 (Re) Reading text (RT)

- 1- Locating the area in the text that occurred in the question or option and then looking for clues to the answer in that context.
Example: (where is this? OK, here it is, let's read **وين هذي خل نشوف هذي هي خل نقرأ ونشوف**)
- 2- Looking back at the passage to search for specific words or phrases which occurred in a question.
Example: (let's go back to the passage and look for 'pigment' **خل نرجع للمقطعة ونبحث عن كلمة pigment**)
- 3- Reading the paragraph first before reading the question.
Example: (I will read the paragraph first **أول خل نقرأ المقطع**)
- 4- Reading the text or part of it to find an answer after reading a question.
Example: (now let's read the paragraph **الحين نروح نقرأ المقطع**)
- 5- Reading part of the text that contains the keyword in a question or option.
Example: (a waste gas of cellular metabolism)
- 6- Rereading part of the text that contains the keyword in a question or option.
Example: (a waste gas of cellular metabolism ... a waste gas of cellular metabolism)
- 7- Rereading the text or part of it not containing a keyword to find an answer after reading a question.
Example: (let's go back and reread... the respiratory system consists of... **خل نرجع ونقرأ المقطع مره ثانية ... the respiratory system consists of**)
- 8- Rereading the text or part of it for clarification.
Example: (air enters the body through the nose... air enters the body through the nose)
- 9- Rereading the target word, i.e. a keyword in a question which appears in the text.
Example: (metabolism metabolism metabolism)
- 10- Rereading part of the text to confirm selection.
Example: (red-coloured pigment haemoglobin, OK red-coloured pigment haemoglobin **تمام**)
- 11- Highlighting or underlining important information in the text related to a specific question.

Example: (let's underline or highlight this *خل نخط خط تحت هذي والا نطلله*)

- 12- Translating the main idea of the text or part of the text related to a specific question into L1 (i.e. Arabic).

Example: (this means air enters from the nose to the nasal cavity then to the pharynx then to the trachea *هذا معناه ان الهواء يدخل من الأنف الى التجويف الأنفي ثم الى القصبة الهوائية*)

- 13- Reasoning / justifying while reading the text related to a specific question.

Example: (sure this happens if the difference in partial pressure increases the rate of diffusion will increase *أكد اذا زاد الاختلاف في الضغط يزيد معدل الانتشار*)

3.2 (Re) Reading a question (RQ)

- 1- Reading the question or part of it after reading the passage or the paragraph.

Example: (let's read the question now *خل نقرأ السؤال الحين*)

- 2- Reading the question or part of it before reading the passage or the paragraph.

Example: (the word this in paragraph 9 refers to, where is paragraph nine)

- 3- Rereads the question or part of it for clarification.

Example: (the word diaphragm means diaphragm means)

- 4- Translating the question or part of it or what is required by the question into L1 (i.e. Arabic).

Example: (which paragraph tells us more about exhaling and inhaling? *يقول السؤال اي مقطع يخبرنا عن الشهيق والزفير*)

3.3 (Re) Reading options (RO)

- 1- Reading an option or part of it after reading the question.

Example: (now read the options *الحين نقرأ الخيارات*)

- 2- Rereading an option or part of it to confirm a selection.

Example: (bees live, OK, where bees live)

- 3- Rereading an option or part of it for clarification.

Example: (bronchi bronchioles alveoli bronchi bronchioles alveoli).

- 4- Continue reading options even when reaching the option thought to be correct.

Example: (this is the right answer but let's read the remaining options *هذي صح بس خل نقرأ الخيارات الباقية*)

- 5- Translating an option or part of it into L1 (i.e. Arabic).

Example: (all chemical processes in a living organism *العمليات الكيميائية في الكائن الحي*)

- 6- Reading an option or part of an option before reading a question.

Example: (first let's read the options *أول خل نشوف الخيارات*)

3.4 Reading test instructions (RI)

- 1- Reading the test instructions before starting the test.

Example: (read the following passage and answer the questions that follow)

- 2- Translating the test instructions into L1 (i.e. Arabic).

Example: (يقول اقرأ القطعة التالية وجاوب على الأسئلة التي بعدها)

4- Selecting or attempting to select an option or part of an option (S)

- 1- Selecting an option by matching keyword/words in it that also appeared in the text.

