THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

A validation framework for an online English language Exit Test: A case study using Moodle as an assessment management system

Zakiya Salim Hamed Al Nadabi

MA in Applied Language Studies

MA in Language Testing

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2017

School of Education

Abstract

Technology-enhanced language tests are increasingly being hosted on course management systems (CMSs) like Moodle. Despite the increased use of CMS-hosted tests and the rising concerns over the reliability and construct validity of computerised tests due to a potential testing mode effect (Chapelle & Douglas, 2006; Fulcher, 2003), validation research on these tests is lacking. Therefore, this study seeks to fill this gap with empirical validation research using a case study of administering and validating a CMS-hosted test. The test was a technology-enhanced English Language Proficiency Exit Test that was hosted on Moodle (hereafter called Moodle-hosted test) and administered to a group of EFL students (*N* = 207) at Sultan Qaboos University in Oman. The overall aim of the study was to provide a validity argument about using a Moodle-hosted test for its intended purpose by empirically establishing reliability and construct validity evidence. To achieve this aim, a study framework was successfully applied following principles of the Assessment Use Argument (AUA) framework of Bachman (2005) and Bachman and Palmer (2010). Applying the framework as a pragmatic tool to conduct validation research led to the structuring of an evidence-based argument about test reliability and construct validity drawing on multiple sources of evidence (Kane, 1992) collected via mixed-method design.

The results of Rasch analysis revealed that a quarter of the test items, which were of the gap-filling type requiring typing of responses, were overly difficult and had high unacceptable measurement error values. Although the study outcomes demonstrated warrants of statistically acceptable reliability estimates, two threats to reliability and construct validity were identified: construct-irrelevance and construct under-representation. The overly difficult items introduced construct-irrelevant difficulty as some test takers found the construct difficult and the resulting scores might have been invalidly low. Thirty percent of the test items also had unacceptable fit statistics, suggesting that they did not contribute independently to test reliability and they inconsistently assessed student performances. Having items with unacceptable fit statistics indicated departure from unidimensionality, as the test might have measured construct-irrelevant sub-dimensions other than the single dimension of language proficiency. Construct under-representation was identified by finding gaps between item difficulty and person ability measures, suggesting that the test did not capture examinees' ability levels well. As difficulty of the items did not match the ability levels of test takers, the test construct might have been under-represented by the set of items and better quality items might be needed to address a range of ability levels. With this evidence that the test had reliability and construct validity issues, the test scores might not be reliable and valid indicators of the target test construct. Further investigation examined a number of factors that could be potential sources of reliability and construct validity issues interfering with test performance results in the Moodle-hosted technology-enhanced testing mode.

Based on a comparison of test scores with examinees' post-test questionnaire responses, the study revealed that test performance was significantly affected by the testing mode due to construct-irrelevant technology-related factors. These were strong rebuttals to reliability and construct validity claims in the validity argument. The study found that some construct-irrelevant technology-related variables significantly affected test performance including: 1) the familiarity and levels of technology experience of test takers, familiarity with Moodle tests, and computer-literacy; 2) the functionality of headphones during the exam; 3) test taker's attitude towards the testing format; 4) the need to type responses for constructed-response test items; and 5) test time sufficiency and the use of a count-down timer. Other construct-irrelevant technology-related issues that did not significantly interfere with test performance were also considered as issues of concern, and these were: 1) screen layout and scrolling; 2) note-taking and text highlighting features; and 3) eye fatigue. Because negative evidence indicated that the testing mode effect threatened reliability and construct validity and created unfairness or bias issues, it was concluded in the validity argument that the Moodle-hosted score-based decisions cannot be justifiably reliable nor valid. The research questions were answered in the validity argument based on combined evidence from the study outputs, including test and post-test questionnaire responses. Therefore, a significant finding from this study was that statistical analysis of test responses alone is insufficient in developing computerised tests that are holistically fit for purpose.

This study contributes knowledge to the field as its findings lay out significant implications and recommendations about the testing mode effect. Practitioners and researchers may wish to adopt these implications and recommendations as guidelines for creating, developing, implementing, and researching reliable and valid large-scale high-stakes tests delivered on Moodle, other course management systems, or any other computerised test delivery tools. To ensure policy-makers are informed about whether using test outcomes can be justifiably fair to students, future validation research studies should be conducted so that potential issues with this testing mode can be further identified and addressed.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Refereed conference paper and presentation

Al Nadabi, Z. (2015). Features of an online English language testing interface. In T. Reiners, B.R. von Konsky, D. Gibson, V. Chang, L. Irving, & K. Clarke (Eds.), *Globally connected, digitally enabled. Proceedings ascilite 2015 in Perth* (pp. CP:17-CP:21). http://www.2015conference.ascilite.org/wp-content/uploads/2015/11/ascilite-2015-proceedings.pdf

Non-refereed conference papers and presentations

Al Nadabi, Z. (2014). *A validation framework for an online English language Exit Test: A case study using Moodle as an assessment management system.* Paper presentation given at ALTAANZ Conference 2014 (Assessing second languages: Linking theory, research, policy and practice), November 27 - 29, 2014, The University of Queensland, St Lucia, Australia. http://www.altaanz.org/uploads/5/9/0/8/5908292/2014_altaanz_conference_booklet.pdf

Al Nadabi, Z. (2014). *A construct validation study on a placement test: Implications for quality management in placement testing.* Paper presentation given at AARE-NZARE 2014 Conference, November 30 - December 4, 2014, Queensland University of Technology, Brisbane, Australia. https://thinkbusiness.conference-services.net/reports/template/onetextabstract.xml?xsl=template/onetextabstract.xsl&conferenceID=3787&abstractID=817332

Al Nadabi, Z. (2015). *Consequential validity and study paths: Students' voices.* Paper presentation at Oman 15th International ELT Conference, Sultan Qaboos University, Oman, April 23 – 24, 2015. http://www.eltconf.com/eltconf15/schedule/abstracts/pp196.html

Al Nadabi, Z. (2015). *A validation pilot study on web-based technology-enhanced assessment*. A paper presentation given at the School of Education Postgraduate Research Community Conference (Charting educational futures: Building on 20 years of research), August 15, 2015, The University of Queensland, St Lucia, Australia.

Al Nadabi, Z. (2015). *A validation framework for investigating the consistency and construct validity of web-based English language tests.* Paper presentation given at Australian Association for Research in Education (AARE) Conference, University of Notre Dame Australia, Fremantle, Australia, SIG: Assessment and Measurement December 3, 2015. https://www2.iceaustralia.com/ei/images/AARE%202015/AARE%202015_Full%20Program.pdf

Publications included in this thesis

“No publications included”.

Contributions by others to the thesis

“No contributions by others.”

Statement of parts of the thesis submitted to qualify for the award of another degree

“None”.

Acknowledgements

I would like to express my appreciation and thanks to my principal advisors Dr Mathew Hillier and Dr Noriko Iwashita and my associate advisor Dr Chris Campbell. I would like to thank you for all your support during my doctoral research. I appreciate your advice and encouragement that helped me improve academically. The priceless opportunities you created for me throughout my candidature led to my academic and professional growth. I am glad to have had the chance to learn from you during my PhD journey. I would also like to express my sincere thanks to Professor Robyn Gillies and Dr Michael Harrington for being my research higher degree committee members. I found your comments and suggestions tremendously useful, and enacting upon them was integral to the successful progress of my research.

I appreciate all the support that I received from my scholarship provider, the Ministry of Higher Education, and from my workplace, Sultan Qaboos University. I would like to thank the administration, teachers, technicians, and most importantly students in the Centre for Preparatory Studies. I would like to thank you all for your support with data collection for my thesis. Thanks to your input in the study, I am hoping that the study outcomes feed into the ongoing institutional development of computerised testing and aid in promoting a positive testing culture for all stakeholders.

Special words of gratitude go to my family. Words cannot express how grateful I am to my husband Sami for his patience and encouragement. We went through tough challenges together and I could not have survived them without your support. My little lovely girls, Farah, Hiba, Sahar and Aisha, you grew up with my studies and learned so much with me. I hope that when you get to read these lines independently one day, you will understand the rationale behind the dedication to my studies. I pray that you pave your way into a bright, happy and peaceful future. To my parents and siblings, I hope the quality of this thesis makes you proud and you forgive me for the precious family moments that I missed.

Keywords

testing mode effect, reliability, construct validity, construct-irrelevance, construct under-representation, course management system, Moodle, validity framework, validation

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 130303, Education Assessment and Evaluation, 50%
ANZSRC code: 130306, Educational Technology and Computing, 20%
ANZSRC code: 200303 English as a Second Language, 30%

Fields of Research (FoR) Classification

FoR code: 1303, SPECIALIST STUDIES IN EDUCATION, 70%
FoR code: 2003, LANGUAGE STUDIES, 30%

Table of Contents

List of Figures and Tables xi

List of Abbreviations xiii

CHAPTER 1. INTRODUCTION 1

1.1. INTRODUCTION 1

1.2. CONTEXT AND BACKGROUND 1

1.2.1. Technology skills accessibility in Oman. 3

1.2.2. Computerised assessment research in Oman. 4

1.3. PROBLEM STATEMENT 5

1.4. STUDY AIMS 7

1.5. THESIS STRUCTURE 8

CHAPTER 2. LITERATURE REVIEW 9

2.1. INTRODUCTION 9

2.2. VALIDATION FRAMEWORK 9

2.3. TEST VALIDATION RESEARCH 12

2.3.1. Earlier development of validation research. 12

2.3.2. Current validation research. 13

2.3.3. CALT validation research. 13

2.3.4. CALT validation aspects. 14

2.3.5. CALT validation studies. 15

2.4. EVALUATION OF CALTS 17

2.5. TESTING MODE EFFECT 18

2.5.1. Computer familiarity. 18

2.5.2. Keyboarding proficiency. 19

2.5.3. Equipment quality. 20

2.5.4. Attitude. 21

2.5.5. Timing. 22

2.5.6. Eye strain. 24

2.5.7. Other factors. 25

2.6. SUMMARY OF LITERATURE REVIEW 26

2.7. RESEARCH QUESTIONS 27

2.8. FORMULATING STUDY FRAMEWORK 28

2.9. CONCLUSION 29

CHAPTER 3. RESEARCH METHODOLOGY 30

3.1. INTRODUCTION 30

3.2. METHODOLOGICAL APPROACH 30

3.3. STUDY DESIGN 32

3.4. PARTICIPANTS 35

3.4.1. Students. 36

3.4.2. Language teachers. 37

3.5. DATA COLLECTION INSTRUMENTS 37

3.5.1. Online test tool. 38

3.5.2. Questionnaires. 42

3.5.3. Interviews. 43

3.5.4. Reflective journals. 43

3.6. DATA ANALYSES 44

3.6.1. RQ1: Statistical test item analyses. 44

3.6.2. RQ2: Descriptive statistics. 46

3.6.3. RQ2: Inferential statistics. 47

3.6.4. RQ2: Non-psychometric analyses. 49

3.7. CONCLUSION 50

CHAPTER 4. RESULTS AND DISCUSSION OF RESEARCH QUESTION 1 52

4.1. INTRODUCTION 52

4.2. RELIABILITY ESTIMATES 52

4.3. ITEM MEASURES 53

4.4. FIT STATISTICS 58

4.5. MEASUREMENT ERROR 60

4.6. SUMMARY OF RESULTS 63

4.7. CONCLUSION 63

CHAPTER 5. RESULTS FOR RESEARCH QUESTION 2 65

5.1. INTRODUCTION 65

5.2. THEME 1: ENDURANCE 66

5.3. THEME 2: EASE OF USE 68

5.4. THEME 3: EXPERIENCE WITH MOODLE TESTS 70

5.5. THEME 4: ATTITUDE AND RESISTANCE TO CHANGE 71

5.6. THEME 5: ENCOUNTERING TECHNICAL ISSUES 73

5.7. THEME 6: SOUND AND HEADPHONES QUALITY 74

5.8. THEME 7: SPLIT SCREEN FOR READING AND NOTE-TAKING 76

5.9. THEME 8: TEST MODE AND FEEDBACK 77

5.10. THEME 9: APPROPRIATENESS FOR TESTING PURPOSE 79

5.11. THEME 10: TIME MANAGEMENT 81

5.12. SUMMARY 83

5.13. CONCLUSION 86

CHAPTER 6. DISCUSSION 88

6.1. INTRODUCTION 88

6.2. RQ1: RELIABILITY AND CONSTRUCT VALIDITY 88

6.2.1. Threats to reliability and construct validity. 88

6.2.2. Implications of reliability and construct validity threats. 91

6.3. RQ2: CONSTRUCT-IRRELEVANT FACTORS 93

6.3.1. Familiarity and experience. 94

6.3.2. Headphones quality 96

6.3.3. Attitude towards testing format. 98

6.3.4. Typing responses for gap-filling items. 100

6.3.5. Test time and length. 103

6.4. DISCUSSION OF VALIDITY ARGUMENT 107

CHAPTER 7. CONCLUSION 112

7.1. INTRODUCTION 112

7.2. OVERALL SUMMARY OF THE FINDINGS 112

7.3. SIGNIFICANCE OF THE STUDY 113

7.4. IMPLICATIONS 114

7.4.1. Familiarity and experience. 115

7.4.2. Typing responses for gap-filling items. 115

7.4.3. Headphones quality 115

7.4.4. Test time and length. 116

7.4.5. Using new features. 117

7.4.6. Eye fatigue 118

7.4.7. Attitude towards testing format. 119

7.5. LIMITATIONS 121

7.5.1. Study design. 121

7.5.2. Study data. 121

7.5.3. Data analyses. 122

7.5.4. Language use in instruments. 122

7.5.5. Examined issues. 122

7.5.6. Test taker characteristics. 123

7.5.7. Documenting details. 123

7.6. CONCLUDING REMARKS 125

REFERENCES 127

APPENDICES 140

APPENDIX A: STUDY VALIDATION FRAMEWORK 140

APPENDIX B: ETHICAL CONSIDERATIONS AND RELEVANT FORMS 143

APPENDIX C: QUESTIONNAIRE FOR UQ STUDENTS IN THE FIRST PROTOTYPE TRIAL 162

APPENDIX D: QUESTIONNAIRE FOR JUDGMENTAL VALIDATION SESSION PARTICIPANTS 164

APPENDIX E: SEMI-STRUCTURED INTERVIEW FOR JUDGMENTAL VALIDATION PARTICIPANTS 166

APPENDIX F: USABILITY STUDY TEST TAKERS' QUESTIONNAIRES 167

APPENDIX G: USABILITY STUDY TEST TAKERS' INTERVIEW 177

Appendix H: Main study examinees' questionnaire 178
Appendix I: Main study invigilators' questionnaire 183
Appendix J: Main study examinees' semi-structured interview 185
Appendix K: Main study invigilators' semi-structured interview 186
Appendix L: Invigilation instructions 187
Appendix M: Pilot study information 188
Appendix N: Descriptive statistics of Moodle-hosted test 191
Appendix O: Person statistics 193
Appendix P: Detailed test takers' questionnaire frequency analyses results 199
Appendix Q: Boxplots of questionnaire analysis results 203
Appendix R: Detailed tables of questionnaire analysis results 242

List of Figures and Tables

Figure 2.1. The structure of an Assessment Use Argument 11
Figure 2.2. Essential components of a CALT interface design process 18
Figure 3.1. The process of developing the Moodle-hosted test validation framework. . 33
Figure 3.2. E-exam accessed through Safe Exam Browser in a full screen mode. 40
Figure 3.3. A snapshot of the split screen mode for reading tests. 40
Figure 3.4. A snapshot of the embedded Matbury's MP3 player for listening tests. 41
Figure 4.1. Item-person map 56
Figure 6.1. Structure of the validity argument about the Moodle-hosted test. 109
Figure 7.1. Screenshot from PARCC reading exam using Notepad to make notes 118
Figure 7.2. Screenshot from PARCC reading exam using Answer Eliminator Tool 118

Table 3.1. Main Study Participating Sample 36
Table 3.2. Examinees' Disciplinary Areas and Courses/Levels 36
Table 3.3. Data Deviating from Normal Distribution 47
Table 4.1. Rasch Item Measures Results: Selected Statistics 54
Table 5.1. Endurance: Comparison with Test Scores 67
Table 5.2. Ease of Use: Comparison with Test Scores 69
Table 5.3. Familiarity and Experience: Comparison with Test Scores 71
Table 5.4. Attitude and Resistance to Change: Comparison with Test Scores 72
Table 5.5. Encountering Technical Issues: Comparison with Test Scores 73
Table 5.6. Sound and Headphones Quality: Comparison with Test Scores 75

TABLE 5.7. SPLIT SCREEN FOR READING AND NOTE-TAKING: COMPARISON WITH TEST SCORES......77

TABLE 5.8. TEST MODE AND FEEDBACK: COMPARISON WITH TEST SCORES......................................78

TABLE 5.9. APPROPRIATENESS FOR TESTING PURPOSE: COMPARISON WITH TEST SCORES................80

TABLE 5.10. TIME MANAGEMENT: COMPARISON WITH TEST SCORES 82

TABLE 5.11. SUMMARY OF RESULTS FOR TECHNOLOGY-RELATED VARIABLES 83

TABLE 6.1. LITERATURE AGREEING AND DISAGREEING WITH STUDY FINDINGS..............................107

TABLE 7.1. PRACTITIONER ADVICE ON ADDRESSING TECHNOLOGY-RELATED ISSUES 120

TABLE 7.2. AGENDA FOR FUTURE RESEARCH 125

List of Abbreviations

AU	Assessment Unit
AUA	Assessment Use Argument
CALT	Computer-Assisted Language Testing
CELP	Credit English Language Program
CMS	Course Management System
CPS	Centre for Preparatory Studies
CTT	Classical Test Theory
EFP	English Foundation Program
ICT	Information and Communication Technology
IRT	Item Response Theory
LC	Language Centre
MTT	Modern Test Theory
RQ1	Research Question One
RQ2	Research Question Two
SEB	Safe Exam Browser
SEM	Standard Error of Measurement
SQU	Sultan Qaboos University
UQ	The University of Queensland

Chapter 1. Introduction

1.1. Introduction

With the increasing demand for twenty-first century skills such as information and communication technology (ICT) skills, technology now plays a major role to facilitate the use of these skills in teaching and assessment practices (Griffin, McGaw, & Care, 2012). E-exams and e-assessments (www.transformingassessment.com and www.transformingexams.com) have become a major component in today's technology-driven world. More specifically, the rising pressure on language testing and the potential efficiencies offered by technology has led to an increasing use of computer-assisted language tests (CALTs) hosted as web-based tests on course management systems (CMSs) like Moodle. Questions over the validity and standards of CALTs and concerns over test fairness have been raised in the literature. Technology-related issues have been reported to affect test performance due to the testing mode effect in the computerised testing environment (Chapelle & Douglas, 2006; Fulcher, 2003). Therefore, in order to provide evidence that students meet language entry standards of institutions, high-stakes computerised assessments upon which inferences and critical decisions affecting students' future are made need to be validated. This study seeks to address the issue by conducting a case study of administering and validating a Moodle CMS-hosted English language proficiency exit test in a specific English as a foreign language context in the Sultanate of Oman. This chapter will introduce the study context and background. The problem statement, study aims, and thesis structure will also be presented.

1.2. Context and Background

Sultan Qaboos University (SQU) in Oman (http://www.squ.edu.om) was established in 1986 as the first national university in Oman. It now has the following nine colleges: Medicine and Health Sciences, Engineering, Agriculture and Marine Science, Education, Sciences, Arts and Social Sciences, Economics and Political Science, Law, and Nursing. Most students are enrolled in undergraduate programs but postgraduate programs at the diploma, masters and doctoral levels have also been introduced.

The transition into the first year upon acceptance to study in undergraduate programs at SQU is important for ensuring quality within the university. Students who join Oman's national university, SQU, are seen as privileged and highly intelligent individuals who worked hard to pass the high school national exams and were able to get a national scholarship to do a bachelor's degree. In the school system, while the Arabic language is the medium of instruction, the English language is taken as a subject and so it is taught as a foreign language. For most of the educational programs at

SQU and most higher education institutions in Oman, English is the primary language of instruction. One would imagine this to be a big shift for graduates of pre-tertiary education, so the support given to them by the university through the provision of a preparatory studies program is needed. These preparatory studies were made mandatory in all of Oman's higher education institutions (universities and colleges) in the 2010/2011 academic year. Newly-enrolled SQU students join the Centre for Preparatory Studies (CPS) to prepare them to study at the university. The CPS now offers preparatory studies in English language, Mathematics, Information Technology, and Study Skills. Preparatory studies are aimed to support students' academic success upon joining their academic programs in their colleges.

The English Department of the CPS has two programs, the English Language Foundation Programme (EFP) and the Credit English Language Program (CELP). The students take the EFP courses to fulfil the English language requirement of their colleges teaching fully or partially in English. The EFP has six main courses, representing six language levels, to be taken over a semester each with 18 hours teaching load. Within each course, students work on developing their English language and study skills. Students are assessed using a combination of continuous assessment and formal mid and end of semester exams. These EFP courses do not count towards the students' GPA but students are expected to successfully progress throughout each level with at least a grade of C- in order to be able to finally exit the EFP and start their university college courses. Students can be exempted from doing the EFP if they score Band 5 in IELTS (with a minimum of 4.5 per skill) or a total of 61 in TOEFL iBT. In the newly-enrolled intake, students can also be exempted from doing the EFP if they score high on a Placement Test and then pass an Exit Test, both administered by the CPS. The CELP 25 courses are usually taken by students in the first two semesters after finishing the EFP and joining their academic English-medium college courses. The CELP courses support students to use their English language skills for communication in their studies. Just like other academic courses, the CELP courses include assessments and count towards the students' GPA. Overall, the CPS with all of its courses and services to students is integral to the success and sustainability of SQU, especially that students need to have a good grasp of the English language in order for them to succeed academically at the university.

Before establishing the CPS, the Language Centre used to run the English language courses at SQU, but it has recently been restructured and accordingly the institutional name has been changed to the CPS. As already mentioned, with the expansion of the CPS, the English language courses are now run by the English Department at the CPS. Despite the change in the name of the institution, the study context throughout this thesis will be referred to as the Language Centre or LC since it is the name used in most of the context-specific literature and for the sake of consistently presenting the documents in this thesis such as ethics approval and consent forms.

This study took place at the LC and involved teachers and students from the EFP. Since this study focuses on the use of technology in assessment, we need to get an overview of technology use in the Omani context and the characteristics of the Omani students with regards to technology skills. So, we first examine accessibility to technology skills development in Oman. We then discuss research on computerised assessment in Oman and provide background information on research and practices concerning the use of the Moodle platform at the study context.

1.2.1. Technology skills accessibility in Oman.

Oman has witnessed political, economic, social, and educational transformations since 1970. In the early years of that transformation, no formal computer education was provided until students reached tertiary education. This has left earlier generations of school leavers without basic technology skills. However, over the last decade this situation has been rapidly evolving due to education reforms. As described by Rassekh (2004) in an International Bureau of Education report on educational reform in Oman, Omani schools are now equipped with computer laboratories and 264 hours of school time is dedicated to computer studies with 120 hours allocated for information technology as subjects taken during the ten years of basic education. Higher educational institutions such as SQU and colleges of education have also improved their educational infrastructure including networks, computer laboratories and the adoption of modern course management systems.

Furthermore, according to a survey on ICT access and usage in Oman conducted by the Information Technology Authority (December, 2012), figures show that in 2011, 52% of Omani employees had ICT skills. In government and private higher educational institutions, 99% were reported to have computers connected to the internet and 94% of employees in the higher educational institutions (including academic and administration staff) had ICT skills as of 2011. In public schools with computer-assisted instruction the rates of computer use is lower. A 2010 survey indicated that the learners-to-computer ratio was 11.8 and only 15% of public schools had ICT-qualified teachers. However, it was also shown that 85% of public schools (grades 1 to 12) had an Internet connection in 2010.

The Science, Technology and Innovation Policy Reviews prepared by the United Nations Conference on Trade and Development (2014) and current ITU (2016a) figures show a rapid increase in computer and internet availability in Oman. In 2011, 58% of Omani households and individuals had access to a computer and 46% had internet access according to 2012 data. This compares to Australia for the same period where 78% of households had an internet connection in 2010-2011 (ABS, 2016). However in 2016, 83% of Omani households now have internet access (ITU, 2016a) while in Australia this figure stands at 85% (ITU, 2016b). These figures indicate that while technology access has been rapidly expanding in Oman, many recent graduates from pre-tertiary education may have lacked access to ICT, computers and internet in educational institutions, especially in schools, and at home until only recently. The effect of lower rates of IT access in pre-tertiary environments implies a relatively lower level of technology skills development for students entering university in Oman. This means that students joining higher education institutions including SQU may confront IT literacy barriers when taking exams and other e-assessment tasks in a computerised testing environment.

1.2.2. Computerised assessment research in Oman.

With the wide-spread technology use in language instruction, the language testing field has seen major developments shifting test administration from the traditional paper-based exams to CALTs using a wide range of technology tools. This shift has sparked researchers’ interests to examine the effect of the new testing mode on test performance and the overall validity of test results. A number of variables have been under investigation with regards to the use of technology in language testing (Chapelle & Douglas, 2006) such as computer familiarity (Taylor, Kirsch, Jamieson, & Eignor, 1999). Chapter 3 will provide an in-depth review of the research conducted in this area. In light of that literature review, the problem statement will be presented in Section 1.3 of this chapter (pp. 5-7). As for computerised assessment research in Oman, the use of technology-enhanced assessment in the Omani higher education context has scarcely been under empirical investigation, especially in the field of language testing. Therefore, further research is needed into the obstacles interfering with technology-enhanced assessment in Oman.

In terms of assessment at SQU and the LC in particular, students have to cope with a heavy testing culture including continuous and summative assessments. Paper-based exams are among these assessments that students have to cope with. Computerised tests are also increasingly being used for student assessment, especially making use of the SQU e-learning Moodle platform. Moodle has been in use at the study context as the learning management system at SQU (https://elearn.squ.edu.om/login/index.php) since 2005 (Al-Ani, 2013). Moodle has been in operation in EFP courses at the LC of SQU since the Fall semester of 2010 (Al-Busiaidi & Tuzlukova, 2013a). The Moodle-based EFP courses in the six levels of the program function as a language learning environment that is student-centered and communicative in that it supports students’ skills development, independent learning, team work and motivation (Al-Busiaidi & Tuzlukova, 2013b).

With the use of Moodle for assessment purposes in the study context (Al Naddabi, 2007; Scully, 2006), it is necessary to evaluate its use in the language curriculum and the way this impacts students' learning and performance. Context-specific research (e.g., Najwani, 2013; Scully, 2013) and practices related to the use of Moodle as a platform for hosting e-assessments on student performance at the study context has mainly focused on formative assessment scenarios. Research has not been done into the use of Moodle for high stakes testing in the study context. Furthermore, the context-specific literature has reported that the introduction and implementation of e-assessment (Moodle-hosted in particular) at the study context comes with a package of issues needing immediate attention. These issues include the lack of accessibility to technology skills among students (see Section 1.2.1, pp. 3-4); the potentially limited technical resources (such as computers with earphones) and the need for a computerised testing infrastructure (Al-Hajri, 2011; Uddin, Ahmar, & Al Raja, 2016); technical failures such as internet or network outages (Al-Ani, 2013); and administration procedures that compromise test security (Najwani, 2013; Scully, 2013). Such issues become particularly significant when considering the use of Moodle for high-stakes testing.

This study explores the use of Moodle for the delivery of a high-stakes test at the study context. Moodle was used for web-based language testing purposes in this study because of the cost-effectiveness of using it as the e-learning platform at SQU; the useful features of the Moodle quiz (*Moodle statistics*, 2017; *Questions,* 2013); Moodle's enhanced testing security; and its capacity to aid testers in doing statistical item analysis (Coy, 2013; Myrick, 2010). Further details on the Moodle-hosted testing interface features used in the study will be provided in the research methodology in Chapter 3. Based on the background information about the study context given in this chapter and the literature to be discussed in detail in Chapter 2, the following section introduces the problem statement.

1.3. Problem Statement

CMSs such as Moodle have become vital course components for a range of educational contexts including English language programs. E-assessments created using online tools such as the quiz module integrated into CMSs like Moodle are increasingly being used for high-stakes language proficiency testing. As a result of the shift from paper to computer, the testing mode effect related to the computerised nature of such exams (Chapelle & Douglas, 2006; Fulcher, 2003) has the potential to impact on their validity. The practice of carrying out validation research as would normally be the case with traditional paper-based exams is needed in the case of computerised exams as well. Yet there currently exists a lack of empirical validation research focused on CMS-hosted language tests. This study seeks to address this gap by using a case study of administering and validating a Moodle CMS-hosted language proficiency test in order to articulate a validity argument about its score-based decisions. As mentioned in Section 1.2.2 (pp. 4-5), within the study context, there is a lack of validation research with respect to the use of technology-enhanced Moodle-hosted tests. This has acted as an incentive to carry out this study. The inquiry is also significant in light of the potential expansion of the use of computerised tests in official high-stakes testing (such as mid and end of semester exams) at an increasingly larger scale.

At the study context, the LC of SQU, technologies have been integrated in the educational program through Moodle-hosted online tests and quizzes. So far such tests have replicated existing testing practices in an electronic form or as Elliot (2007) describes *e-assessments 1.5,* in the stages of integrating technologies in assessment (Elliot, 2007; Gruba & Hinkleman, 2012). Most of the assessment practices at the study context are still done entirely in the paper-based traditional format. There have been many initiatives to integrate technology into assessment practices such as using Moodle tests and quizzes, Moodlereader, and e-portfolios (Scully, 2006, 2008, 2013). These Moodle activities are used in a blended learning approach (Al-Busaidi & Tuzlukova, 2013a, 2013b; Scully, 2006, 2008, 2013) either as practice materials or as informal assessment tools for some course components. Students' marks on Moodle summative tests or quizzes contribute to their final course grades. From the researcher's prior observations, the Moodle tests used at the study context seem to vary in their difficulty levels for students of a particular course. There does not seem to be pre-set guidelines for the development and administration procedures of these tests, and better control over the quality of these tests is needed. We also find that there is no validation research conducted in relation to the level of tests and quizzes administered online. The researcher believes that empirical research that evaluates and validates these online Moodle-hosted assessment practices is needed in order to indicate whether they measure the intended learning outcomes.

Moreover, as will be presented in Chapter 2, the language testing literature overall lacks guidelines for good practice on the development of a CALT interface (Fulcher, 2003). Such guidelines can be of use to practitioners developing a CALT interface as they help address the technology-related issues that can interfere with test performance. The focus of CALT validation research has to date been in the form of cross-mode comparison, that is, paper versus computer tests (e.g., Al-Amri, 2007; Fulcher, 1999; Weir, O'Sullivan, Jin, & Bax, 2007). There remains a need to examine other validity aspects that are more relevant to the computerised testing mode. This means that we need to investigate features unique to the computerised testing mode by examining technology-related factors that have the potential to interfere with the tested construct (language abilities). CALT has been reported to lead to a testing mode effect posing reliability and construct validity threats and bias concerns in the computerised testing environment (Chapelle & Douglas, 2006; Fulcher, 2003). Hence, researchers need to identify the unreliable and construct-irrelevant sources by studying technology-related variables that can unfairly affect test performance (Chapelle & Douglas, 2006; Douglas & Hegelheimer, 2007; Stoynoff, 2012), taking into consideration key test taker characteristics (Stoynoff, 2012). This is a gap in the literature specifically focusing on test validation and assessing language through technology.

In sum, given the need to investigate reliability and construct validity threats that are idiosyncratic to the features of the computerised testing mode (Chapelle, 2008), this validation research study is focused on the research problem of the testing mode effect that can result from integrating technology in assessment using Moodle-hosted exams. Taking a direction different from the dominant cross-mode comparative validation research, the focus of this study is on the threats posed by the testing mode given the features of the Moodle-hosted testing interface in relation to the characteristics of the test takers such as computer familiarity at the study context. The study will address a number of technology-related issues that might be of concern such as familiarity with technology, technical failures, and administration procedures. Following recommendations in the literature (Chapelle & Douglas, 2006) to utilise an evidence-based interpretive approach in test validation research, these negative aspects or threats of integrating technology in assessment will be incorporated in a structured argument about test score interpretation and use following a specific validation study framework. This framework will be guided by principles of the Assessment Use Argument framework (Bachman, 2005; Bachman & Palmer, 2010), as will be outlined in Chapter 2. Examining this problem can help us understand what testing mode technology-related construct-irrelevant issues can unfairly affect test performance in this testing environment. Guidelines for good practice will also be formulated as by-products of the outcomes of this study.

1.4. Study Aims

This empirical study aimed to address the above-mentioned research problem and to fill in the gap in the literature that lacks validation research on CALTs that are hosted on CMSs like Moodle.

By establishing empirical evidence from the case study of administering and validating the Moodle-hosted test, the overall aim of the study is to articulate a supported validity argument about using the Moodle-hosted test for its intended purpose. The validity argument is intended to be disseminated to stakeholders at the study context as research outcomes highlighting potential issues with this testing mode. This study aims to contribute knowledge to the field based on the implications and recommendations of the study findings about the testing mode effect. Such implications and recommendations may then be considered by practitioners and researchers as guidelines for creating, developing, implementing, and researching reliable and valid large-scale high-stakes tests delivered on Moodle, other CMSs, or any other test delivery technologies. By enhancing our understanding of the testing mode effect and by developing guidelines for good practice when using CALTs on the Moodle CMS, the study outputs can help bring the plans to have large-scale and high-stakes CALTs at the study context to reality.

1.5. Thesis Structure

This introductory chapter has provided a broad overview of the research program. Chapter 2 provides an extensive review of the relevant literature focusing on two main aspects: test validation research and technology-enhanced language assessment. Guided by the literature review as well as the aims and questions of the research study, the study framework will be presented in Chapter 2. Chapter 3 outlines the research methodology including the methodological approach, study design, participants, data collection instruments, and data analysis procedures.

The results of the study will be reported in two chapters. Chapter 4 reports and discusses the results of statistical analysis carried out on the test score data. Chapter 5 presents the findings of statistical and thematic analyses of test taker’s questionnaire responses and the results of comparing these responses to test performance data. Chapter 6 provides a discussion of the study findings. Finally, Chapter 7 is the concluding chapter that provides an overall summary of the study. It also covers other aspects including: study significance and contribution to knowledge, implications and recommendations for practice and future research as made from the study findings, and study limitations.

Chapter 2. Literature Review

2.1. Introduction

The overall aim of this study is to provide a validity argument about using a Moodle-hosted test for its intended purpose by empirically examining reliability and construct validity evidence. This chapter presents a review of the literature on validation frameworks and provides the rationale for selecting the study validation framework. The chapter also provides a review of the literature focusing on two main aspects: test validation research and computer-assisted language testing (CALT) research. Based on the review of the literature, the research gap and research questions will be outlined. Informed by the review of the literature, this chapter will also present the validation framework guiding the study in the construction process of the validity argument.

2.2. Validation Framework

In order to formulate a study validation framework aimed to generate a validity argument and to provide a sound rationale for its selection, the literature on test validation frameworks was reviewed. The review of relevant literature highlighted the use of test validation frameworks to validate language tests as a prominent trend in the field of language assessment. Such validity frameworks have been formulated to guide test developers in their test development process to ensure that certain good testing practices or qualities of good tests are adhered to. Validity frameworks have evolved from the traditional view of validity, exemplified by different types of validity including content, criterion and construct validity types (Messick, 1993), to the Messickian unitary conceptualization of validity (Messick, 1989) and Bachman and Palmer’s (1996) test usefulness framework. Recent validity frameworks include the argument-based validity approaches (Kane, Crooks, & Cohen, 1999; Kane, 1992, 2011); the Evidence-Centered Design framework (Mislevy, Steinberg, & Almond, 2002); and the evidence-based interpretive validity argument approaches including: 1) Bachman’s (2005) and Bachman and Palmer’s (2010) Assessment Use Argument (AUA) framework and 2) Weir’s (2005) socio-cognitive framework.

Bachman (2005) emphasises the significance of considering test use in test design and development, as reflected in the existing literature. Central to this view is the addition of the consequences of test use as an essential element to validity of score-based interpretations (Messick, 1989). Emphasising test fairness has also led to the development of a test fairness framework (Kunnan, 2004), which supports the use of tests in a fair manner. This perspective on fair test use is also connected to the discussions of ethics and validity (Lynch, 2001) and the development of the concept and principles of critical language testing (Shohamy, 1998, 2001). To ensure test fairness, test developers need to consider various factors and issues in their assessment development process and conceptualise all of these in light of a specific systematic framework for their assessment practices (Weir, 2005).

In recent validity frameworks (Bachman, 2005; Kane, et al., 1999; Kane, 1992; Mislevy, et al., 2002; Weir, 2005), there seems to be an agreement on the need to provide evidence to support claims made from score-based interpretations. The assessment validation practice of some language test developers responsible for large-scale assessments used for high-stakes purposes is that they gather evidence to support the validity of score-based interpretations. However, Bachman (2005) criticises that such efforts "are frequently shopping lists of correlations, content analyses, and other evidence collected more or less as time and resources permit [and so, there is a need for] a much more focused, efficient program for collecting the most critical evidence" (p. 32). Any existing data related to an assessment cannot be considered evidence in support of interpretations and uses since "'Data' become 'evidence' only when their relevance to some hypothesis, some inference, some claim, is established" (Mislevy et al., 2002, p. 492). To establish a strong argument, as Bachman (2005) argues, test developers need to gather evidence to refute rebuttals or counterclaims that potentially act as sources of invalidity and may result in what Messick (1989) referred to as construct-irrelevant variance and construct under-representation.

Examination boards and language testers around the world currently adapt such validation frameworks or models to guide their test design and validation practices, especially the AUA (Bachman, 2005; Bachman & Palmer, 2010). In the AUA, Toulmin's (2003) basic argument structure forms the basis for articulating a validity argument for a given assessment. The AUA is depicted in Figure 2.1. The structure of an assessment argument is made of two parts. One part is the assessment validity argument linking assessment performance to assessment-based interpretations. The other part is the assessment utilisation argument linking the assessment-based interpretations or inferences to the intended uses of assessment or claims (decisions to be made), where utilisation refers to making an inference from a score interpretation and linking it to a decision (Bachman, 2005).