Example: (between the bronchioles and alveoli, this one it's in passage هذي هي موجودة في القطعة)

- 2- Selecting an option because it is stated in the text.
Example: (lie down in front of the older, this is stated exactly in the passage هذي مذكورة بالنص في القطعة)
 - 3- Selecting an option based on understanding the material read though it is not stated in the text.
Example: (in the passage it says..., OK it's A (حسب القطعة الإجابة هي
 - 4- Selecting an option relying on background knowledge.
Example: (According to my knowledge the answer is this (حسب معلوماتي الإجابة هذي
 - 5- Selecting an option using a rule of thumb.
Example: (this is the answer and I won't change it the first choice is always right هذي هي الإجابة ولا أغيرها دائما (الجواب الأول صح
 - 6- Returning to the text to look for the correct answer, after reading the question and all or some of the options.
Example: (let's go back to the text to find the answer (الحين نرجع للنص نشوف الإجابة
 - 7- Returning to the text to look for the correct answer after reading the question or part of it but before reading the options.
Example: (before I see the options let's reread the paragraph (قبل نشوف الخيارات خل نرجع ونقرا النص مره ثانية
 - 8- Returning to the text to confirm an answer.
Example: (let's confirm it (خل نتأكد
 - 9- Selecting an option based on inference from the text read.
Example: (because it starts with nose so A has the right order (بما أنها بدت بالأنف اذا الإجابة الأولى هي الترتيب الصحيح
 - 10- Reasoning/justifying selecting an option.
Example: (sure C is right because paragraph 3 5 and 9 did not talk about inhaling and exhaling (الثالثة لأن مقطع ثلاثة وخمسة وتسعة ما تكلموا عن الشهيق والزفير
 - 11- Selecting an option using blind guessing.
Example: (I do not know the answer but I will choose D (والله ما عرف الإجابة بس حظ الرابعة ومشني حالك
 - 12- Selecting an option by recalling information from the text.
Example: (it says in the text that it diffuses from the cells into the capillaries so the first option is right (في القطعة أنه ينتشر من الخلايا الى الأوعية الدموية اذا الإجابة الأولى صح
 - 13- Attempting to select an option by recalling information from the text.
Example: (it could be C based on the passage but (ممکن تكون الإجابة الثالثة حسب القطعة ولكن
 - 14- Returning to the text to look for a word and find an answer with the help of the keyboard shortcuts (i.e. Ctrl+F).
Example: (let's use the keyboard to find it (خل نستخدم لوحة المفاتيح ونبحث عنها
- 5- Rejecting or attempting to reject an option (RE)**
- 1- Discarding an option(s) based on background knowledge.
Example: (this one is wrong it's known (هذي خطأ معروفة
 - 2- Discarding option(s) based on sentence, paragraph, or passage overall meaning.
Example: (according to the passage this is wrong (حسب النص هذي خطأ
 - 3- Rejecting an option for unstated reason.

Appendix L: Strategy Counting Template

Strategies Coding Template

Overall Test-Level Strategies	
O1	
O2	
O3	
O4	
O5	
O6	

Initial Reading of Passage (when applicable)	
R1	
R2	
R3	
R4	
R5	
R6	
R7	
R8	
R9	

Test Taking Strategies										
	1	2	3	4	5	6	7	8	9	10
T1										
T2										
T3										
T4										
T5										
T6										
T7										
T8										
T9										
T10										
T11										
T12										
T13										
T14										
T15										
T16										
T17										
T18										
T19										
T20										
T21										
T22										
T23										
T24										

	1	2	3	4	5	6	7	8	9	10
T25										
T26										
T27										
T28										
T29										
T30										
T31										
TW1										
TW2										
SUP1										
SUP2										
SUP3										
Ex1										
Ex2										
Ev1										
Ev2										
Ev3										
Inf1										
Inf2										
Inf3										
Aff										

P C

Strategies Coding Template

Overall Test-Level Strategies	
O1	1
O2	
O3	
O4	
O5	
O6	

Initial Reading of Passage (when applicable)	
R1	
R2	
R3	1
R4	
R5	1
R6	
R7	
R8	
R9	1

Test Taking Strategies										
	1	2	3	4	5	6	7	8	9	10
T1	1	1	1	1	1	1	1	1		
T2	2				2		2	1		
T3										
T4	3				3					
T5					4					
T6					4			2		
T7										
T8	4	3	12	2	3	2	3			
T9	5			3	3	3/5	4/10	3		
T10						7		6		
T11										
T12										
T13	6			4						
T14										
T15	11			6	11		11	7		
T16			13	3		3				
T17		8		3						
T18										
T19										
T20										
T21				7		10		8		
T22		10								
T23										
T24										