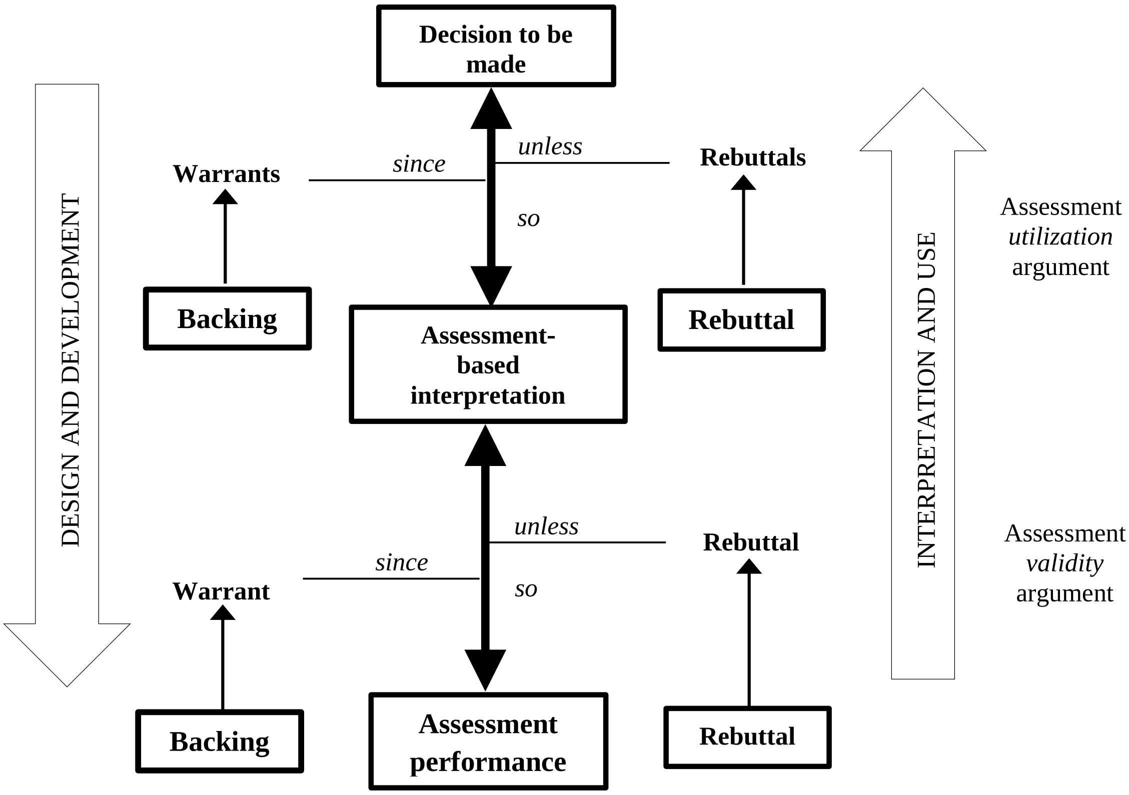


Figure 2.1. The structure of an Assessment Use Argument (adapted from Bachman, 2005, p. 25).

As argued by Bachman (2005), warrants justify the claims or decisions made on the basis of an interpretation. Backing represents the evidence that supports the warrants and can be gathered from a number of sources including results of prior research and conducting specific validation research. Rebuttals act as counterclaims refuting specific warrants. Evidence collected from various sources that rejects the rebuttals can be the backing in support of the claims. The AUA accounts for the qualities of test usefulness developed earlier by Bachman and Palmer (1996) as comprised of reliability, construct validity, authenticity, interactiveness, and impact. Practicality is also reflected here as a quality of the test development process. Bachman (2005) places these qualities of test usefulness (Bachman & Palmer, 1996) in the AUA framework as either warrants or backing evidence to support the claims or decisions to be made. Referring to the principles or features of critical language testing (Shohamy, 2001), Bachman (2005) also argues that warrants such as impact in a utilisation argument take into account the questions of critical language testing on the purposes and uses of assessments. Qualities of fairness (Kunnan, 2004) can also be traced in the AUA, as stated by Bachman (2005). An example of one of the qualities in Kunnan’s (2004) fairness framework is absence of bias, which Bachman’s (2005) AUA accounts for by arguing that potential sources of bias could be considered as rebuttals about unintended consequences that need to be rejected by backing evidence.

As a test validation model, the AUA framework (Bachman, 2005; Bachman & Palmer, 2010) has notable strengths because it reflects a body of empirically-based research in the area. It encompasses commonly researched and cited aspects in test validation tasks including qualities of test usefulness such as reliability and construct validity, test fairness, and critical language testing. Added to that, the AUA framework provides a structured approach to presenting evidence for accepting or rejecting validity. Providing pieces of evidence to support claims, hypotheses, or inferences we make from test scores has been a principle for good language testing practice agreed upon by many researchers (e.g., Bachman & Palmer, 2010; Kane, 1992, 2011; Messick, 1989; Weir, 2005). However, there is no consensus on how this is to be achieved. What is certain is that the existing literature clearly shows that validation studies should employ a combination of research tools to triangulate data in support of a conclusion through the use of multiple evidence sources (Kane, 1992). If a conclusive argument is to be made, it should then be articulated through warrants and rebuttals that are backed with multiple pieces of evidence.

Based on this review of validation frameworks, the framework guiding this study will be specified towards the end of this chapter (Section 2.8, pp. 28-29). The validity aspects to be examined in this study were specified based on a review of test validation research including CALT validation research and studies and CALT research in Oman, which will be addressed in the next section.

2.3. Test Validation Research

Test validation research has evolved from its earlier development to the current views on validation. CALT validation research has also focused on a number of validation aspects.

2.3.1. Earlier development of validation research.

Over the years, test validation research has gone through major developments. Earlier development of test validation research methods witnessed treatment of separate types of validity. In the 1950s and 1960s, content validity and criterion-related validity dominated validation research on discrete-point tests in particular (Lado, 1961). By correlating scores of a language test with scores of another valid criterion or test, test validity can be established indirectly. Content validity evaluates the extent to which test items represent a real life problem. Evidence on criterion-related validity can be established by correlating performance on test items with items measuring the same problem in the criterion test. In that period, reliability was seen as a prerequisite for validity and the methods of establishing reliability, including internal consistency and test-retest reliability, were introduced by Lado (Xi, 2008).

In the 1970s, validation research focused on concurrent or predictive validity and content or face validity (Clark, 1978). Inter-rater reliability became important in validation research of subjectively-scored tests. In earlier developments of validation research, validity was considered as different types and researchers considered establishing one validity type in their validation task sufficient to support the use of a test. Test validation methods in the 1960s and 1970s focused on analysing test items using content and correlational analyses. Factor analytic techniques were common as well (Xi, 2008). In the 1980s, the focus shifted from predictive validity research to studies on the processes of test-taking and the factors influencing test performance (Bachman, 2000; Xi, 2008).

2.3.2. Current validation research.

The current unitary view of construct validity (Messick, 1989) considers different validity types as pieces of evidence that would support a specific test use. The view of construct validity of score meaning or interpretation was expanded by Messick to include the evaluation of test use social consequences and the value implications of test interpretation. Bachman and Palmer (1996) then proposed the test usefulness framework that includes six qualities: validity, reliability, interactiveness, authenticity, impact, and practicality. By applying this framework, practitioners put Messick's unitary concept of construct validity into action in their empirical validation studies. Empirical validation studies started to address other validity aspects such as factors that can affect test performance including test takers, strategies and processes. In these investigations, the focus was shifted from the test to score interpretation for a specific test use, to indicate whether empirically-established evidence can support validity claims for the intended test use. Because of this shift, other aspects that are not considered test qualities became part of the investigations on language test quality. These aspects include fairness (Kunnan, 2004), ethical issues, impact or consequences of test use (Kunnan, 1998), critical language testing (Shohamy, 2001), as well as social and policy considerations (McNamara, 2006; McNamara & Roever, 2006). Such aspects concern wider social and testing policy issues. Triangulated quantitative and qualitative methodologies also became common in validation research (Xi, 2008), especially given the current view of validity types as pieces of evidence to support a particular test use.

2.3.3. CALT validation research.

Part of language test validation research is focused on CALT and deals in one way or another with validity aspects or factors that can affect test performance within the term *testing mode effect*. In addition, from the point of view of fairness and avoidance of bias, the *Standards for educational and psychological testing* (1999) by the National Council on Measurement in Education, American

Psychological Association, and American Educational Research Association point to the need to address what is termed *construct-irrelevant variance* associated with CALT (such as examinees' familiarity with technology and test format) in test design and use. Construct-irrelevant variance and construct-irrelevant (technology-related) factors that can contribute to test performance in a CALT refer to the *testing mode effect* term, although studies might not have directly referred to the testing mode effect as a construct-irrelevant variance. According to the *Standards for educational and psychological testing* (1999), construct-irrelevant variance "refers to the degree to which test scores are affected by processes that are extraneous to its intended construct" (p. 10). Davies, Brown, Elder, Hill, Lumley, and McNamara (1999) also state that construct-irrelevant variance is a

> type of systematic measurement error where there is some variance in the test scores that is due to factors other than the construct in question.…Such variance contaminates the interpretations that are made on the basis of test scores, and hence negatively affects the construct validity of the test (pp. 32-33).

As argued by Brown (2005), test score variance exhibits "meaningful variance" and "error variance" (p. 290). Score variance can be meaningful variance that is attributed to the test purpose or a measurement error variance that is attributed to other sources such as problems in test items, personal issues, and scoring procedures. As explained by Brown (2005), a reliability estimate of, for example, 0.91 indicates that 91% of the test variance is reliable and the remaining 9% variance is measurement error that stems from sources irrelevant to the test construct. Examining the sources of measurement error tells testers about unreliability, which reveals the true reliability estimate. Sources of measurement errors that can reduce reliability are attributable to candidates' characteristics (such as guessing, anxiety, motivation, and test wiseness); testing situation (such as environment factors as in noise, space, and lighting; and factors associated with test administration procedures as in equipment, directions, and timing); scoring procedures (such as scoring errors and scorer biases); and factors associated with the test and items (such as test security, test format, item types, and clarity) (Brown, 2005; Davies, et al., 1999). Test reliability and construct validity will be affected by the level of error in test results and bias or systematic error gets introduced as a test turns out to be systematically measuring something other than the intended test construct (Davies, et al., 1999).

2.3.4. CALT validation aspects.

Just like paper-based exams, validation studies on computerised web-based exams have reported incorporating reliability analyses such as a validation study by Chapelle, Jamieson and Hegelheimer (2003) of a web-based English as a second language test (ESL) and another validation study of a web-based test of ESL pragmalinguistics by Roever (2006). When it comes to the relationship between reliability and validity aspects in validation research, a test must be reliable for it to be valid as scores should reflect test takers' actual differences or otherwise they would be due to chance (Roever, 2006). This is supported by the argument made by Brown (2005) that reliability is a validity precondition since a test should be proven to be consistent to claim that it is systematically measuring what it is purported to measure. In the computerised testing mode, as the construct represented by the test may change due to the testing mode effect (Fulcher, 1999), construct-irrelevant variance should not be reflected in the test scores as the test should mirror the construct being tested only. As emphasised by Roever (2006), though a high reliability estimate is necessary to support claims of construct validity, it is not a sufficient condition. This means that although obtaining reliability estimates in validation tasks tells researchers that the test is systematically testing the construct being measured, it is essential to identify potential sources of construct-irrelevant (unreliable) variance that can jeopardise construct validity.

Therefore, CALT validation research or CALT evaluative research looking into the testing mode effect needs to reflect the new view of validity as a unitary concept. Given that the construct-irrelevant technology-related factors can be sources of measurement error, researchers need to examine such factors in order to establish evidence on CALT validation aspects such as reliability and construct validity in support of test use following a sound validation framework like the AUA.

2.3.5. CALT validation studies.

When reviewing CALT validation studies, we find that although language testers and researchers reporting on using technology for language assessment praise technology affordances and advantages (e.g., Chapelle & Douglas, 2006; Weir, O'Sullivan, Jin, & Bax, 2007; Yu, 2010), studies still report some disadvantages and issues or threats to validity (Chapelle & Douglas, 2006). Arguments for test fairness and avoidance of bias get raised when a test or item feature that is irrelevant to what is being tested advantages or disadvantages a particular test taker group(s) (Brown, 2005). In light of these arguments, minimising sources of unfair technology-related issues (or the test mode effect threatening validity) in the test design and implementation stages becomes critical and it should be thoroughly investigated (Douglas & Hegelheimer, 2007; Fulcher, 2003).

Ideally, as noted by Chapelle and Douglas (2006), when evaluating a CALT using a specific framework such as Bachman and Palmer's (1996) test usefulness framework, researchers need to provide evidence of the six test qualities: construct validity, reliability, authenticity, practicality, impact, and interactiveness. Studies (Chapelle, 2001; Chapelle, et al., 2003) have used this approach in their CALT evaluation. Chapelle's (2001) study identified a number of technology-related positive and negative CALT features. Chapelle and Douglas (2006) later updated their work listing these features. Under the impact quality, they argued that examinees without extensive technology use experience might get test anxiety. Referring to construct validity, they state that examinees' performance on a CALT might not reflect the same ability measured by other types of assessment. As mentioned earlier, collecting combined evidence on both reliability and construct validity is recommended in language testing validation research to address concerns with the testing mode effect construct-irrelevant technology-related issues (Roever, 2006).

When reviewing CALT research in the higher education context in Oman, we find that the use of technology-enhanced assessments, especially in the field of language testing, has scarcely been under empirical investigation. Exceptionally, Al-Hajri's (2011) study examined the social and psychological factors that might affect Omani higher education students' performance when taking an English language computerised assessment. Factors that were envisaged as irrelevant to the test construct were investigated including test takers' gender, college of study and geographical region, computer experience, and computer self-efficacy. In a more recent study, Uddin, Ahmar, and Al Raja (2016) also surveyed perceptions towards online examinations. One hundred students in the management major at the College of Commerce and Business Administration, Dhofar University in Oman were surveyed. The findings showed that students prefer e-examinations over traditional paper-based tests. The study reported that students agreed that computerised tests enable them prompt access to their results; improve the quality and standard of examination results; and eliminate biases in test administration and scoring. It was also reported that students disagreed that computerised tests will facilitate paperless examination in the university; will help in identifying students who demonstrate best abilities in various courses; will help identify students with learning difficulties; and will eliminate examination frauds and other unethical behaviors. Both studies (Al-Hajri, 2011; Uddin, et al., 2016) recommended the provision of sufficient material and human resources infrastructure for computerised testing in Oman and to prepare test takers for taking exams in this testing mode.

As outlined in Chapter 1, validation research at the level of high-stakes language tests is needed to address potential testing mode effect issues at the study context. Such research is necessary given that a number of issues have been reported about the use of computerised tests (including Moodle-hosted exams) at the study context. These issues include the students' lack of accessibility to technology skills; the limited technical resources and the need for a sound infrastructure for computerised testing (Al-Hajri, 2011; Uddin, Ahmar, & Al Raja, 2016); technical failures such as internet or network outages (Al-Ani, 2013); and test administration procedures that compromise test security (Najwani, 2013; Scully, 2013).

In light of the review of CALT validation studies, including the research conducted in Oman, we conclude that with the limited research on technology-enhanced assessment in Oman, further research is needed to look into the obstacles interfering with this testing mode in Oman. The need for such research is also echoed in the overall test validation literature that has highlighted a number of technology-related construct-irrelevant factors related to the testing mode effect, as will be outlined in Section 2.5 (pp. 18-26). To unpack these factors, CALTs need to be evaluated, as discussed in the next section.

2.4. Evaluation of CALTs

When adopting CALTs, testing researchers and practitioners need to have clear guidelines on how to evaluate these tests (Chapelle & Douglas, 2006). Attempting to construct and follow certain guidelines and standards in the evaluation of CALTs, testing researchers and practitioners have built on general principles followed in the field to evaluate the quality of tests. However, the quality of CALTs, no doubt, feature unique properties that should be carefully considered. The first evaluation criteria relevant to CALT were established by the *Guidelines for computer-based tests and interpretations* (1986) authored by the American Psychological Association’s Committee on Professional Standards and Committee on Psychological Tests and Assessment. The evaluative research on CALT has focused on the advantages versus potential disadvantages or threats of CALT (Chapelle, 2001; Chapelle & Douglas, 2006). As Chapelle and Douglas (2006) recommend, any threats or negative aspects of using technology in language assessment need to be integrated in an overall argument for test score-based interpretation and use following an argument-based evidence-supported interpretive approach such as the AUA (Bachman, 2005; Bachman & Palmer, 2010). This means that evaluation research of CALTs that follows an approach such as the AUA framework needs to incorporate evidence on the potential disadvantages or threats of CALT into the argument for test score interpretations and uses.

Researchers (e.g., Douglas & Hegelheimer, 2007; Fulcher, 2003) have noted the need to control technical aspects in a technology-based test environment (or the effect of the test delivery mode using computer or technology). This is to avoid technical aspects becoming sources of construct-irrelevant variance affecting test performance. If test performance is affected by these construct-irrelevant technology-related factors, scores will be worthless (Fulcher, 2003). Pointing to the lack of guidelines for good practice in the language testing literature on the development of a CALT interface, Fulcher (2003) proposes a model of such guidelines for good practice in a process of three phases: 1) initial planning and design; 2) usability testing; and 3) piloting or field testing and fine-tuning. Fulcher (2003) further argues that practitioners can avoid the threat of creating interface-related construct-irrelevant variance in test scores by following an interface design principled approach. Test developers need such guidelines to consider for good interface design and development of CALTs. Figure 2.2 demonstrates essential components of a CALT interface design process in which validity evidence can be provided to support test use based on an evaluation of CALT.

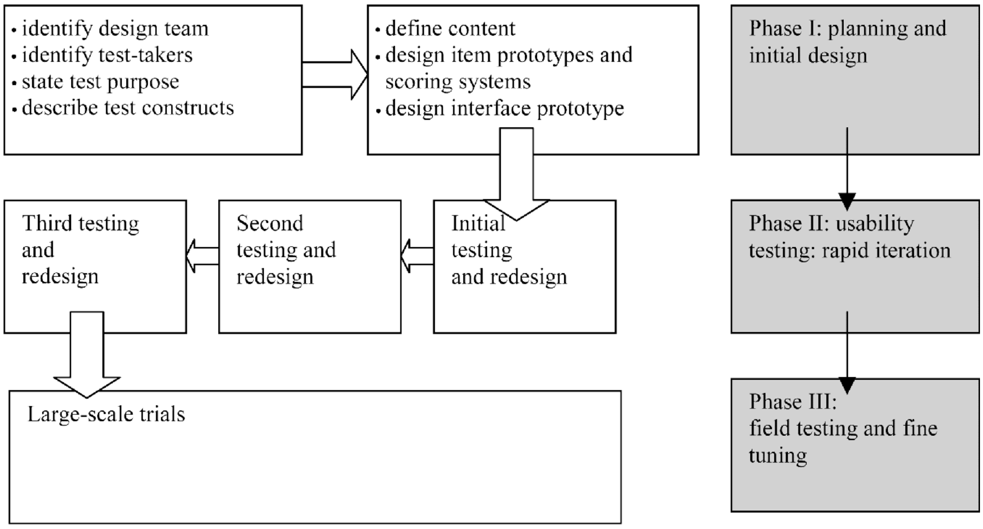

• identify design team
• identify test-takers
• state test purpose
• describe test constructs
• define content
• design item prototypes and scoring systems
• design interface prototype
Phase I: planning and initial design
Third testing and redesign
Second testing and redesign
Initial testing and redesign
Phase II: usability testing: rapid iteration
Large-scale trials
Phase III: field testing and fine tuning

Figure 2.2. Essential components of a CALT interface design process (Fulcher, 2003, p. 386).

2.5. Testing Mode Effect

A considerable proportion of the language testing literature that reports on the validity of tests does so by conducting comparability studies (e.g., Al-Amri, 2007; Wagner, 2010; Weir, et al., 2007). These tend to a) compare test performance on paper-based and computer-based tests as two testing modes; b) establish equivalence between the two modes; and c) compare whether the two modes measure the same construct. As such, validation research has focused on how the testing mode affects validity of score-based inferences with reference to a paper-based mode. The comparative studies looked at factors such as gender-related differences, regional differences, computer familiarity levels (Al-Hajri, 2011; Coniam, 2006; Taylor, Kirsch, Jamieson, & Eignor, 1999), first language, and socioeconomic status (Stoynoff, 2012). A number of factors related to the testing mode effect were identified as technology-related construct-irrelevant factors such as computer familiarity, keyboarding proficiency, equipment quality, attitude, and timing.

2.5.1. Computer familiarity.

The importance of the familiarity and experience variable has been addressed by researchers (e.g., Kirsch, Jamieson, Taylor, & Eignor, 1998; Weir, et al., 2007; Taylor, Jamieson, Eignor, & Kirsch, 1998). While some studies established that test performance was not affected by the lack of prior experience with computers or computer familiarity (Maycock & Green, 2005; Taylor et al.; 1998;

Weir, et al., 2007), other studies (Fulcher, 1999; Russell, 1999) have found a link between the familiarity variable and test performance.

Taylor et al. (1998) studied the relationship between computer familiarity and performance on computer-based TOEFL test tasks. The main finding of their study was that the evidence did not suggest that the lack of prior experience with computers affected performance on the computer-based TOEFL. In Taylor et al.'s (1998) study, having been given an introductory tutorial, no meaningful differences in test scores were found between candidates familiar and non-familiar with computers. Similarly, in a study by Maycock and Green (2005) investigating the impact of computer familiarity and attitudes towards computer-based IELTS on test performance, computer familiarity had no significant effect on test scores. Examinees were also familiarised with this computer-based IELTS by an introductory tutorial and sample materials. In another study examining the IELTS writing paper-based and computer-based versions in which no such introductory tutorial was given to test takers, Weir, et al. (2007) established no connection between performance on the test and computer familiarity. However, even in the event of not finding a significant effect of computer familiarity on test performance results, Weir, et al. (2007) argue that computer familiarity cannot be overlooked when comparing paper-based and computer-based tests.

In Fulcher's (1999) study of an ESL placement test that examined the presentation mode effect, mean score differences were found significant on a web-based test, but not significant on a paper-based test. In another study by Russell (1999), performance of test takers with more keyboarding experience was better on open-ended test items of a computer-based test. Both studies (Fulcher, 1999; Russell, 1999) reported familiarity and experience as a variable that can significantly affect test performance, considering it an indicator of bias and an equity issue (Fulcher, 1999).

2.5.2. Keyboarding proficiency.

Typing responses for constructed-response test items is another technology-related factor that has attracted attention in the literature. This factor is connected to candidates' typing and keyboarding skills as well as their computer familiarity and experience. Typing responses in computer-based tests is an issue that has been examined in a number of studies. Hillier (2015) reported student opinions on computerised testing through surveys conducted prior, during, and after mid-semester trials on an e-exam system. Students had a choice of typing or handwriting. Among the participating students' views, there was a range of positive and negative perspectives. One of the concerns that were voiced was "typing proficiency" (p. 582) as students who typed their exams in the trials reported that typing would be more time efficient for them and their good typing skills would put them at an advantage. On the other hand, students who hand-wrote their exams in the trials reported they had poor typing skills.

Arguing that differences in the typing speed of test takers can be sources of error variance in test scores, Roever (2001) reported a study in which examinees who were given 60 seconds per item were able to complete 99% of each of the two multiple-choice sections of the test. However, although they were given 90 seconds per item, they could only complete 83% of the section in which they had to type brief responses. Examinees were not tested for handwriting the responses. Although it was recognised that raising the time for responding to items requiring typing would be an option, Roever reported that the native speaker comparison group did not have a problem with typing speed. Therefore, Roever's (2001) findings raise concerns about what impact second language students' varying levels of keyboarding skills including typing speed can have on time-limited test performance.

Furthermore, in another study by Coniam (1999), students had a positive attitude towards taking a computer-based test when the testing task was limited to just selecting an answer in a multiple-choice type test. When the testing task was more demanding as it required test takers to type in words or phrases, test takers' preference was more for a paper-based version of the test. As argued in Coniam (2006), this is an indication that examinees' negative views towards taking computer-based tests might not be attributed to computer familiarity and accessibility only, but question type (such as multiple-choice or constructed-response) is also of importance in shaping these views. Investigating TOEFL-iBT writing tasks, Barkaoui (2014) also found that the keyboarding skill had a significant but a small effect on test scores. Barkaoui concluded that test performance mainly reflected test taker English language proficiency and writing ability, but argued for redefining the construct tested by these writing tasks to include keyboarding skills. This argument is based on the increased use of language through computers in academic contexts. Hence, the literature overall signals the need for further examination of new item types especially the constructed-response ones requiring typing of responses in the computerised testing mode, in order to understand whether this variable interferes with test performance.

2.5.3. Equipment quality.

Another technology-related variable that has been studied is encountering technical glitches during CALT administration such as problems with the headphones used for listening tests. For example, in a field trial device effect study of the National Assessment Program for Literacy and Numeracy (NAPLAN) in Australian schools, Davis, Janiszewska, Schwartz, and Holland (2016) reported that there were four technical issues in the use of headphones in listening to the audio part of the spelling test in NAPLAN. One problem was that some headphones had to be replaced as they did not work. Some students also experienced difficulty hearing the audio recording and had to replay the recording due to the poor quality of the headphones. For a few other students, the ability to hear the recording was affected due to the headphones quality, but their headphones did not need to be replaced. Another issue was to do with the size of the headphones as some Year 3 students' headphones were too large and uncomfortable. With NAPLAN in Australia being changed from a paper-based mode to a computer-based mode, assessments can be conducted online or onscreen without an internet connection starting in 2017. Hence, this device effect study informed the Australian Curriculum, Assessment and Reporting Authority (ACARA) (2016) about the minimum technical requirements for conducting NAPLAN online including device screen display; headphones, earphones, or earbuds; keyboards; pointing devices; network; and security using the NAP secure browser application. The headphones quality issues highlighted by the device effect study (Davis, et al., 2016) reflect what might happen in any other real-life testing situations as technical problems might be inevitable.

In research by Choi, Kim, and Boo (2003) that compared test performance across modes, the computerised testing mode significantly affected listening test performance. Hamouda (2013) also identified that poor quality equipment resulted in poor sound quality and interfered with students' listening comprehension. In Arnold's (2000) study, students also experienced anxiety while processing the listening test input, suggesting "acoustic inadequacies" as a factor that leads to such interference with test performance (p. 779). Using headphones with specifications of three sample rates (44 kHz, 22 kHz and 11 kHz) and two sample depths (16 bit and 8 bit), Yang (2009) also found statistically significant differences in test performance among students. Such research outputs indicate the significance of carrying out research that looks into technical issues that can be encountered during CALT administration.

2.5.4. Attitude.

Another technology-related factor is the attitude towards the two testing modes, paper-based and computer-based. A number of studies have investigated students' attitudes to digital delivery of testing tasks. In Fulcher's (1999) study, test takers' attitudes were examined by asking them if they preferred the paper-based testing format or the Internet-based testing format. Test takers were also requested to indicate on which test they would perform best and to nominate which they would choose if given the choice. Fulcher's (1999) study found that regardless of preferences for a testing mode over another, test taker attitudes had no significant effect on the computer-based test scores.

In another study, Singer and Alexander (2017) examined differences in reading across mediums using digital and print versions of book excerpts and newspaper articles. Before reading the texts, students' topic knowledge and their medium preferences were assessed. After reading, students were asked about which reading medium they comprehended best. Results indicated that 69% of students expected their comprehension to be better when reading digital texts, but their comprehension task performance outcomes were not as consistent with their views. While no differences in performance across mediums were shown in the task of identifying the main idea of the text, students did better when reading in print in the task of recalling key points and other relevant information. The researchers state they cannot assume that the mere preference for reading in a digital environment means that students are well-prepared to comprehend digital reading texts.

Findings of a study by Fan and Ji (2014) supported that personal characteristics including attitudinal factors can affect test performance as it was reported that a significantly small percentage of test score variance was explained by attitudinal factors. In Maycock and Green's (2005) study, varying attitudes were reported as 35% of respondents indicated preferences for the paper-based version of IELTS writing, while 41% of respondents indicated preference for the computer-based version and 24% did not report a preference. For the item asking about the preference for the computer-based test to the paper-based test, no statistically significant effect on test performance was found. Furthermore, research by Stricker, Wilder, and Rock (2004) reported positive attitudes towards the computer-based TOEFL among test takers. These attitudes were reported to have a moderate correlation with test performance.

As students come to the testing room with different attitudes and preferences towards the testing modes, whether these attitudes affect their test performance or vice versa still remains a grey area that needs to be studied.

2.5.5. Timing.

Timing exams and the sufficiency of the allocated test time can be deemed a construct-irrelevant factor in any testing mode because they reflect student test time management ability and not their abilities on the tested construct. Hence, test timing has been extensively referred to in the literature. The effect of test time and by default the effect of test length have been issues of concern echoed in many studies. Although having more items on a test can get more information about student ability (Green, 2013), lengthy tests require more time which may affect test performance.

Yamamoto (1995) reported a study that examined the effect of TOEFL test length and test time using a HYBRID model (Yamamoto, 1990). The study evaluated "test speediness" (p. i) by estimating the proportions of test takers who switch from a response strategy based on their ability to a random or guessing response strategy at any time during the test because of being confounded by the time limit to respond. Test length had a small effect on the proportion of test takers affected by test speediness. The proportion of examinees that were affected by test speediness was greater when taking a shorter test that was limited to 50 minutes than when taking a test of 55 or 60 minutes. The study also reported that after finishing 80% of the shorter duration exam, about 20% of the test takers responded randomly to the TOEFL multiple-choice test items. Therefore, their true language abilities were not reflected by the last 20% of the test. The findings of Yamamoto's (1995) study suggest that the time limit can affect test performance and when this is inadequate, test takers resort to a guessing response strategy. The inadequate available time can therefore become a confounding factor that is extraneous to the tested construct. To avoid test speededness, Parshall, Spray, Kalohn and Davey (2002) recommend setting a maximum time limit so that all examinees get sufficient time to finish the test. They also state that taking longer tests can amplify examinees' fatigue even if adequate time is provided to answer all questions. Hence, both test length and test speededness must be considered together to ensure the test reflects examinees' true ability levels.

Hale's (1992) research on the *Test of Written English* also reported that student test performance under the 45-minute test condition was significantly higher by about 1/4 to 1/3 point (on a 6-point scale) than it was in the 30-minutes test condition. Furthermore, Powers and Fowles (1996) found that allowing more time had a positive effect on test takers' performance, which was significantly better on a GRE writing essay test when taken in 60 minutes than it was when taken in 40 minutes. When comparing 15 and 30 minutes testing time conditions, Crone, Wright, and Baron (1993) also found that students scored significantly better when given more time on the SAT II writing task.

In another study by Kroll (1990), a small but insignificant difference was found in test scores obtained from 60-minute timed essays versus take-home essays written over an extended period of 10-14 days. Livingston (1987) also reported that essay test scores increased slightly (with a small effect) by increasing the time limit from 20 to 30 minutes, and that the more proficient students tended to be affected by the test time limit by about half a point (on the 2 to 12 scale). However, other studies did not find test performance differences under different test time conditions. For example, Knoch and Elder's (2010) study did not find examinees' scores on a writing test to be significantly different under long (55 minutes) and short (30 minutes) time conditions. Furthermore, Ghanbari, Karampourchangi, and Shamsaddini (2015) considered the time pressure variable as a non-linguistic factor that had no effect on writing test performance.

In a CALT mode, other mediating factors such as typing ability, usability of test software and familiarity with the testing system may alter the ideal duration and speediness of the test. Therefore, this variable needs to be studied further in the context of CALT validation research.

2.5.6. Eye strain.

The issue of eye strain or eye fatigue is another factor identified in CALT testing mode effect literature. Eyestrain is a symptom of Computer Vision Syndrome and "refers to computer users' subjective complaints about uncomfortable, painful, and/or irritable visual experiences" (Yan, Hu, Chen, & Lu, 2008, p. 2030). This issue has been identified by Dillon (1992) as one of the factors examined by ergonomic research on the presentation mode (paper versus computer screen presentation modes) on reading. The issue of eye strain also emerged as a theme in students' comments in Hillier's (2015) pre-exam survey results that were conducted prior to administering a series of bring-your-own-device (BYOD) e-exams.

Furthermore, a study in a Norwegian school context by Mangen, Walgermo, and Bronnick (2013) looked at the impact the technological interface had on reading comprehension. There were two student groups, one of which read two texts in print format and the other group read them in a PDF format on a computer screen. The findings showed that students reading in print scored significantly better on the reading comprehension test than students reading from laptop computer screens. The researchers argue that reading comprehension may be impeded by particular features of digital screen text display. They also imply that reading performance might be obstructed by scrolling through texts longer than a page and by the "the lack of spatiotemporal markers of the digital texts to aid memory and reading comprehension" (p. 67). In their study, it was not possible to determine whether visual fatigue could affect participants' reading performance when using laptop computer screens. However, they argue that reading processes including identifying letters and words rely on visual text legibility, which can be influenced by a number of factors such as screen resolution, contrast levels, and backlighting. Therefore, visual processing of digital texts can negatively impact higher-level processes including reading comprehension.

In another study by Singer and Alexander (2017), students performed better when reading in print than when reading texts digitally on a computer screen. In their discussion, they address how visual challenges in digital mediums can add to demands triggered by navigation tasks using scrolling and page turning. Nevertheless, they conclude that the findings of their study did not provide evidence on what impact visual ergonomics of the computer screen had on students' performance. Furthermore, time length (or duration/time of the test here) is an important factor in computer use. As reported by Trusiewicz, Niesluchowska, and Makszewska-Chetnik (1995), using the computer for longer periods of time decreased visual functions and could cause eyestrain. In another study, Benedetto, Drai-Zerbib, Pedrotti, Tissier, and Baccino (2013) examined the effects of two display technologies, the electronic ink (E-ink) and the liquid crystal display (LCD), on visual fatigue in lengthy reading sessions. The researchers used an eye tracking technology to objectively measure eye blinks per second and also used a “Visual Fatigue Scale” (p. 4), which is a rating scale of visual fatigue as a subjective measure. They found that higher visual strain was triggered when reading on the LCD of Kindle Fire HD e-reader device compared to when reading from a Kindle Paperwhite E-ink device and a paper book. It was also shown that reading from the E-ink and paper were very similar. Though such findings are device-specific, they have sound implications for the effects of visual fatigue on reading processes and performance especially in prolonged visual activity in lengthy reading sessions.

Clearly, eye fatigue may also be triggered by reading from a book or paper when there is poor light in the room, just like screen light can cause this problem. In the case of CALT, as eye fatigue is not part of the tested construct, it is a technology-related factor that needs to be considered in the implementation and validation of computerised exams.

2.5.7. Other factors.

The factors so far discussed have been seen largely by CALT research through the lens of cross-mode comparative study designs comparing paper and equivalent on-screen testing formats. However, there are additional test features only afforded when using post-paper computerised test designs. CALT research is needed that focuses on additional sources of validity threats that may become apparent when post-paper CALT testing mode features are used. This is because technology introduces a number of new construct-irrelevant factors in the CALT testing mode. One factor that is idiosyncratic to the computerised testing mode relates to screen layout and scrolling features of the computerised testing interface. Dyson and Kipping (1998) and Fulcher (2003) argue that readers who are unfamiliar with scrolling can get distracted by the way the reading text is presented. It is, therefore, recommended to keep scrolling to a minimum (Fulcher, 2003).

Therefore, as recommended by Care, Luo, Awwal, and Yasotha (2015), specific layout and scrolling features used in a testing interface should be examined and the use of other technical features such as note-taking and text highlighting in a computerised exam also need to be investigated as part of the testing mode effect research. Other factors include font size, window size, window flexibility, navigation between stimuli and between questions, inclusion of multimedia, and including interactive tools and new question types such as drag and drop and hot-spot. Since new technologies allow the design of new and innovative item types, research is also needed to examine how well these item types function in a CALT (Chapelle & Douglas, 2006; Douglas & Hegelheimer, 2007).

Test security compromising CALT validity inferences is also another area that needs to be further investigated and resolved by more advanced technologies (Chapelle & Douglas, 2006; Douglas & Hegelheimer, 2007). The use of unfamiliar security technologies in the testing event might also interfere with the test takers' performance, making this area even more in need for further research to identify their contribution to test performance.

Scoring and its inaccuracies posed by the use of technology is another area that calls for researchers' attention. Comparative research on machine versus human scoring still has to continue to arrive at more accurate automatic response scoring systems (Chapelle & Douglas, 2006; Jordan, 2008) that minimise invalid test score interpretations and negative test consequences. When it comes to reliability, yielding consistent and reproducible test scores is an advantage CALT has over paper-based testing modes, especially when the latter do not provide objective scoring procedures and are often associated with human errors (Noijons, 1994). Relevant to scoring inaccuracies is the concept of consequential validity, which is an important piece of evidence in support of inferences made from test scores (Bachman, 2005; Bachman & Palmer, 2010; Weir, 2005). Hence, consequences of CALT exhibited in test bias and negative impact must be at the forefront of the CALT research agenda (Douglas & Hegelheimer, 2007; Stoynoff, 2012) as negative consequences can be a threat to validity.

Overall, further research on the testing mode effect on test taker performance has been called for by researchers (Chapelle & Douglas, 2006; Douglas & Hegelheimer, 2007; Stoynoff, 2012) in relation to particular characteristics of the test taker population taking large representative samples (Stoynoff, 2012). Within the frame of the concept of systematic measurement error or bias and its potential sources in a CALT situation, it is argued here that validation research focusing on the testing mode effect needs to include potential sources of technology-related construct-irrelevant variance. This is because the testing mode introduces a range of features, a number idiosyncratic to CALT (e.g., test security provisions, interface clarity, equipment or hardware and software provisions) that test takers of different characteristics (e.g., computer familiarity, gender, age, and attitude) will encounter in the test situation.

2.6. Summary of Literature Review

As set out in the literature review, technology affordances and features may be taken by some critics of CALT as introducing construct-irrelevance variance due to the testing mode effect. The presence of these sources of technology-related construct-irrelevant variance may lead to debates about the test construct definition and arguments against the reliability and validity of CALT score interpretations and uses. Relevant literature (e.g., Brown, 2005; Fulcher, 1999, 2003; Taylor, et al., 1999) has made the call for future studies to investigate how particular technology-related variables pertinent to the test mode effect can contribute construct-irrelevant variance into test scores. Such studies should aim to understand these variables so that practitioners can deal with them more effectively in order to minimise or eliminate the testing mode effect.

In summary, the literature review of test validation research informed the study by identifying the need to articulate a validity argument from the study findings following the principles of an evidence-based validation framework, namely the AUA (Bachman, 2005; Bachman & Palmer, 2010). Research on the assessment of language through technology also helped identify guidelines for developing a technology-enhanced language testing interface, especially the need to investigate and provide evidence about the testing mode effect construct-irrelevant technology-related factors. Earlier examination (Chapter 1) of context-specific practices and research identified the need for further in-depth research on the use of Moodle as the technology to host the aspired-for high stakes large-scale e-assessments at the study context. To bring it all together, the next section describes the gap in the literature and research questions.

2.7. Research Questions

As mentioned in the previous section, based on the literature review focusing on test validation and assessing language through technology, further CALT validation research is needed to address the testing mode effect issue that can threaten reliability and construct validity of test score interpretations (Chapelle & Douglas, 2006; Fulcher, 2003). As discussed in the problem statement (Section 1.3 in Chapter 1, pp. 5-7), there is also a lack of research addressing the issue when using Moodle-hosted language tests at the study context. To bridge this research gap, the research problem was investigated in this study by administering a Moodle-hosted language proficiency Exit Test and articulating a validity argument about its score-based decisions following the AUA framework (Bachman, 2005; Bachman & Palmer, 2010). The study framework will be described in the next section. The research questions guided the study to investigate the research problem of the testing mode effect and achieve the overall study aim (Section 1.4 in Chapter 1, pp. 7-8).

The overall study aim was to provide a validity argument about using a Moodle-hosted test for its intended purpose by empirically examining reliability and construct validity evidence.

To achieve this overall aim of the study, the study was guided by two research questions:

- RQ1: To what extent can the Moodle-hosted test scores be reliable and valid indicators of the tested construct?
- RQ2: To what extent can technology-related construct-irrelevant factors affect the reliability and construct validity of the Moodle-hosted test?

RQ1 examines the extent to which the test scores can be reliable and valid indicators of the tested construct by applying a statistical test data technique that is specified in in the research methodology in Chapter 3. The assumption or claim here is that statistical evidence of high reliability estimates, fit and highly discriminating items, and acceptable low Standard Error of Measurement (SEM) will be a warrant that the Moodle-hosted test scores will be reliable and valid indicators of the tested construct. RQ2 investigates the extent to which technology-related construct-irrelevant variance factors associated with the testing mode effect can interfere with test results and become potential sources of measurement error variance and hence impact the reliability and construct validity of the Moodle-hosted test. The assumption here is that statistical and non-psychometric types of evidence will be a warrant that the test is measuring what it is supposed to measure, which is English language abilities. Such evidence should support the claim that the testing mode does not introduce construct-irrelevant variance or measurement error, and that test performance is not to a great extent influenced by construct-irrelevant technology-related factors. The overall aim of the study can then be achieved based on the evidence established by examining reliability and construct validity aspects as outlined in RQ1 and RQ2.