Appendix M: Think Aloud Protocol PBT

790 PBT

(In the name of God, the Beneficent, the Merciful)

When did Newman first work in the? (.) What is the name of the? (.) Car racing start. Just taking a quick view at the questions. Paul Newman was born in Cleveland, Ohio, in 1925, and did some acting in high school and college, but never seriously considered making it his future career. However, after graduating, he started working in the theatre and on several TV shows in New York. When he was thirty, he went to Los Angeles and made his first film. It was what he called an 'uncomfortable' start in the movie, in the role of a Greek slave. The next film he chose was his big break. He played the role of the boxer, Rocky in the film 'Someone up There Likes Me'. Newman is a method actor who believe in the role before beginning the film. He spent days – from morning till night – with Graziano. No (.) (difficult name) (.) He studied the boxer's speech and watched him box. The picture brought Newman stardom overnight. Newman went on to make films such as 'Cat on a Hot Tin Roof', 'Butch Cassidy and the Sundance Kid', 'The Sting' and 'Towering Inferno'. He was made over forty-five films and has won many awards, but he has never won an Oscar. He was living in Los Angeles when he became engaged to Joanne Woodward, an actress whom he had first known in New York. Newman and Miss Woodward were married in Las Vegas (Newman and Miss Woodward were married in Las Vegas). His marriage to Woodward is one of the longest and strangest in Hollywood. They have co-starred in six films. Ever since the film 'Winning', Newman has been interested in car racing, and in 19 (nineteen) he came second in the twenty-four hour Le Mans race. He has a strong social conscience, and has supported causes such as driver. All the money from 'Newman's (.) When did Newman first work in the theatre? Student begins to speak Arabic:(In the beginning it's asking about his work in the theatre. He said: I didn't find it so I will read it again) <rereads part of passage>(.) Oh, when. (.) When did he start? (.) This was in the thirties. (.) I have to put in capital letters. I have to put A.D because this is in the past. < writes answer> (.) He started (.) (I forgot the date. I believe, but there is no date here so I want to put when he started he was thirty years old) It's ok written.(.) What's the name of his company? I remember I read it.(.) Be patient. (.) He went back to Los Angeles. Then the next film he.. (I want to ask the teacher but I'm confused) What's the name of Newman's company? He played(.)He

studies (.) (What was his beginning, what did he do next?) Then he went to Los Angeles (.) I can't find it (.) I don't think that is it (.) (The next paragraph is about his films. I'm going to read it and pay attention) When did Newman's interest in car racing start? Let me read, maybe I will find something. (.) He got married. They joined together in 6 films (.) Car racing (.) I think this is the answer (.) What did interest in the car? When, oh, when. (.) This question is easy and clear. 1979. Very easy. (.) <writes answer> I will write in briefly (.) In 1979 (.) How many films did Newman and Woodward make together? This is about the films they did together (.) It is easy I don't have to return to the text because I have memorized it (.) I have to look for the word films because I don't remember it. I'm very bad in spelling. (.) Where did Newman first know Woodward from? (.) This was in New York. (.) <writes answer> This goes with a capital because it is a city. The n is capital and the y is capital. (.) Then I have to put periods which I forgot to put in all my answers (.) What is a method actor? I'm going to go back to the question to the end of it. A method actor I found in the second paragraph (.) He believes in (.) what is it? (.) I will start from the beginning so that it is clear so that the person who is reading understands that I understand <writes answer> (.) I wrote it with a bigger font. My spelling is not good. I'm looking for believe. I E V E. Living. I know this word. E I (.) no it's correct. Living in the (.) I will write it at the bottom (.) Before beginning the film (.) I didn't memorize it (.) I double n, I N G the film. This is easy. Full stop (.) Which film made Newman a star? This is when it mentioned about him having a break (.) He played the boxer 'Someone up There Likes Me'. The break was in the beginning of the second paragraph (.) The second line (.) This is a film <writes answer> There likes me. L I A S M E (.) How many questions left? I think 3 (.) When did Newman make his first film? There is a mistake here (.) I will write the same answer <writes answer> When he was thirty (.) In his thirties or thirty? <checks passage> No, he was thirty (.) The first question I wrote a wrong answer. I didn't know until I finished the 8th question (.) It's a long word <writes answer> D U A. D I N G. (.) The 9th question (.) "It was what he called an 'uncomfortable' start." What does "It" refer to in line 4? <counting lines> 1, 2, 3, 4. (.) When he was thirty he went to Los Angeles, it was what he called an uncomfortable start in the movie. (.) I will return to it later <counting lines again> 1, 2, 3, 4, 5, 6, 7, 8, 9. (What is its name? Ok) (.) I don't think it is American, it may be Mexican (.) What is the name of company? I think it's in the end (.) Newman's Own (.) This seems like company name <writes