2.8. Formulating Study Framework

The literature review of validation theory and technology-enhanced assessment helped put together a study framework aimed to generate a validity argument about (not for) a Moodle-hosted English Language Exit Test. The framework can be considered a pragmatic approach to examine the research questions and achieve the study aims. To provide a validity argument, the study employed a validation framework, which is outlined in Appendix A (pp. 140-142). The AUA framework principles and concepts (Bachman, 2005; Bachman & Palmer, 2010) were used in this study as a pragmatic approach to articulate a specific evidence-based interpretive validity argument. The main drive for using the AUA framework in the study stems from its strengths as an argument model, as discussed in Section 2.2 of this chapter (pp. 9-12).

Applying the if-then rule, as Kane (2011) recommends to use for argument-based evidence-based validation approaches, research questions and validity claims or assumptions are given in Appendix A (pp. 140-142) along with the concepts of warrants, rebuttals, and backing evidence supporting or refuting claims (Bachman, 2005; Bachman & Palmer, 2010; Kane, 2011). Ideas expressed in the

AUA (Bachman, 2005; Bachman & Palmer, 2010) were incorporated where relevant to this study to formulate a validation research model. Research instruments to be employed for examining the validation research questions are described in the framework as a mixed-method research paradigm reflecting the need for multiple sources of evidence to support a conclusion (Kane, 1992). The framework reflects components of research methodology including the study design, participants, data collection procedures and instruments, and data analyses. All details related to research methodology are provided in Chapter 3.

The aim of using the validation framework in this study was to provide a validity argument based on examining the extent to which the Moodle-hosted test score-based decisions can be valid and reliable for the intended test score use, by empirically examining reliability and construct validity evidence. As mentioned in Appendix A (pp. 140-142), some of the testing mode effect technology-related construct-irrelevant factors addressed in the literature review were examined in the study (such as computer familiarity, test timing and length, and attitude). Reliability was included as an aspect to examine in the validation task using the study framework because, as discussed in Section 2.3 (pp. 12-17), a test needs to be proven to be reliable or consistent to claim that it is systematically testing what it is purported to measure (Brown, 2005). In other words, reliability and validity are intertwined because for a test to be valid, it must be reliable as its scores systematically reflect what is being tested and are not due to chance (Roever, 2006). Since the effect of the computerised testing mode may change the tested construct, it is essential to examine whether technology-related construct-irrelevant variance is reflected in the test scores (Fulcher, 1999). Hence, construct validity was another aspect examined in the study, especially that being highly reliable is not a sufficient support for a construct validity hypothesis (Roever, 2006). This means that although obtaining reliability estimates in the validation task informs us whether the test is systematically testing the construct being measured, it is still necessary to obtain evidence on whether potential sources of construct-irrelevant (unreliable) variance can jeopardise construct validity.

2.9. Conclusion

This study addresses the research issue of the testing mode effect by presenting a case study of administering and validating the Moodle-hosted test. The review of the literature aided the creation of a validation framework that guided the study to articulate the validity argument. By examining the research questions, the study will provide validity evidence from multiple sources using quantitative and qualitative data collection and analysis procedures in support of the validity argument, as detailed in Appendix A (pp. 140-142) and as described next in Chapter 3.

Chapter 3. Research Methodology

3.1. Introduction

The overall aim of this study is to provide a validity argument about using a Moodle-hosted test for its intended purpose by empirically examining reliability and construct validity evidence. As mentioned in Chapters 1 and 2, this study addresses the research problem of the testing mode effect by presenting a case study of administering and validating a Moodle-hosted test. This chapter describes the methodology followed to examine the research questions that are provided in Chapter 2. The first research question examines the extent to which the test scores can be reliable and valid indicators of the tested construct. The second research question investigates the extent to which technology-related construct-irrelevant factors affect the reliability and construct validity of the Moodle-hosted test. As outlined in Chapter 2, a validation framework (Appendix A, pp. 140-142) based on principles of the Assessment Use Argument framework (AUA) (Bachman, 2005; Bachman & Palmer, 2010) guided the design of the study to examine the research questions.

This chapter outlines and justifies the methods used to answer the research questions – that is, this chapter argues that the methods used to collect evidence for the validity argument are well-suited to the task. The chapter first begins by outlining the methodological approach that is then followed by the study design, participants, data collection instruments, and data analysis procedures.

3.2. Methodological Approach

An instrumental case study (Creswell, 2012) was used because it is well-suited to the descriptive, exploratory and explanatory purposes (Putney, 2010) of this validation study. The case study is intended to capture an in-depth picture of the research issue and compile a detailed description of the case study context (Creswell, 2012). The case study can be defined as a design logic for examining the issue (Putney, 2010). Using terms from Creswell (2012), the unit of analysis in this case study relates to the study of the test administration event or activity as a system or an entity bounded by a certain time and place. This is also compatible with the view that the case study can be a methodology for inquiry. Therefore, the case study here sufficiently captures the research issue from the test administration event and from the incurred participants' lived test experience.

From the pragmatic perspective, it was desirable to capture multiple information sources (Creswell, 2012) in order to fully examine the validity research framework aspects outlined in Chapter 2. This included establishing evidence for validity and reliability by using statistical techniques and by gathering contextual information from study participants on the impact of technology factors.

Therefore, both positivism and constructivism have set the guiding principles in this study with a mix of qualitative and quantitative procedures used in data collection and analysis. This has enabled the researcher to capture a deeper and wider picture of the case study (Pinto, 2010; Staller, 2010). Capturing such a picture of the case study has provided a richer more nuanced view of the case than would be available from quantitative data alone. The qualitative aspects shed light on complex human thought, attitude and purposeful behavior.

Applying two types of methods, that is, quantitative and qualitative, can widen the scope of examining the research problem and can gain us insights into human experience (Pinto, 2010). The quantitative positivist approach is justifiable here since there is a need to follow traditional testing and measurement practices when establishing statistical reliability evidence of test scores. Such statistical measures seek a *single truth* about the thing being measured. However, a positivist stance with its objectivist ontological perspective of the nature of reality overlooks the context within which human experiences take place. Taking a constructivist stance allows the researcher to account for the meanings research participants attach to their experiences (Creswell, 2014; Staller, 2010). In this case, it draws upon the perspectives of participants with respect to the influence that various technology factors can have in the testing process. When following a constructivist epistemological position, the complexity of the studied phenomena or issues can be captured much more deeply. The constructivist ontological perspective recognises that humans socially construct their reality (Staller, 2010). Therefore, including a qualitative constructivist element in this study is justified.

These being the rationales for following both approaches, the pragmatic perspective of using a mixed-method research design captures what works best to fully address the research issue within a specific context (Pinto, 2010). Following a pure paradigm, either positivism or constructivism, cannot gain a detailed picture of the research issue in the specified context. Rather than focusing on a singular paradigm (such as positivism, post-positivism, or constructivism) and applying its methods in the study, the pragmatist epistemological stance uses methods from compatible paradigms to fit the purpose of the inquiry. The use of this mixed design provides a greater variety of data sources and analysis methods. As Pinto (2010) emphasises, qualitative data can be transformed or “quantitized” by converting them into numeric codes and conducting statistical analyses. Quantitative data can also be “qualitized” by converting them into textual data such as narratives and analysing them qualitatively (p. 814). Quantitative and qualitative data types gained from the mixed method approach followed in this study have lent themselves to quantitative and qualitative data analysis techniques where appropriate, which has provided a rich case study to present. These psychometric and non-psychometric analysis techniques were used as methods to establish reliability and construct validity evidence in the study.

The researcher’s role has been to compile a rich descriptive case study through interpreting the study results and being an active participant in the collection of data from participants and by recording observations and field notes (Heigham & Croker, 2009). To ensure the “Credibility” in qualitative research and “Validity” in quantitative research, Brown (2008) recommends drawing from multiple sources of data or triangulation in order to reduce sources of researcher bias. In this study, data were triangulated using participant, methodological, and theory triangulation options, as will be described in this chapter. When establishing validity evidence through the mixed method approach to justifiably support test use (Kane, 2012), the evidence can be both for and/or against a validity position. This means that while positive evidence can support warrants of reliability and construct validity claims in the validity argument (Chapter 2, Section 2.8, pp. 28-29), negative evidence (Chapelle, Jamieson, & Hegelheimer, 2003; Wang, Choi, Schmidgall, & Bachman, 2012) pointing to technology-related construct-irrelevant factors being sources of the testing mode effect can still be the rebuttals to these claims. By addressing the research issue using such types of evidence established via these mixed methods, triangulation of the findings can be enhanced.

When reporting and interpreting the study findings, a full rich description of the case study should be given, accounting for the study design logic and the research process that was followed; the themes or issues revealed by the study; and the “lessons learned” (Creswell, 2012, p. 99). In articulating a validity argument for stakeholders, a “thick description” of the context and the study design procedures as well as the meaningfulness of the study results can also enhance its “Transferability” in terms of qualitative research and “Generalizability” as termed in quantitative research. This makes the study findings applicable to a wider range of contexts (Brown, 2008, pp. 294-295). It is acknowledged that there are limitations to the applicability of findings from in-depth case studies given they are firmly rooted in context (see Chapter 7, Section 7.5, pp. 121-125 for study limitations). A significant by-product of this research is to provide improved guidelines for creating, developing, implementing, and researching large-scale high-stakes Moodle-hosted exams that yield more reliable and valid score-based decisions. Such guidelines will have implications for practitioners and researchers working in similar testing contexts. Following the described methodological approach, the validation research framework of this study will contribute to the validation theory literature as a pragmatic methodological tool to conduct validation research.

3.3. Study Design

Figure 3.1 illustrates the process of developing the Moodle-hosted test in the pilot study phase moving on to the main study phase events in light of the validation framework. The diagram shows the study phases and demonstrates the process used for test development and how data collection from the pilot study informed the main study.

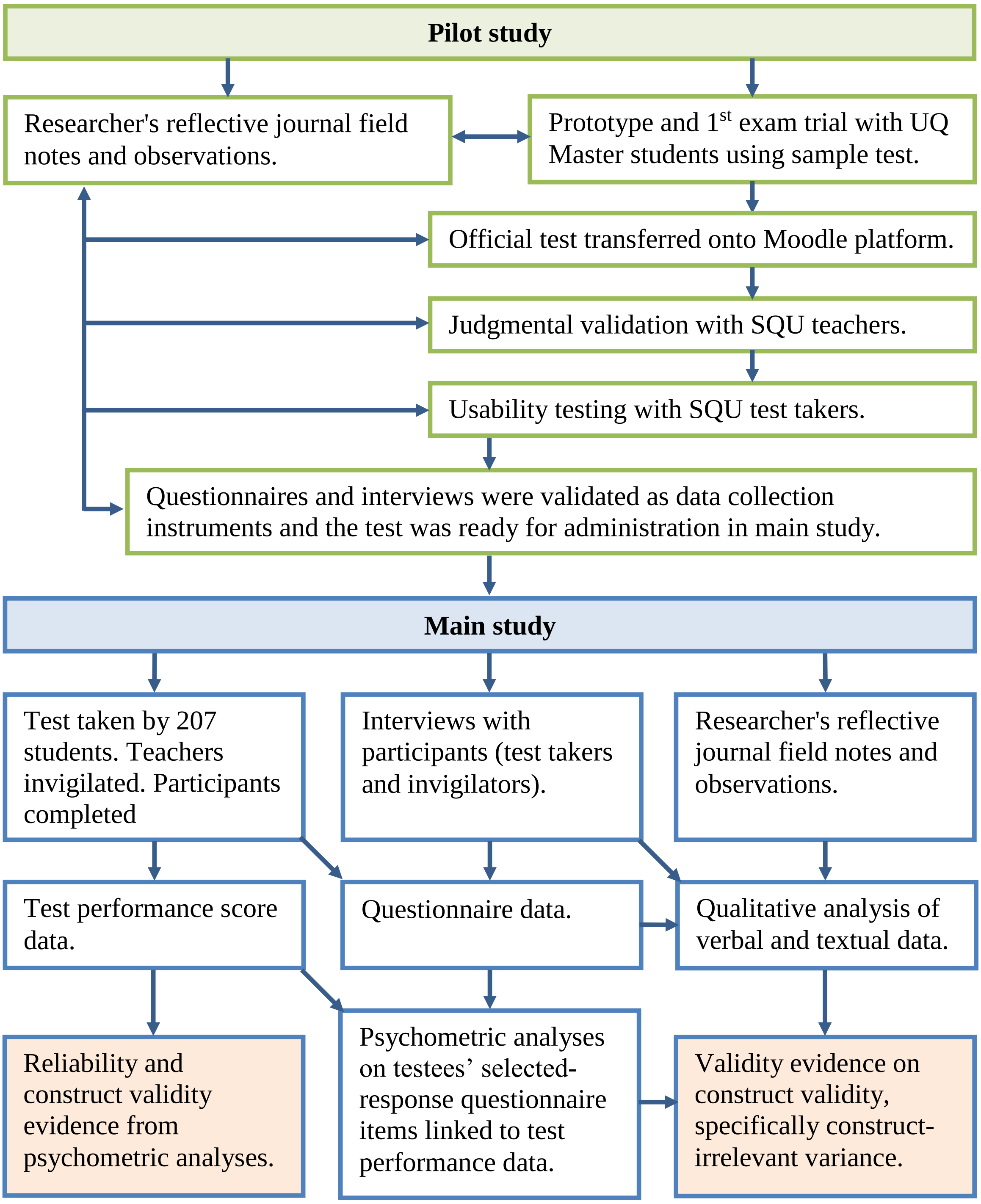

Pilot study
Researcher's reflective journal field notes and observations.
Prototype and 1st exam trial with UQ Master students using sample test.
Official test transferred onto Moodle platform.
Judgmental validation with SQU teachers.
Usability testing with SQU test takers.
Questionnaires and interviews were validated as data collection instruments and the test was ready for administration in main study.
Main study
Test taken by 207 students. Teachers invigilated. Participants completed
Interviews with participants (test takers and invigilators).
Researcher's reflective journal field notes and observations.
Test performance score data.
Questionnaire data.
Qualitative analysis of verbal and textual data.
Reliability and construct validity evidence from psychometric analyses.
Psychometric analyses on testees' selected-response questionnaire items linked to test performance data.
Validity evidence on construct validity, specifically construct-irrelevant variance.

Figure 3.1. The process of developing the Moodle-hosted test validation framework.

As displayed in Figure 3.1, in preparing for the main study through the pilot study, a prototype using a sample paper-based Exit Test was first transferred onto Moodle. This prototype was trialled in October 2014 with 23 volunteering students in a Master program at The University of Queensland (UQ) in Australia. The pilot used a USB based e-exam system capable of running

Moodle without a network connection on students' laptops. The system was developed as part of an Australian Government funded project (Hillier et al, 2015, software available from transformingassessment.com). It was decided to use this e-exam since it was not possible to grant these UQ students access to the Moodle-hosted test on the SQU server because they were not SQU enrolled students. The purpose of this exam trial was to examine potential technology-related construct-irrelevant factors that might be present in all aspects of the Moodle-hosted test administration. Participants sat the test and then provided feedback on questionnaires (Appendix C, pp. 162-163). This exam trial confirmed that technical issues (such as pictures, listening audio files, headphones, and other equipment) can affect test performance if such problems creep into the test administration setting.

Later on, in March and April of 2015, an official paper-based Exit Test was transferred onto the Moodle platform at the study context, the LC at SQU in Oman. The researcher developed a technology-enhanced Moodle-hosted interface with special technical features useful for online language testing purposes, as will be described in this chapter. The development of this interface went into stages and involved participants from the study context to gain their feedback on the usability of the interface. In examining the usability of the interface, a group of four language teachers from the study context participated in a judgmental validation session. These teachers trialled the test and provided their feedback on a questionnaire (Appendix D, pp. 164-165) and a focus group semi-structured interview (Appendix E, p. 166). Utilizing the feedback received in this judgmental validation session, modifications were made to the interface to accommodate the participating teachers' suggestions and concerns. Usability testing sessions were then conducted with 25 test takers from the study context. These examinees sat the test, filled in questionnaires (Appendix F, pp. 167-176), and took part in focus group semi-structured interviews (Appendix G, p. 177). These examinees' feedback was also useful in improving the interface and validating the research instruments.

In this pilot study, a problem resolution approach (Fulcher, 2003) was followed to tackle issues with the technology-enhanced Moodle-hosted testing interface and prepare it for formal use in the main study. Questionnaires and interviews were validated (trialled and refined) and prepared for formal use as data collection instruments in the main study to get participants' retrospective accounts of their test experience. Existing questions in these instruments were either edited or deleted and new questions to address potential issues were added. Field notes and observations recorded by the researcher on reflective journals aided to take decisions and actions during this process.

In the main study, an official version of a Moodle-hosted test was administered to a sample of 207 volunteering examinees from the study context, the LC at SQU in Oman. In each testing session the Moodle-hosted test was administered in a computer laboratory to a sample of examinees supervised by an invigilator. Participants' feedback (including examinees and invigilators) was elicited using post-test questionnaires (Appendices H and I, pp. 178-184). A sample of test takers and invigilators later took part in respective semi-structured interviews (Appendices J and K, pp. 185-186) to talk about the testing experience. The single test administration was done in different testing sessions based on logistical arrangements regarding lab bookings and participants' commitments, where *single* here means that every examinee took the test only once. Appendix N (pp. 191-192) mentioned in Chapter 5 (RQ2 results) shows data collected on the venue (location), level, section, course code and discipline area. Table N1 in Appendix N (p. 191) gives a summary of the data collected. The data obtained from the main study participants and from the researcher's observations and field notes were intended to be analysed and interpreted to examine the research questions (Appendix A, pp. 140-142). However, due to time limitations only the test score data and test takers' questionnaires were used to contribute to the validity argument. The questionnaires were selected for further analysis because they provided collective insights about the test-taking experience using feedback from 89.9% (n = 186) of the test takers and they were the closest in time to the experience of the test event.

As laid out in Figure 3.1, in light of the validation framework, the study was designed to go through a preparatory pilot study phase before the Moodle-hosted test was administered in the main study and data were collected from participants. We should emphasize here that the informal pilot study procedures should not be considered part of the body of evidence gathered in the main study to address the research questions in light of the validation research framework explained in Chapter 2. However, the pilot study is reflected here as an essential component in the story of the research process used in setting the ground for the main study and is itself reflective of good test development procedure. This is because it is a basic principle of good language testing practice to cater for all phases of the design and development of a language test (O'Sullivan, 2012). The following sections outline the study participants, data collection instruments, and data analysis procedures, focusing mostly on the main study. Appendix M (pp. 188-190) provides further detailed information about the pilot study participants and data analysis procedures.

3.4. Participants

Table 3.1 summarises the main study participating sample. The sample includes both student and teacher participants.

Table 3.1. *Main Study Participating Sample*

	Test Administration			Questionnaires[a]		Interviews[b]	
Student EFP level	No. of Classes	Test Takers	Invigilators	Test Takers	Invigilators	Test Takers	Invigilators
4	4	52	4	51	3	2	2
5	4	59	3	50	3	5	3
6	1	23	1	19	1	1	0
6	5	73	4	66	3	6	2
Totals	14	207	12	186	10	14	7

Notes. [a]Questionnaires returned from test takers and invigilators. [b]Semi-structured interviews conducted with test takers and invigilators.

3.4.1. Students.

As seen in Table 3.1, a total of 207 students were recruited. All 207 volunteer test takers sat the Moodle-hosted test, after which the majority (89.9%; *n* =186) returned follow-up questionnaires. Table 3.2 gives details on the examinees' disciplinary areas and courses or levels.

Table 3.2. *Examinees' Disciplinary Areas and Courses/Levels*

Level	Course Code	GEN	COM	SCI	MED/ NUR	ENG	Law	AGR	EEAL	Totals By Level
4	340	52								52
5	450		16	19			11	13		59
6	560				23					96
6	604		9	13		23		19	9	
Totals By Discipline		52	25	32	23	23	11	32	9	207

Notes. GEN = General English; COM = Commerce; SCI = Sciences; MED/NURS = Medicine/Nursing; ENG = Engineering; AGR = Agriculture; EEAL = Education, English specialists, Arts, and Law.

As seen in Table 3.2, student participants were drawn from Levels 4, 5 and 6. A quarter came from the general language preparation program in Level 4 with the remainder coming from discipline specific groups in Levels 5 (Intermediate English) and 6 (Advanced English). Overall, the sample was representative of the levels eligible to sit the Exit Test with the majority enrolled in Level 6, which is normally the case in the regular paper-based Exit Test administrations. A lesser number of volunteers participated in interviews but were still representative of the three levels tested.

The gender and age ratios in the test population could not be controlled given the voluntary nature of participation. Test takers' detailed profiles were obtained via a demographics section on the post-test questionnaires. These profiles are related to gender, current course of study or overall levels of English, and self-assessment of the levels of familiarity with Moodle testing and computer literacy. These detailed profiles of the study participants are reported as results of the questionnaire analysis in Chapter 5 (RQ2 results).

3.4.2. Language teachers.

Language teachers from the SQU English Language Foundation Program were invited to participate via email. Those that volunteered invigilated their students' Moodle-hosted test sessions in the main study. The teacher participants then provided an account of their experience via a follow-up questionnaire (Appendix I, pp. 183-184) and semi-structured interviews (Appendix K, p. 186). Invigilation instructions (Appendix L, p. 187) were given to these teachers prior to commencing the testing session. Each testing session was supposed to be supervised by an invigilating class teacher as well as the researcher. However, as reported in Table 3.1, two testing sessions in Levels 5 and 6 were supervised by the researcher alone. There were a total of 12 invigilators with ten returning a completed questionnaire. Seven of the ten invigilators were interviewed individually by the researcher. The age and gender ratios of teacher participants could not be controlled as well given the voluntary nature of participation.

Ethical clearance was obtained from the School of Education at UQ to conduct the study. Ethical clearance was also obtained from the main study context at the LC, SQU in Oman. The researcher provided each participant an Information Sheet and Consent Form to sign prior to their participation in the study. Following ethical procedures, students were given the chance to leave the testing session (opt out of the study). Refer to Appendix B (pp. 143-161) on ethical considerations and relevant sheets.

3.5. Data Collection Instruments

Data collection instruments were first developed and trialled in the pilot study, and refined for the main study. The instruments include 1) a Moodle-hosted test; 2) questionnaires; 3) focus group semi-structured interviews; and 4) the researcher's observations and field notes recorded in reflective journals. In the main study, the data collected from the instruments above contain:

1) the Moodle-hosted test score data comprised of test takers' scores on the overall test, subtests, and individual items; and the item statistics report on Moodle;
2) retrospective verbal protocols of participants (test takers and invigilating teachers) from questionnaires and interviews; and
3) the researchers' field notes and observations reported in her reflective journals.

The following subsections describe the data collection instruments in detail.

3.5.1. Online test tool.

A sample Exit Test given to the researcher by the Assessment Unit (AU) from the study context at the LC was initially used as a prototype in the first trial of the Moodle-hosted online test tool at UQ. An official Exit Test from the same source was then administered in the pilot study. The user interface and system features were enhanced via the prototype and pilot testing stages to better suit the needs of language testing for use in the main study. These improvements are outlined later in this section.

The technology-enhanced web-based test in this study utilised the objective part of a working paper-based version of the Exit Test. The paper-based test was transferred onto the Moodle CMS version 1.9 platform. The test construct draws on the intended testable learning outcomes of English language skills (proficiency) reflected in the *Foundation programme English language curriculum document* (2012-2013). The Exit Test is a criterion-referenced test (Brown, 2005; Brown & Hudson, 2002) that is used for the purpose of classifying test takers into two decision categories as pass or fail decisions based on a cut-point score. Those who pass the test can exit the English Foundation Program and commence college credit academic courses. Those who fail the test will remain in the English Foundation Programme since they are deemed to require further English language support.

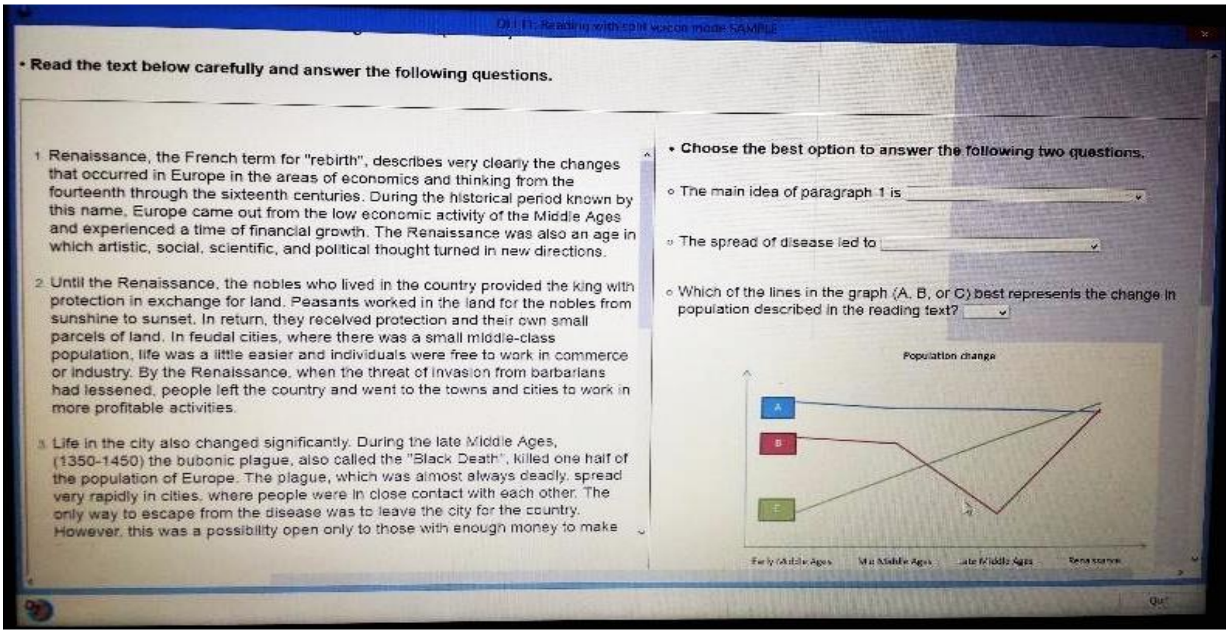
The Exit Test is described as a large-scale test because a large test population (hundreds of candidates) may sit the test in a single administration. The test is also high-stakes in the sense that students' study paths at university will be determined based on their results on this test. The paper-based Exit Test is 120 minutes in duration. It comprises objective sub-tests for reading and language use (60 minutes), objective listening subtests (30 minutes) and a writing test (30 minutes). The Moodle-hosted test in this study utilised only the objective parts for a duration of 90 minutes and contained 60 items weighted at one point each.

Transferring the test onto the Moodle platform was informed by guidelines for good practice for computer-based interface design suggested by Fulcher (2003). These include the basic principle of ensuring that the testing mode effect is put under control in order to avoid introducing construct-irrelevant score variance that has been considered as a threat to validity of score-based interpretations. The technology-enhanced Moodle-hosted interface was designed to include technical features useful for online language testing purposes. These features include enhanced test security settings aided by Safe Exam Browser; an embedded MP3 player for listening; and a split screen mode for reading tests. All of these features were intended to serve the purpose of limiting or eliminating the testing mode effect and were informed by the pilot study. One example of this is the inclusion of the split screen mode for reading tests after it was suggested to the researcher by teacher participants during the judgmental validation session.

The features of the Moodle-hosted testing interface used in this study are described in Al Nadabi (2015). The most important feature of the interface is applying enhanced test security settings. The standard settings on the Moodle platform allow designers to create password-protected tests. This limits access to individuals or classes with knowledge of a common password (for example; the password can be displayed at the front of the room once all candidates are seated in the exam room). These tests can also be timed and a count-down timer can be displayed to each examinee. The number of attempts allowed for each test can also be set.

Heightened test security can be accomplished by using Moodle in conjunction with a security browser called Safe Exam Browser (SEB, version 2.0.3). This browser is an open source application that displays online exams in a full screen mode and allows access to specified computer functions and web resources during these exams. SEB prevents the use of shortcuts and functions such as right-click to copy or print screen with task manager and program switching disabled to prevent cheating during the exam. See *Safe exam browser* (2015) for details on this browser. The traditional approach of supervising the exam to prevent cheating is still recommended when using this type of computerised exams. As Coy (2013) and Myrick (2010) recommend, for high-stakes tests, such security measures provided by Moodle settings should be combined with test proctoring or invigilation to achieve high security (Coy, 2013). Refer to Figure 3.2 for a snapshot of the use of SEB on a sample e-exam.

The use of enhanced test security settings aided by SEB can limit the effect of construct-irrelevant technology-relevant sources of measurement error, leading to a better testing experience where examinees' cheating behavior is monitored much more closely.



• Read the text below carefully and answer the following questions.

1 Renaissance, the French term for "rebirth", describes very clearly the changes that occurred in Europe in the areas of economics and thinking from the fourteenth through the sixteenth centuries. During the historical period known by this name, Europe came out from the low economic activity of the Middle Ages and experienced a time of financial growth. The Renaissance was also an age in which artistic, social, scientific, and political thought turned in new directions.

2 Until the Renaissance, the nobles who lived in the country provided the king with protection in exchange for land. Peasants worked in the land for the nobles from sunshine to sunset. In return, they received protection and their own small parcels of land. In feudal cities, where there was a small middle-class population, life was a little easier and individuals were free to work in commerce or industry. By the Renaissance, when the threat of invasion from barbarians had lessened, people left the country and went to the towns and cities to work in more profitable activities.

3 Life in the city also changed significantly. During the late Middle Ages, (1350-1450) the bubonic plague, also called the "Black Death", killed one half of the population of Europe. The plague, which was almost always deadly, spread very rapidly in cities, where people were in close contact with each other. The only way to escape from the disease was to leave the city for the country. However, this was a possibility open only to those with enough money to make

• Choose the best option to answer the following two questions.

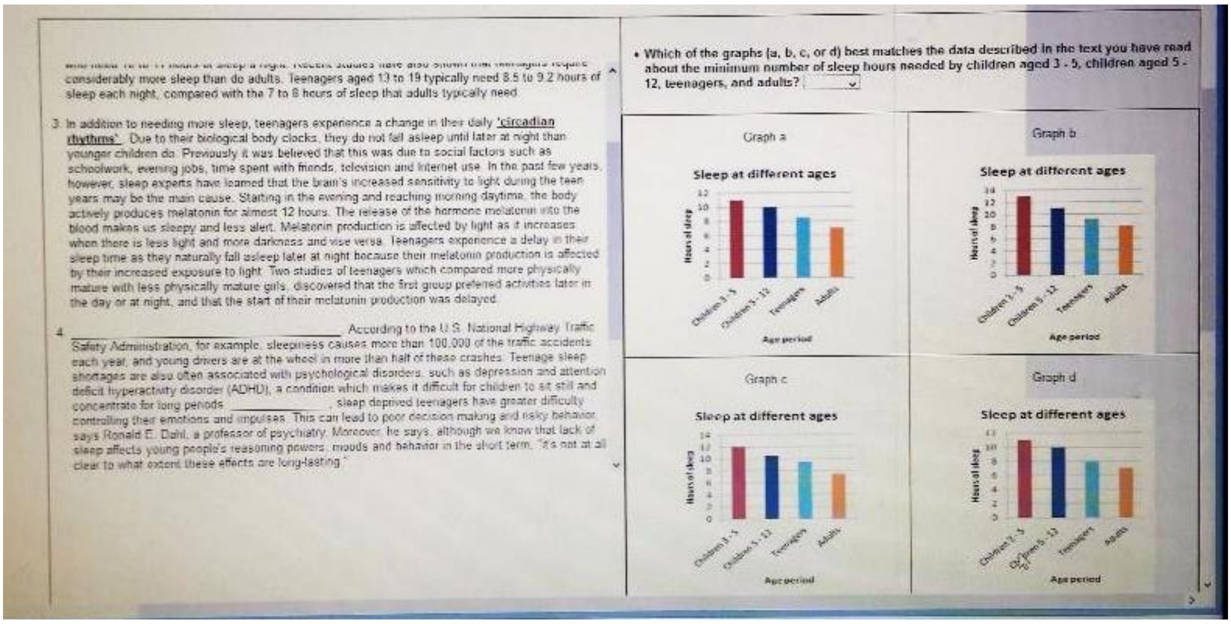
○ The main idea of paragraph 1 is

○ The spread of disease led to

○ Which of the lines in the graph (A, B, or C) best represents the change in population described in the reading text?

Figure 3.2. E-exam accessed through Safe Exam Browser in a full screen mode.

A split screen feature allows examinees to simultaneously access reading paragraphs on the left side of the screen and the relevant questions on the screen right side. The two independently scrolling content regions allow for two separate user selected sections of a larger amount of content to be visible on the one screen. Figure 3.3 shows the reading test split screen mode feature.



who need 10 to 11 hours of sleep a night. Recent studies have also shown that teenagers require considerably more sleep than do adults. Teenagers aged 13 to 19 typically need 8.5 to 9.2 hours of sleep each night, compared with the 7 to 8 hours of sleep that adults typically need.

3 In addition to needing more sleep, teenagers experience a change in their daily 'circadian rhythms'. Due to their biological body clocks, they do not fall asleep until later at night than younger children do. Previously it was believed that this was due to social factors such as schoolwork, evening jobs, time spent with friends, television and Internet use. In the past few years, however, sleep experts have learned that the brain's increased sensitivity to light during the teen years may be the main cause. Starting in the evening and reaching morning daytime, the body actively produces melatonin for almost 12 hours. The release of the hormone melatonin into the blood makes us sleepy and less alert. Melatonin production is affected by light as it increases when there is less light and more darkness and vise versa. Teenagers experience a delay in their sleep time as they naturally fall asleep later at night because their melatonin production is affected by their increased exposure to light. Two studies of teenagers which compared more physically mature with less physically mature girls, discovered that the first group preferred activities later in the day or at night, and that the start of their melatonin production was delayed.

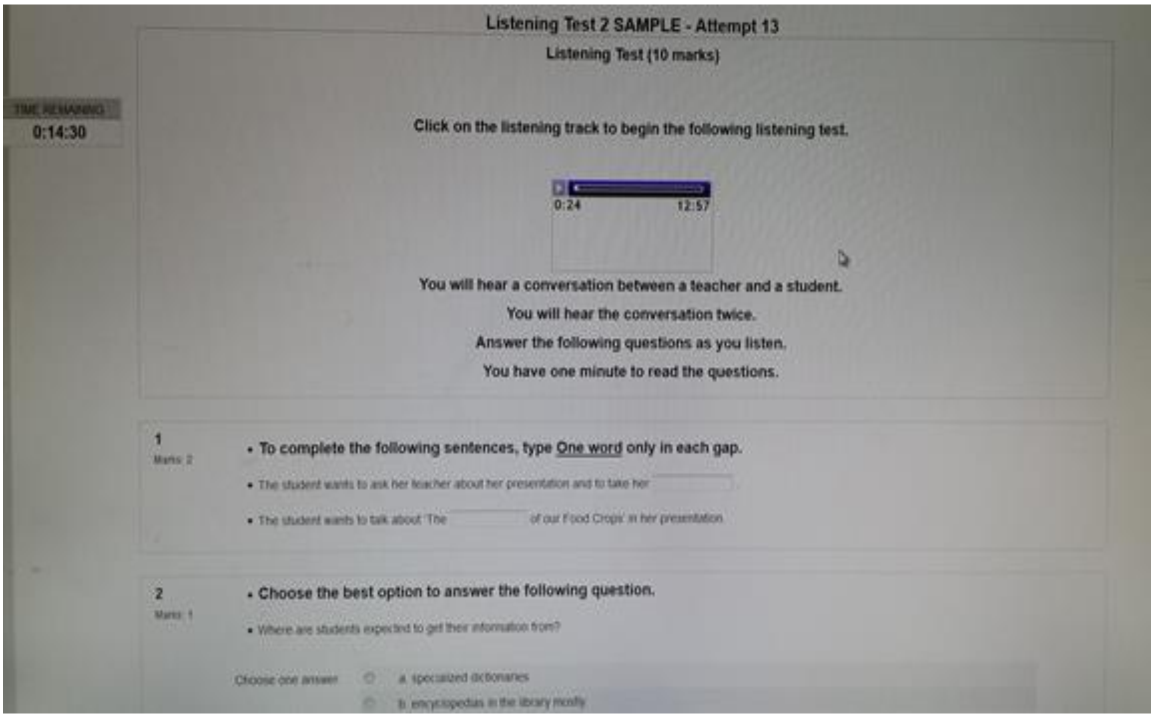
4 ____________________ According to the U.S. National Highway Traffic Safety Administration, for example, sleepiness causes more than 100,000 of the traffic accidents each year, and young drivers are at the wheel in more than half of these crashes. Teenage sleep shortages are also often associated with psychological disorders, such as depression and attention deficit hyperactivity disorder (ADHD), a condition which makes it difficult for children to sit still and concentrate for long periods. ____________, sleep deprived teenagers have greater difficulty controlling their emotions and impulses. This can lead to poor decision making and risky behavior says Ronald E. Dahl, a professor of psychiatry. Moreover, he says, although we know that lack of sleep affects young people's reasoning powers, moods and behavior in the short term, "it's not at all clear to what extent these effects are long-lasting."

• Which of the graphs (a, b, c, or d) best matches the data described in the text you have read about the minimum number of sleep hours needed by children aged 3 - 5, children aged 5 - 12, teenagers, and adults?

Figure 3.3. A snapshot of the split screen mode for reading tests.

The split screen mode used for the reading test is contrary to many paper-based exams in which examinees have to flip pages to connect the test questions and the reading. Sweller's (1994) cognitive load theory (Ayres & Sweller, 2005) supports the argument that presenting the reading test materials in a split screen interface will aid concentration during exams. This can reduce split attention and cognitive load demands caused by presenting all material to examinees on the one screen (i.e., reading passage and subsequent questions). Examinees in the pilot study expressed their satisfaction with the split screen mode for the reading tests pointing to a much more positive testing experience than for similar paper-based exams.

The use of Matbury's MP3 player for listening tests (Matbury, 2010a, 2010b) helped to ensure that all examinees are exposed to the listening materials in a consistent way. Figure 3.4 displays Matbury's MP3 player for listening tests.



Listening Test 2 SAMPLE - Attempt 13

Listening Test (10 marks)

TIME REMAINING
0:14:30

Click on the listening track to begin the following listening test.

0:24 12:57

You will hear a conversation between a teacher and a student.

You will hear the conversation twice.

Answer the following questions as you listen.

You have one minute to read the questions.

1
Marks: 2

- To complete the following sentences, type One word only in each gap.
- The student wants to ask her teacher about her presentation and to take her
- The student wants to talk about 'The of our Food Crops' in her presentation

2
Marks: 1

- Choose the best option to answer the following question.
- Where are students expected to get their information from?

Choose one answer ○ a. specialized dictionaries
○ b. encyclopedias in the library mostly

Figure 3.4. A snapshot of the embedded Matbury's MP3 player for listening tests.

To meet test fairness principles (Chapelle & Douglas, 2006; Fulcher, 2003; Kunnan, 2004) in this context means that examinees are exposed to the listening material the same number of times (such as once or twice only) and are each able to control playback in the same manner (i.e., not permitted to pause, stop, or use rewind or fast forward). The Matbury's MP3 player allows the test administrator to control the mode of playback. Of course, if it turns out that major issues in the use of such a player make it difficult to meet such test fairness goals, reliability and validity may still be questioned.