answer> I will now go to last question (.) When he was thirty he went to (.) Till now it is not clear to me (.) I will read the sentence again (.) He went to Los Angeles and made his first film. (.) I don't think it's Los Angeles, no way (.) I have no idea. I haven't found anything (.) Let me read it from the beginning <reads passage> (.) I believe The role of a Greek slave (.) I think this is it. Yes < writes answer> (.) let me check the spelling, correct (.) Slave S L A V E (.) Let me check if I have filled in all the questions. (.) Done (.)

Appendix N: Think-Aloud Protocol CBT

110 CBT

Paul Newman was born in Cleveland, Ohio*(=Ohio), in 1925, and did some acting in high school and college, but never seriously considered making it his future career. However, after graduating, he started working in the theat*(=theatre) and on several TV shows in New York. When he was thirty, he went to Los Angeles and made his first film. It was what he called an 'uncomfortable' start in the movies, in the role of a Greek slave. The next film he choose*(=chose) was his big break. He played the role of the boxer, Rocky Graziano*(=Graziano) in the film 'Someone up There Likes Me'. Newman is a method actor who believes in the*(=N/A) living role before beginning the film. He spent days from morning till night with Graziano*(=Graziano). He studied the boxer's speech and watched him box. The picture brought Newman stardom overnight. Newman went on to make films such as 'Cat on a Hot Tin Roof', 'The Hustler', 'Butch Cassidy and the Sundance Kid', 'The Sting*(=Sting)' and 'Towering Inferno'. He has made*(=made), he has made over forty-five films and has own*(=won) many awards, but he has never won an Oscar. He was living in Los Angeles when he become*(=became) engaged to Johnny*(=Joanne) Woodward, an actress whom he had first known in New York. Newman and Miss Woodward were married in Las Vegas in 1958, 1985th*(=1958) His marriage to Woodward is one of the longest and strongest in Hollywood. They have co-started, starred in six films. Ever since the film 'Winning', Newman has been interested in car racing*(=racing), racing, and in 1979 he become*(=became) second in the twenty-four hour Le Mans race. He was a strong social conscience, and has supported causes such as the anti-nuclear*(=nuclear) movement, the environment, and driver education. All the money from 'Newman's Own' salad dressing, popcorn, and spaghetti sauce, now a multi-million-dollar business, goes to charity.

First question (.) When did Newman first work in the theatre? When did Newman first work in the theatre? (.) Come back here <scans passage> (.) After graduating he started working in the theatre. (.) After graduating <types answer> G - r - a - d, (.) g - r - a - d - u - a - t - i - n - g. (.) What's name of Newman's company? Wait, wait. I come back to it. It was <goes directly to location in passage> he went to Los Angeles and made his film. It was what he called an uncomfortable start. (.) No. Actually, I don't know the answer (.) Ok you go next one and you come back (.) When did Newman's interest in car racing? (.) Yes. Newman has been interest in racing and (.) wait. (.) They have co-starred in 6 films. Ever since the film winning (.) Since the film winning, Ok (.) How many films did Newman and