Prior to every testing session the researcher prepared each laboratory computer and set up the headphones for the listening component. Students entered the room where computers were already switched on with the SEB window open ready for them to start the test. The researcher explained the testing process and the test environment by conducting a walk-through demonstration of the login-process via the screen projector.

The Moodle-hosted test enhanced by the three main features described here was used as a testing instrument to generate score data. This testing instrument was also the springboard for feedback on the study participants' online testing lived experience obtained through questionnaires and interviews. Thus, the single administration of this test formed the basis of the case study.

3.5.2. Questionnaires.

The development of questionnaires followed a staged process. First, a questionnaire (Appendix C, pp. 162-163) was used to gather feedback from the participants in the first exam trial at UQ. The feedback obtained from participants helped the researcher construct questionnaires for the second stage of the pilot including a questionnaire for the judgmental validation participants (Appendix D, pp. 164-165) and a questionnaire for the usability study (Appendix F, pp. 167-176). Changes to the questionnaires resulting from the pilot included new items added to cover features incorporated into the online testing interface such as the split screen mode and to address issues such as staring at the computer screen for a long time and its effects. Items were also edited for clarity, deleted or combined with other items to avoid redundancy.

The final version of the questionnaire used for test takers (Appendix H, pp. 178-182) in the main study is made up of 36 items including four background information items and a combination of five open-ended items and 27 five-point Likert scale items that asked respondents to indicate their level of agreement with each statement (5=strongly agree; 4=agree; 3=neutral; 2=disagree; and 1=strongly disagree). Questions sought opinions from test takers on a number of aspects such as the user interface (e.g., background colours, navigation, and clarity of font and pictures), test system features (e.g., split screen mode, listening test sound and headphones quality, and instant test feedback/results), administration procedures (e.g., login process and test procedures and instructions), test takers' familiarity and experience with computers and Moodle tests, and preferences between paper and computerised tests. The questionnaires were provided in a bilingual format that included examinees' mother tongue language, Arabic, as well as English. This was to allow participants to express themselves freely so that the researcher would have easier access to the meanings they attach to their experience (Sunuodula, Feng, & Adamson, 2015). The reliability statistics analysis on this questionnaire data showed a Cronbach's Alpha of .692, i.e. approximately .70, which is just at the lower edge of the acceptable range of values for Cronbach's Alpha between .70 and .80 (Pallant, 2007). It should be noted here that, in order to maintain the reliability of the questionnaire, the total number of 186 questionnaire respondents was reduced to 174 test-takers after eliminating a number of cases (test-takers) from the sample due to incomplete responses. One respondent did not do the listening section of the test, while eleven other respondents did not complete a large number of questionnaire items.

The invigilators filled in a ten-item questionnaire (Appendix I, pp. 183-184) that consisted of eight open-ended constructed-response items in addition to the two background information items. Questions covered topics such as overall experience with the Moodle-hosted test and its invigilation, practicality of running the test, efficiency of computer laboratories, technical issues faced during the testing session, use of Moodle for official exams, and Moodle automatic test marking.

3.5.3. Interviews.

Audio-recorded interviews and focus groups were also essential data collection instruments in the study with a phased approach used to improve interview protocols. The pilot study interviews assisted the researcher in developing interview questions for the main study to elicit feedback from test taker and invigilator participants. The interview protocol used with four judgmental validation participants to discuss their insights about their Moodle-hosted testing experience appears in Appendix E (p. 166). The interview protocol used for 25 test takers in the usability study sessions that explored their Moodle-hosted test-taking experience appears in Appendix G (p. 177).

In the main study, 14 audio-recorded semi-structured interviews were conducted with test takers in Arabic. Each interview lasted for about 30 minutes and covered issues such as examinees' test-taking experience, use of Moodle for official exams to take decisions about English language proficiency, preference of testing mode (i.e., paper-based or Moodle-hosted), technical issues affecting test performance, and Moodle scoring and feedback functionality. The interview protocol is in Appendix J (p. 185).

To capture the invigilators' detailed feedback on the test experience, seven audio-recorded semi-structured interviews were also conducted after the testing event. Topics discussed included overall experience with the Moodle-hosted test, testing invigilation experience, technical issues during the testing session, efficiency of computer laboratories, use of Moodle for official exams, and Moodle automatic test marking. Appendix K (p. 186) shows sample interview questions.

3.5.4. Reflective journals.

The researcher kept reflective journals for the duration of the research process. These provided evidence of learning, aided in the problem resolution approach and helped the researcher prepare for the main study. The researcher recorded observations and took field notes during the pilot study.

The researcher consulted these records to assist in resolving issues with the testing interface and in updating the other research instruments including questionnaires and interviews. In the main study, these records assisted the researcher in coding, organizing, analysing, triangulating, and interpreting the case study data. This type of narrative inquiry is advocated as an essential research method in qualitative research (Pinot, 2010). The data collected via the reflective journals were aimed to address the research questions by acting as empirical evidence (Staller, 2010). However, just like some of the data collected from questionnaires and interviews, these journals were not incorporated as a source of evidence in the validity argument due to logistical and time constraints.

The data that were obtained from the main study by administering the test and examinees' questionnaires were then analysed and interpreted to discuss the research questions in light of the validity argument framework (Chapter 2). The procedures that were used for data analyses are discussed in the next section.

3.6. Data Analyses

The data gathered from the main study using the data collection instruments were analysed to achieve the overall study aim. Both psychometric and non-psychometric approaches were utilized to address the research questions. As mentioned in Section 2.7 (pp. 27-28), RQ1 examines the extent to which the Moodle-hosted test scores can be reliable and valid indicators of the tested construct. RQ2 investigates the extent to which technology-related construct-irrelevant variance factors associated with the testing mode effect can interfere with test results. By examining reliability and construct validity aspects through the RQ1 and RQ2 data analyses, separate pieces of evidence are combined to achieve the overall study aim – that is providing a validity argument about the extent to which the Moodle-hosted test score-based decisions can be reliable and valid for the intended test score use.

3.6.1. RQ1: Statistical test item analyses.

The test takers' overall scores on the Moodle-hosted test, subtests scores, responses to individual items, and the item statistics report on Moodle formed the score data. This was statistically analysed in order to establish evidence for a response to RQ1 focusing on reliability.

The term *reliability* is used here to refer to the reproducibility of the test scores. This is a quality that assures test users that the test results can consistently be replicable if test takers were to take the same test again in similar conditions (Fulcher & Davidson, 2007). To establish internal reliability of the Moodle-hosted test, Rasch analysis was conducted on the score data obtained from the single test administration. The Rasch Item Response Theory (IRT) measurement model from Modern Test

Theory (MTT) can yield information on reliability (Green, 2013) as well as construct validity (Baghaei, 2008). Therefore, Winsteps software version 3.91.0 (http://www.winsteps.com/winsteps.htm) was used to do Rasch analysis to obtain information on reliability and to link item difficulty to person ability. As stated by Bachman (2004), the two factors that indicate a test taker's performance based on IRT are: (1) item characteristics, and (2) the test taker's ability level on an "underlying ('latent') trait" (p. 141). Rasch analysis on Winsteps (Linacre, 2012) can help test developers match person ability estimates with item difficulty measures (Al Naddabi, 2012; Bond & Fox, 2007; Green, 2013). Running the Rasch analysis produces outputs in the form of estimates of item reliability and person reliability. Other Rasch outputs also include item and person variable maps that graphically represent the match between item difficulty and person ability. Other useful outputs include measures of item difficulty and person ability in the form of fit indices as well as their associated Standard Error of Measurement (SEM). Of particular importance here is the measurement error. As emphasised by Brown (2005), a low SEM value showing the amount of error in the measure is desirable because it indicates more consistent or reliable test results. Contrary to Classical Test Theory that reports an average SEM for the test taker sample as a Test Reliability Index, Rasch on Winsteps reports a SEM for every item measure since every test score or measure has a different SEM (Linacre, 2014). By knowing the measurement error associated with each item, we know how much confidence we can place on test scores obtained from these items. Large errors are not acceptable for claiming reliability as increased error reduces reliability (Castle, 2016).

Rasch analysis was used as a validation tool in the present study because it is useful for evaluating both reliability and construct validity. Besides obtaining reliability estimates (Green, 2013), Rasch analysis outputs give information on construct-irrelevance and construct under-representation (Baghaei, 2008) identified by Messick (1989) as two construct validity issues. Construct-irrelevance can be identified by obtaining item and person fit indices through Rasch analysis (Wright & Stone, 1999). Items that do not fit the Rasch model can be considered as not consistently discriminating between test takers. Such items do not contribute to measuring the single measurement dimension or target construct intended by the construct theory (McNamara, 1990). Therefore, such items point to construct-irrelevant variance or multidimensionality, indicating the need to modify or discard the items (Baghaei, 2008). Construct-irrelevant variance has two distinct forms: construct-irrelevant easiness and construct-irrelevant difficulty. By including tasks or items that make the construct difficult, test takers' scores may become invalidly low. Scores of other test takers may be invalidly high as a result of the presence of construct-irrelevant easy items that test-wise examinees can easily answer (Baghaei, 2008). Construct under-representation can be identified through finding gaps on the Wright item-person map where a mismatch exists between item difficulty and person ability. Finding such gaps indicates that test items are not targeting test taker ability, which means that these items might be insufficient to measure persons across ability ranges or that items of better quality might be needed (Baghaei, 2008; Bond, 2003).

Employing the Rasch analysis for establishing reliability and construct validity in this study is in line with the methods followed by other studies that reported Rasch analysis findings on reliability and construct validity (e.g., Akiyama, 2001; Aryadoust & Goh, 2009; McNamara, 1990, 1991). For example, in Aryadoust and Goh’s (2009) study, construct-irrelevance, was identified through Rasch analysis fit statistics results. Instances of construct under-representation were also identified through the Wright item-person map. In sum, following a similar approach, the present study utilised Rasch analysis since it is useful in the examination of reliability and construct validity, making use of person and item reliability estimates, as well as the variable maps outputs and fit statistics tables to identify construct-irrelevance and construct under-representation. Rasch analysis also provided evidence on reliability by determining discrimination of test items and their contribution to reliability. Both reliability and construct validity are essential components in the validity argument framework (Section 2.8, pp. 28-29 and Appendix A, pp. 140-142).

To establish evidence in an answer to RQ1, the Rasch statistical test score data analysis approach described above can indicate the degree to which score-based decisions can be reliable and valid. As discussed in the validity argument framework in Section 2.8 (pp. 28-29), the assumption or claim here is that statistical evidence of high reliability estimates, fit and highly discriminating items, and acceptable low SEM will be a warrant that the Moodle-hosted test scores will be reliable and valid indicators of the tested construct. The results of this analysis carried out on the Moodle-hosted test score data to answer RQ1 are presented in Chapter 4.

3.6.2. RQ2: Descriptive statistics.

RQ2 addresses the extent to which technology-related construct-irrelevant variance factors can affect reliability and construct validity. Quantitative and qualitative data analyses were carried out to arrive at this type of evidence to report in the validity argument. These data analyses procedures are laid out in the validation research framework in Chapter 2.

First of all, SPSS was used to obtain descriptive statistics, frequencies in particular, to report the frequencies of responses to questionnaire items (Brown, 2005; Green, 2013) such as Likert scale items. Further analysis following advice by Green (2013) involved the comparison of selected questionnaire items (test taker’s perception of a given technology issue) to respondents’ test performance data (test mean scores). This linking analysis allowed the researcher to link test takers' perceptions of their testing experience and any reported technology-related issues to their test performance. For instance, a questionnaire item "The headphones worked properly during the exam" was linked to respondents' Listening Subtest scores and the Total Test score. If a large percentage of respondents 'disagreed' or 'strongly disagreed' with the statement and therefore reported technical problems with the headphones, then examining test takers' mean scores on the Listening Subtest will provide an indication of the extent to which performance on the Listening Subtest was affected by such technical issues. Such technical problems with the test are considered in the validity framework (Appendix A, pp. 140-142) as potential technology-related sources of measurement error that point to construct-irrelevant variance. This analysis was undertaken by grouping respondents according to the option they selected (e.g., 1–5 on the Likert scale). Chapter 5 (RQ2 results) presents the outcomes of all of these statistical analyses showing the differences in the test performance of these respondent groups.

3.6.3. RQ2: Inferential statistics.

The findings of this study were tested for significance to identify whether the dependent variable of test performance was affected by the technology-related independent variables investigated by the questionnaire items. Exploratory analyses on SPSS showed that the data did not meet normality and homogeneity of variance assumptions of parametric tests.

The results of testing normality assumptions (Table 3.3) using the Shapiro-Wilk test identified that some of the data deviate from normality. Although the Kolmogorov-Smirnov is another test of normality that is reported alongside the Shapiro-Wilk Test, the latter is reported in Table 3.3 because it is considered a better test of normality (Field, 2009). This procedure is to test normality of a dependent variable (test scores) across all levels of an independent variable (responses to questionnaire items). The dependent variable needs to be approximately normally distributed for each category or level of the independent variable.

Table 3.3. *Data Deviating from Normal Distribution*

Data Element	No. of Questionnaire Response Options (Student Group)	Shapiro-Wilk Test Sig. Value	Skewness (z-Value)	Kurtosis (z-Value)
Q6	3	0.030	2.29	2.96
Q12	3	0.038	1.39	-0.48
Q18	1	0.006	1.73	-1.63
Q19	1	0.004	1.85	-1.53
Q20	5	0.047	2.04	1.29
Q27	3	0.032	.88	-1.39
Q29	5	0.049	1.33	-1.08

Q31	3	0.028	.52	-0.82
Q35	2	0.005	2.06	-1.32
Q22	2 Listening Test	0.032	1.36	-0.32
	3 Total Test	0.031	1.74	-0.13
Q11	4 Listening Test	0.037	1.12	-0.18
	4 Total Test	0.021	1.54	-1.13
Q8	3	0.021	1.70	-0.26
Q14	3 Language Use Test	0.000	2.18	1.12
	4 Total Test	0.003	1.13	1.59

Furthermore, for a normal distribution, the skewness and kurtosis z-values should be within the range of -1.96 to +1.96 and the Shapiro-Wilk Test *p*-value (Sig.) should be above 0.05. Otherwise, the null hypothesis of a normal distribution gets rejected. To calculate the z-values for the skewness and kurtosis, we need to divide their values by the standard errors for each. The obtained z-values from this should be within the range of -1.96 to +1.96. Based on that, we can conclude whether the data is skewed or kurtotic. If the values are within the acceptable range, we can conclude that they do not differ significantly from normality and that the data are approximately normally distributed. Once it is checked that the dependent variable is approximately normally distributed for each category of the independent variable, parametric tests can be used. Inferential statistics classified as non-parametric methods will need to be used if it is not normally distributed as they do not make assumptions about the distributions. To summarise the results reported in Table 3.3, testing the normality assumption indicated that this assumption was violated, which means that non-parametric tests should be used to test for significance of the study findings.

The homogeneity of variance assumption was tested in the data using Leven statistic on SPSS. The assumption is that the data should show equal variances across the groups (Green, 2013). The results of testing this assumption revealed that the data exhibited equal variances across the groups for most of the questionnaire items. The data items that were found to have variances that are significantly different in different groups were Q25 for the total test, Q31 for the total test, and Q14 for the language use test. The significance level of the Levene statistic was less than 0.05 for these items. This indicated that the groups responding to these items showed significantly different variances, which violated the assumption of heterogeneity of variances. Violation of this assumption suggests the use of a non-parametric method to test group differences (Field, 2009).

In this study, non-parametric statistical tests were chosen to test for statistical significance since the study data did not meet the stringent assumptions (especially normal distribution and homogeneity of variance) of the parametric tests and because the data were measured on ordinal ranked scales (Likert for the independent variables). For nominal and non-normally distributed data, non-parametric tests are techniques that should be used to test for significant differences between groups (Bachman, 2004). The assumptions of non-parametric techniques are random sampling and independence of observations, which were met in the case of this study data.

The Kruskal-Wallis Test was the statistical test chosen to assess significance of the findings obtained from statistically comparing the test performance of the respondent groups. This test is the non-parametric alternative to one-way between-groups analysis of variance (one-way ANOVA) (Pallant, 2010). The Kruskal-Wallis Test fits the type of data we have in the study since it allows the comparison of one dependent continuous variable (test scores) and one categorical independent variable with two to five ordinal Likert scale categories (each of the technology-related factors represented by the questionnaire items).

Post hoc pairwise comparisons using the Dunn-Bonferroni method were conducted on the variables showing significant Kruskal-Wallis Tests. Where significant results were found on the Kruskal-Wallis Tests and the Dunn-Bonferroni post hoc pairwise comparisons, effect sizes (*r*) were also calculated. Calculation of the effect sizes when using the non-parametric Kruskal-Wallis Tests and their post hoc comparisons was based on Cohen's eta squared effect size statistics (Pallant, 2010). These effect size statistics indicate the strength of association or magnitude of the differences between the examined means by showing how much of the dependent variable variance that the independent variable can explain (Tabachnick & Fidell 2007). In this study, the effect size calculation results were categorised based on Cohen's (1988, p. 22) guidelines into small (*r* between .01 and .05), medium (*r* between .06 and .13), and large effect ($r \geq .138$, approximately .14). As such, the testing for significance indicated the extent to which the technology-related issues (independent variables) affected test results (dependent variable).

The findings of these analyses for each technology-related independent variable are reported in the RQ2 results chapter (Chapter 5) to provide evidence on reliability and construct validity in an answer to RQ2. Appendices Q to R (pp. 203-250) which are referred to in Chapter 5 provide detailed results of these statistical analyses.

3.6.4. RQ2: Non-psychometric analyses.

A number of statistical procedures exist in the literature to investigate construct validity such as factor analytic techniques (Green, 2013; Kunnan, 1992; Pallant, 2010), unidimentionality studies (Brown, 2005), analysis of differential item functioning (DIF) (Zumbo, 2007), ANOVA designs (Brown, 2005; Green, 2013; Pallant, 2010), multi-trait-multimethod studies (Brown, 2001), and generalizability studies (Brennan, 1992). However, this study used a basic statistical approach in order to provide just one part of a high resolution snapshot on the extent to which technology-related issues can impact test performance and create bias concerns. Due to the scope of the study, a bias analysis was not done and instead a simple statistical analysis approach was followed. Suggestions as to how this could be addressed are provided as further research in Section 7.5.7 (p. 124). Advocating the use of a non-psychometric approach to examine construct relevance or irrelevance of variance and its sources (Cohen, 2012), Davidson (2000) argues that statistical "evidence should be fused with and weighted against evidence that derives from other sources" (p. 615). Therefore, statistical evidence in this study was complemented with qualitative evidence on the testing mode effect to complete the high resolution snapshot.

Qualitative data that were gathered in the study included responses to open-ended constructed-response items on test takers' questionnaires and invigilators' open-ended items; transcribed audio-recorded interviews of testees and invigilators; and the researchers' field notes and observations on the reflective journals. However, as mentioned in Section 3.3 of this chapter (pp. 32-35), only test takers' questionnaires were used as the focus for further analysis.

Thematic induction (Bazeley, 2010; Bazeley & Jackson, 2013) was used to analyse comments provided by test takers in the open-ended parts of the questionnaire items (Q18, Q19, Q34, Q35, and Q36). Common themes and patterns emerged from the data via a process of coding, organizing, triangulating, and interpreting the data. This method is commonly used in narrative inquiry for the analysis of verbal and open text responses and serves to enrich the thick description of the case study. The identified themes were used as a framework to present the results in relation to RQ2 in Chapter 5 that included technology-related issues investigated by the questionnaire. As such, the study used psychometric and non-psychometric analyses procedures as methods to establish reliability and construct validity evidence.