Woodward make together? (.) Yes. They did, they did, they did. Yes (.) Las Vegas. (.) They have co-started. (.) What's mean starred? (.) Wait, wait. He comes, no. (.) Above (.) wait, wait, wait (.) He has made over (.) No (.) Became engaged to actress (.) Newman and Ms. Woodward (.) No. (.) I think. No, no, no. I think the best answer for this 6 films <types answer> (.) Where did Newman first know Woodward from? In Los Angeles (.) I think the answer in Los Angeles <goes to location in passage and starts reading> First known in New York, Ok (.) What is a method actor? (.) I read this <goes directly to location in passage> Believe in living (.) Wait (.) believe in living the role before begin the film (.) Wait (.) in the film someone up there likes me. Newman is a method actor. What is a method actor? A method actor believes in living the role before beginning the film. (Wait) I believe, Wait (.) Yes, a method actor who believes in living the role before beginning the film. (I don't understand what is meaning of this, Ok <types answer> Believes in living the role before beginning. Ok (.) Which film made Newman star, a star? < reads passage> He spent from morning until night. (.) No, no, no. {UV} He has made over 45 films but never won Oscar. (.) Wait, wait. I'm reading from beginning (.) The next film he chose was his big break. He played role of boxer. Ok, he spent days from morning until night. (No) Studied the boxer's speech, No (.) The picture brought Newman stardom overnight, no, no, no, no (.) He made over 45 films, No (.) I will move to question {UV} (.) When did Newman make his first film? (When oh you grateful person!) He went to Los Angeles and made his first film. (The best answer for this when he was 30 he went to Los Angeles, when he was 30) When he went to Los Angeles. When he was 30 he went to Los Angeles and made his first film. I think the answer when he was 30 (.) "It was what he called an 'uncomfortable' start." What does "It" refer to in line 4? It was what he called an uncomfortable start in the role. Ok (.) when he was 30 he went to Los Angeles and made his first film. What it, it, it refer to? Ok (.) It refer to when he went to Los Angeles made his film (.) Ok, making his film. It refer to making his film <types answer> His first film, Ok (.) "He studied the boxer's speech and watched him box." What does "him" refer to in line 9? (Refer to boxer) <counts lines> 2, 9, 9, this is 4, 5, 6, 7. Wait (.) Yes, studied the boxer speech and watch him box. (.) Him refers to the boxer (.) He spend days from morning till (.) Graziano. I think this is that name of boxer (.) Ok Rocky Graziano. Refer to Rocky Graziano <types answer> Z - I - A - N - O. (.) Ok my friend. Let me check (.) Done (.)

Appendix O: Cohen & Upton's (2007) Strategies

Table 7

Revised Reading Strategies Coding Rubric (R)

Strategy	Description
Approaches to reading the passage	
R1	Plans a goal for the passage.
R2	Makes a mental note of what is learned from the prereading.
R3	Considers prior knowledge of the topic.
R4	Reads the <u>whole</u> passage <u>carefully</u> .
R5	Reads the <u>whole</u> passage <u>rapidly</u> .
R6	Reads a <u>portion</u> of the passage <u>carefully</u> .
R7	Reads a <u>portion</u> of the passage <u>rapidly</u> looking for specific information .
R8	Looks for markers of meaning in the passage (e.g., definitions, examples, indicators of key ideas, guides to paragraph development).
R9	Repeats, paraphrases, or translates words, phrases, or sentences—or summarizes paragraphs/passage—to aid or improve understanding.
R10	Identifies an unknown word or phrase.
R11	Identifies unknown sentence meaning.
Uses of the passage and the main ideas to help in understanding	
R12	During reading rereads to clarify the idea.
R13	During reading asks self about the overall meaning of the passage/portion.
R14	During reading monitors understanding of the passage/portion's discourse structure (e.g., compare/contrast, description, definition).
R15	Adjusts comprehension of the passage as more is read: Asks if previous understanding is still accurate given new information.
R16	Adjusts comprehension of the passage as more is read: Identifies the specific new information that does or does not support previous understanding.
R17	Confirms final understanding of the passage based on the content and/or the discourse structure.

(Table continues)

Table 7 (continued)

Strategy	Description
Identification of important information and the discourse structure of the passage	
R18	Uses terms already known in building an understanding of new terms.
R19	Identifies and learns the key words of the passage.
R20	Looks for sentences that convey the main ideas.
R21	Uses knowledge of the passage/portion: Notes the discourse structure of the passage /portion (cause/effect, compare/contrast, etc.).
R22	Uses knowledge of the passage/portion: Notes the different parts of the passage (introduction, examples, transitions, etc.) and how they interrelate (“Is this still part of the introduction or is this the first topic?” “This sounds like a summary—is it the conclusion?”).
R23	Uses knowledge of the passage/portion: Uses logical connectors to clarify content and passage organization (e.g., “First of all,” “On the other hand,” “In conclusion”).
R24	Uses other parts of the passage to help in understanding a given portion: Reads ahead to look for information that will help in understanding what has already been read.
R25	Uses other parts of the passage to help in understanding a given portion: Goes back in the passage to review/understand information that may be important to the remaining passage.
Inferences	
R26	Verifies the referent of a pronoun.
R27	Infers the meanings of new words by using work attack skills: Internal (root words, prefixes, etc.).
R28	Infers the meanings of new words by using work attack skills: External context (neighboring words/sentences/overall passage).

Table 8***Test-Management Strategies Coding Rubric (T)***

Strategy	Description
T1	Goes back to the question for clarification: Rereads the question.
T2	Goes back to the question for clarification: Paraphrases (or confirms) the question or task.
T3	Goes back to the question for clarification: Wrestles with the question intent.
T4	Reads the question and considers the options before going back to the passage/portion.
T5	Reads the question and then reads the passage/portion to look for clues to the answer, either before or while considering options.
T6	Predicts or produces own answer after reading the portion of the text referred to by the question.
T7	Predicts or produces own answer after reading the question and then looks at the options (before returning to text).
T8	Predicts or produces own answer after reading questions that require text insertion (I-it types).
T9	Considers the options and identifies an option with an unknown vocabulary.
T10	Considers the options and checks the vocabulary option in context.
T11	Considers the options and focuses on a familiar option.
T12	Considers the options and selects preliminary option(s) (lack of certainty indicated).
T13	Considers the options and defines the vocabulary option.
T14	Considers the options and paraphrases the meaning.
T15	Considers the options and drags and considers the new sentence in context (I-it).
T16	Considers the options and postpones consideration of the option.
T17	Considers the options and wrestles with the option meaning.

(Table continues)

Table 8 (continued)

Strategy	Description
T18	Makes an educated guess (e.g., using background knowledge or extra-textual knowledge).
T19	Reconsiders or double-checks the response.
T20	Looks at the vocabulary item and locates the item in context.
T21	Selects options through background knowledge.
T22	Selects options through vocabulary, sentence, paragraph, or passage <u>overall meaning</u> (depending on item type).
T23	Selects options through elimination of other option(s) as unreasonable based on background knowledge.
T24	Selects options through elimination of other option(s) as unreasonable based on paragraph/overall passage meaning.
T25	Selects options through elimination of other option(s) as similar or overlapping and not as comprehensive.
T26	Selects options through their discourse structure.
T27	Discards option(s) based on background knowledge.
T28	Discards option(s) based on vocabulary, sentence, paragraph, or passage <u>overall meaning</u> as well as <u>discourse structure</u> .

Table 9

Test-Wiseness Strategies Coding Rubric (TW)

Strategy	Description
TW1	Uses the process of elimination (i.e., selecting an option even though it is not understood, out of a vague sense that the other options couldn't be correct).
TW2	Uses clues in other items to answer an item under consideration.
TW3	Selects the option because it appears to have a word or phrase from the passage in it—possibly a key word.

Appendix P: This Study's Identified Strategies (Template)

Strategies Coding Template

Overall Test-Level Strategies	
O1	Reads passage first then answers questions
O2	Starts to read passage then skips to questions before finishing reading
O3	Reads all questions first, then reads passage then answers questions
O4	Reads and answers one question at a time

Initial Reading of Passage (when applicable)	
R1	Reads whole passage carefully
R2	Reads whole passage rapidly
R3	Reads a portion of the passage carefully
R4	Reads a portion of the passage rapidly looking for specific information
R5	Pauses & thinks about reading
R6	Repeats word(s)/phrase(s)/sentence(s) in reading passage to aid in comprehension
R7	Paraphrases/Summarizes portion(s) of reading passage to aid in comprehension
R8	Translates word(s)/phrase(s)/sentence(s) to aid in comprehension
R9	Indicates that he doesn't understand word/phrase meaning in passage

Test Taking Strategies	
T1	Reads Question
T2	Rereads word/phrase in question stem for clarification
T3	Translates word/phrase in question stem for comprehension
T4	Paraphrases question stem for clarification
T5	Guesses meaning of unknown word(s) in question
T6	Reads question stem and then scans passage for keyword(s)
T7	Reads question stem and then search reads the passage/portion to look for clues to the answer
T8	Uses spatial memory to locate key words
T9	When found keyword(s)/clue(s), reads sentence containing clue(s) keyword(s) carefully
T10	Rereads sentence containing clue(s)/keyword(s) for clarification
T11	Paraphrases sentence containing clue/keyword(s) for clarification
T12	Translates word(s)/phrase(s) in sentence containing clue(s)/ keyword(s) for clarification
T13	Reads sentence before/after sentence containing key information for contextual clarification
T14	Guesses meaning of unknown words in passage
T15	Utilized strategies generate correct answer
T16	Utilized strategies generate incorrect answer
T17	Uses background knowledge to aid in answering question
T18	Provides correct answer to question from memory
T19	Provides wrong answer to question from memory
T20	Guesses answer
T21	Reconsiders or double checks response
T22	Discovers answer to item later on and goes back to change previous answer
T23	Changes correct answer into incorrect answer after rereading
T24	Changes incorrect answer into correct answer after rereading
T25	Moves to next question without answering item
T26	Goes back to question for clarification