3.7. Conclusion

This chapter has laid out the guiding principles for the use of the research methods and how these methods have been used to answer the research questions. In doing so this chapter has argued that the methods used to collect evidence for the validity argument are justified and well-suited to the task. This chapter included sections on the methodological approach, the study design, participants, data collection instruments, and data analysis procedures. This chapter also highlighted the contribution that the pilot phase played in the development of the Moodle-hosted test and the research instruments used in the main study. The main study results are the focus of the remainder of this thesis and will be presented in relation to the research questions they are intended to address in Chapters 4 (RQ1 results) and Chapter 5 (RQ2 results) and these results will be discussed in Chapter 6 (Discussion).

Chapter 4. Results and Discussion of Research Question 1

4.1. Introduction

The overall aim of this study was to provide a validity argument about using a Moodle-hosted test for its intended purpose by empirically examining reliability and construct validity evidence. This chapter explicitly states the evidence that will be used in the validity argument in relation to the first research question (RQ1): *To what extent can the Moodle-hosted test scores be reliable and valid indicators of the tested construct*? Note that Chapter 5 will report the findings (evidence) in relation to the impact of technology-related construct-irrelevant factors that could affect the reliability and construct validity of the Moodle-hosted test.

As described in Section 3.6.1 in the Methodology Chapter (pp. 44-46), Rasch item analysis was the statistical test employed to establish reliability and construct validity evidence in order to answer RQ1. As presented in the validity framework (Section 2.8, Chapter 2, pp. 28-29) and in the RQ1 data analyses section (Chapter 3, Section 3.6.1, pp. 44-46), the assumption here is that statistical evidence of high reliability estimates, fit and highly discriminating items, and acceptable low Standard Error of Measurement (SEM) will be a warrant that the Moodle-hosted test scores will be reliable and valid indicators of the tested construct.

Information on descriptive statistics is given in Appendix N (pp. 191-192) including a table and a histogram showing descriptive statistics of the Moodle-hosted test such as the mean and standard deviation. We first begin this chapter by exploring the results and discussion of Rasch analysis and then summarise and discuss these results. The Rasch produced a number of statistical outputs that provided information on reliability estimates, item difficulty measures, fit statistics, and measurement error.

4.2. Reliability Estimates

Through the Rasch item analysis, two reliability estimates were obtained in Winsteps Convergence Table, item reliability of 0.96 and person reliability of 0.80. When interpreting these values, we come to the conclusion that the reliability estimates for the entire test are within the acceptable range. This inference is made because internal reliability values that are at or above 0.70 are acceptable and values above 0.80 are usually preferred (Pallant, 2013).

According to Green (2013), the high item reliability value of 0.96 suggests that we can have a good amount of confidence that performance of test items can be replicated when tested under similar conditions on another test population. This high value also indicates that the study sample is large enough for reliability analysis. Person reliability provides an estimate of the amount of confidence we can have in the test takers' scores and the extent to which their performance will be replicated when taking a similar test in similar conditions. The person reliability value of 0.80 is a marginal value indicating that the abilities of test takers are not sufficiently measured by this set of items. This means that we need better quality test items that target different ranges of ability.

When running the Rasch analysis using the score data of each subtest, the analysis produced different reliability for each subtest. For the Reading Subtest, person reliability was 0.60 and item reliability was 0.94. In the Language Use Subtest, person reliability was 0.58 and item reliability was 0.93. For the Listening Subtest, person reliability was 0.57 and item reliability was 0.97. The high item reliability figures mean that we can have a good amount of confidence that performance of items in each subtest can be replicated when tested on another sample under similar conditions. These high figures also suggest that the test sample size is sufficient for reliability analysis. The low person reliability figures mean that we can place a lower amount of confidence on the test takers' scores on each subtest and the extent to which their performance will be replicated when taking similar subtests in similar conditions. These low person reliability values also indicate that the set of items in each subtest did not sufficiently measure the abilities of test takers, and better quality test items are needed to target different ability ranges (Green, 2013). Consistent with the reliability estimates for the whole test, estimates of person reliability were lower than the high item reliability estimates for all three subtests. Person reliability estimates for all three subtests were not acceptable as they were below the lowest acceptable reliability figure of 0.70 (Pallant, 2013).

The Rasch analysis provides an accurate picture of how item difficulty and person ability match each other by producing the two estimates for persons and items. The following sections (4.3 to 4.5) present detailed Rasch results in relation to item measures, item fit statistics, and measurement error.

4.3. Item Measures

The model reflected by this item analysis can be considered a Rasch dichotomous model (Green, 2013). In this model, it is likely that a person answers an item correctly as a function of the person ability and item difficulty. The results of the Rasch analysis are produced in tables of item and person measures as well as graphical representations in the form of Variable maps called an Item map and a Person map. These Rasch outputs map out the ability range of test takers, difficulty range of test items, and the extent to which the two variables of person ability and item difficulty match each other.

Table 4.1 presents selected Rasch item measures statistical results for each item including the total correct responses, item difficulty, measurement error, and fit statistics. The selection of these statistics was made based on their usefulness for the intended reliability analysis in terms of providing information about item difficulty, discrimination, and the associated measurement error. The difficulty column in Table 4.1 lists item difficulty levels in logits (also called item measures). These item measures ranged from +5.88 logits (item Q22LU1) to -2.34 logits (item Q16R3). The range of 8.22 logits is large, indicating that item difficulty levels varied. The Rasch analysis results are also plotted on the item map (Figure 4.1).

Table 4.1. *Rasch Item Measures Results: Selected Statistics*

Item	Total Correct	Item Difficulty	Error[d]	Fit[e]
Q22LU[a]1	0	5.88	**1.83**	0.00**
Q42LS[b]1	3	3.56	**0.58**	0.65**
Q35LU2	4	3.26	**0.51**	1.05
Q26LU1	7	2.68	**0.39**	0.78**
Q41LS1	7	2.68	**0.39**	1.00
Q32LU2	8	2.54	**0.36**	0.65**
Q25LU1	9	2.41	**0.34**	1.22*
Q27LU1	17	1.71	**0.26**	0.58**
Q37LU2	21	1.47	**0.24**	0.81
Q34LU2	23	1.36	**0.23**	1.25*
Q24LU1	24	1.31	**0.22**	1.12
Q38LU2	24	1.31	**0.22**	0.74**
Q23LU1	28	1.13	**0.21**	0.61**
Q40LU2	30	1.04	**0.20**	0.67**
Q52LS2	30	1.04	**0.20**	0.74**
Q30LU1	35	0.84	0.19	0.79**
Q36LU2	35	0.84	0.19	0.92
Q39LU2	35	0.84	0.19	0.74**
Q57LS2	44	0.54	0.18	1.06
Q44LS1	51	0.33	0.17	1.22*
Q11R[c]2	52	0.3	0.17	0.96
Q54LS2	54	0.25	0.17	0.91
Q53LS2	55	0.22	0.17	0.91
Q56LS2	59	0.11	0.16	1.14
Q6R1	64	-0.02	0.16	1.11
Q20R3	67	-0.09	0.16	1.08
Q3R1	69	-0.14	0.16	0.91
Q14R2	69	-0.14	0.16	1.03

Q31LU2	69	-0.14	0.16	0.81
Q21LU1	71	-0.19	0.15	0.99
Q15R2	74	-0.26	0.15	1.21*
Q51LS2	75	-0.28	0.15	0.99
Q60LS2	75	-0.28	0.15	0.98
Q45LS1	78	-0.35	0.15	1.02
Q8R1	83	-0.47	0.15	1.14
Q33LU2	87	-0.55	0.15	0.74**
Q48LS1	89	-0.6	0.15	1.05
Q9R2	92	-0.67	0.15	0.98
Q47LS1	95	-0.73	0.15	1.19
Q18R3	96	-0.75	0.15	1.01
Q17R3	97	-0.78	0.15	1.13
Q19R3	98	-0.8	0.15	0.99
Q2R1	99	-0.82	0.15	0.99
Q10R2	100	-0.84	0.15	1.02
Q59LS2	101	-0.86	0.15	1.24*
Q50LS1	106	-0.97	0.15	1.02
Q28LU1	107	-0.99	0.15	0.81
Q1R1	113	-1.12	0.15	1.12
Q4R1	113	-1.12	0.15	1.10
Q13R2	116	-1.19	0.15	0.96
Q46LS1	124	-1.37	0.15	1.09
Q29LU1	125	-1.39	0.15	0.88
Q12R2	126	-1.41	0.15	1.18
Q5R1	131	-1.53	0.15	0.99
Q43LS1	132	-1.55	0.15	1.03
Q58LS2	135	-1.62	0.15	1.04
Q49LS1	136	-1.65	0.15	1.02
Q55LS2	145	-1.87	0.16	0.82
Q7R1	146	-1.89	0.16	1.07
Q16R3	162	-2.34	0.18	0.71**

Notes. [a]Language Use; [b]Listening; [c]Reading; [d]Error acceptable value = less than 0.20; large unacceptable error values in bold; [e]acceptable fit range for high-stakes test = 0.80 to 1.20 (1.0 is perfect fit; *misfit= over 1.20; **overfit = below 0.80).

```
Measure Person - Map - Item
               <more>|<rare>
    4                +  Q22LU1
                     |
                     |
                     |
                     |  Q42LS1
                     |
                     |  Q35LU2
                     |
    3                +
                     |
                     |T
                     |  Q26LU1  Q41LS1
                     |  Q32LU2
                     |  Q25LU1
                     |
                     |
    2                +
                     |
                     |  Q27LU1
                     |
                     |  Q37LU2
                     |S Q34LU2
                     |  Q24LU1  Q38LU2
                     |  Q23LU1
    1                +  Q40LU2  Q52LS2
                  .  |  Q30LU1  Q36LU2  Q39LU2
                     |
                  .  |
                 .# T|  Q57LS2
                .##  |  Q44LS1
               ####  |  Q11R2   Q53LS2  Q54LS2
                  #  |  Q56LS2
    0          ####  +M Q6R1
               .### S|  Q14R2   Q20R3   Q31LU2  Q3R1
                .##  |  Q15R2   Q21LU1  Q51LS2  Q60LS2
          #########  |  Q45LS1
               .###  |  Q33LU2  Q8R1
        ###########  |  Q48LS1  Q9R2
              #####  |  Q17R3   Q18R3   Q19R3   Q47LS1
              .#### M|  Q10R2   Q2R1    Q59LS2
   -1    .#########  +  Q28LU1  Q50LS1
               .###  |  Q1R1    Q4R1
                ###  |  Q13R2
             .#####  |S Q12R2   Q29LU1  Q46LS1
              #####  |  Q43LS1  Q5R1
        .########## S|  Q49LS1  Q58LS2
                ###  |
                 ##  |  Q55LS2  Q7R1
   -2                +
                ###  |
                .##  |
                 .# T|  Q16R3
                  .  |
                     |
                  #  |T
                     |
   -3                +
               <less>|<freq>
EACH "#" IS 2: EACH "." IS 1
```

Figure 4.1. Item-person map

Figure 4.1 shows that the more difficult items are at the top right side of the map and that no persons are plotted against them on the opposite side of the line. Q22LU1 was the most difficult item and Q16R3 was the easiest item. Fifteen items were found at the top of the item map with no persons on the other side, which means that these items were the most difficult on the test. These items were Q22LU1, Q42LS1, Q35LU2, Q26LU1, Q41LS1, Q32LU2, Q25LU1, Q27LU1, Q37LU2, Q34LU2, Q24LU1, Q38LU2, Q23LU1, Q40LU2, and Q52LS2. This finding indicates that the difficulty level of 25.0% of the test (15 items out of a total 60) did not match the ability levels of the test takers. Moreover, when we look at the bottom of the map, we find that the test was too difficult for 8.2% (*n* = 17) of the persons situated between the lowest logits of -2 and -3, with only one item (Q16R3) matching the person ability measures.

These results suggest that the varying range of ability levels of the test sample were not matched by these too difficult items. These items were gap-filling items in Language Use and Listening Subtests. Given this evidence, the test sample found it difficult to respond to gap-filling item types in Language Use and Listening Subtests. In the Rasch analysis reported in this chapter, no specific pattern explained why these items were high in difficulty levels. However, we might explain this trend by assuming that typing responses to these items might be challenging for test takers. Task difficulty of constructing words to fit into the given context in such item types might also explain this trend as test takers were not provided with hints from a list in this task. Chapter 5 will present more evidence on the test takers' performance on gap-filling items through the questionnaire analysis results and the comparison of test performance with questionnaire responses.

Going back to Figure 4.1 (p. 56) showing the item map, we find gaps between items. This finding suggests the need to include items that target a range of test takers' ability levels at these gaps. The gaps can be seen where no items were matching the ability levels of 11 test takers placed at logit points between item Q16R3 and items Q55LS2 and Q7R1. None of the items at a difficulty level between items Q49LS1 and Q58LS2 and items Q55LS2 and Q7R1 matched the ability levels of six other test takers. Likewise, the ability levels of three test takers were not targeted with items easier than Q16R3. There are also gaps between the most difficult items at the top of the map. Such gaps in item difficulty indicate a mismatch between item difficulty and person ability measures, which could be considered instances of construct under-representation. To achieve precise measurement, item difficulty should match person ability levels and there should not be such big gaps between the items on the map as these gaps mean that more items are needed to precisely measure the untargeted person abilities (Baghaei, 2008). Hence, these Rasch results provided evidence of threats to construct validity, namely construct-irrelevance and construct under-representation.

4.4. Fit Statistics

Fit statistics shown on the last column of Table 4.1 (p. 54) are also useful Rasch outputs. Fit statistics provide information on how each item contributes to the tested construct and identify whether most of the test items measure the targeted ability. Based on fit indices, items get classified as fit, misfit, and overfit items. A misfit item assesses student performances inconsistently, which can be observed in the response pattern to this item not corresponding to the response pattern expected by the Rasch model. As such, a misfit item does not discriminate between low and high ability test takers. An overfit item does not contribute much to measuring the test construct in that it is a redundant and dependent item that does not function independently of other test items (Akiyama, 2001).

McNamara (1996) suggests that fit values ranging from 0.80 to 1.30 are appropriate or acceptable. However, for high stakes tests, Linacre (2014) limits acceptable fit values to be within a range from 0.80 to 1.20. Therefore, in interpreting the results of Rasch fit statistics in this study, items with fit values below 0.80 were considered overfitting and items with values above 1.20 were considered misfitting. Fit values are shown in the last column of Table 4.1 (p. 54) and marked on the item map (Figure 4.1, p. 56) with bold for overfit items and underlining for misfit items.

Each of the test items that are within the acceptable range of the fit index for a high-stakes test makes an independent contribution to the tested construct (McNamara, 1996). When examining fit statistics, 70.0% (*n* = 42) of the test items were found to be within the acceptable fit range of 0.80 and 1.20. On the other hand, 30.0% (*n* = 18) of the items had unacceptable fit values. These findings suggest that as expected by the Rasch response model, each of the 70.0% of the items made an independent contribution to the tested construct and consistently assessed student performances.

As shown in Table 4.1 (p. 54) and Figure 4.1 (p. 56), of the 30.0% (*n* = 18) unacceptable fit items, five items (Q25LU1, Q34LU2, Q44LS1, Q15R2, Q59LS2) were misfit since their fit indices were above 1.20. Finding the five misfit items means that the response pattern of 8.3% of the test items (*n* = 5) did not correspond to the response pattern expected by the Rasch model, so these items were less predictable. For instance, the least able test takers predicted to answer these items incorrectly unexpectedly answered them correctly while the more able test takers answered them incorrectly. Finding such misfitting items in the test signals unwanted noise in the data. Such items need to be carefully revised or discarded from the test because they do not contribute much to the testing instrument (McNamara, 1996). Unlike the rest of the acceptable fit items, misfit items do not consistently discriminate between low and high ability students (Akiyama, 2001).

Misfit items are considered a threat to construct validity because they indicate departure from test unidimensionality. *Unidimensionality* is the term used to refer to measuring one single dimension (language proficiency here) by the defined test construct. Items that fit the Rasch model indicate that they measure the single dimension intended by the construct theory. On the other hand, items that do not fit the Rasch model are indicators of multidimensionality and might need to be modified or discarded as such items do not contribute to measuring the tested construct. This means that items that do not fit the Rasch model do not only measure the single construct of language proficiency since other sub-dimensions that are irrelevant to the construct are being measured as well (Baghaei, 2008).

Furthermore, of the 30.0% (*n* = 18) unacceptable fit items, 21.7% (*n* = 13) of the test items (Table 4.1, p. 54) were found overfitting (Q22LU1, Q42LS1, Q26LU1, Q32LU2, Q27LU1, Q38LU2, Q23LU1, Q40LU2, Q52LS2, Q30LU1, Q39LU2, Q33LU2, Q16R3). Overfit items are considered redundant items that are dependent on other items and point to a lack of local independence (Akiyama, 2001). Such items do not contribute independently to the test as is the case when items that are based on the same information (such as items based on a paragraph) do not work independently of each other. The dependence of the overfit items on other items might be explained by the fact that all these items except Q16R3 (from Reading Test Three) are from the Language Use and Listening Test Subtests. Responding to the items in these subtests required access to the information presented in context such as the Language Use Subtest passages and the Listening Subtest script. Both model fit and local independence are Rasch modeling principles that support the unidimensionality assumption (Bond & Fox, 2007). Finding 30.0% (*n* = 18) of the items misfit and overfit signals that these items do not make independent contributions to constructing the ability to be tested. Results of overfit items are discussed in Chapter 6 (p. 88).

These results also confirm what Linacre (2014) mentioned that more discriminating items tend to be overfit (low fit index; less than the 0.80 cut-off) while the less discriminating items tend to be misfit (high fit index; greater than the 1.20 cut-off). Therefore, misfit items are problematic, but overfit items might have a high level of discrimination. As such, fit statistics have provided information on item discrimination. Based on the results reported in Table 4.1 (p. 54), this means that the five misfit items were less discriminating and the 13 overfit items were more discriminating. These item fit and discrimination results will be discussed further when presenting measurement error results in Section 4.5.

The Rasch analysis also generated person fit statistics. Finding misfitting persons indicates that the testing instrument did not capture their ability levels well (Knoch & McNamara, 2015). As can be seen in the person measure table (Appendix O, Table O1, pp. 193-196) and the person map (Appendix O, Figure O1, pp. 197-198), of the 207 test takers, 55.1% (*n* = 114) fit the Rasch model as their fit indices were within the acceptable fit range for high-stakes tests (0.8 to 1.20). On the other hand, 44.9% (*n* = 93) of the examinees did not conform to the acceptable fit indices because 30.4% (*n* = 63) were overfit persons and 14.5% (*n* = 30) were misfit persons. These findings indicate that the ability levels of the persons with unacceptable fit were not captured well by the test. When it comes to error values for persons, they are all larger than the lowest acceptable value of 0.20. We can infer from these findings that a high ability level examinee's true ability might not be reliably tested given the large error value and the unacceptably large fit index. Overall, the results suggested that the test might not have measured takers' true language ability reliably as 44.9% (*n* = 93) of the test takers did not fit the Rasch-expected response model and their measurement error values were unacceptable.

4.5. Measurement Error

To identify the amount of measurement error in the test and the amount of test unreliability, we need to examine the SEM value. As mentioned in the Methodology (Section 3.6.1, pp. 44-46), it is desirable to have a low SEM to have more reliable test results. The SEM value here points to the amount of measurement error in the test and consequently reflects its unreliability aspect, which can be more informative than a reliability estimate (Brown, 1999). Measurement error is the difference between Observed and True scores. The Observed score is the test taker's actual score obtained in the exam. The True score is the test taker's actual ability. It is important to measure error because reliability decreases when there is more error in the observed scores. The opposite holds true as decreased measurement error leads to increased reliability. This means a reliability estimate would be the relationship between True score variance and Observed score variance (Castle, 2016). As explained in Brown (1999), an estimate of internal reliability estimates the proportion of variance in the actual test scores that would be attributable to true score variance. The SEM is useful as it gives an estimate of how much variability in actual test scores can be expected around a cut pass score to be due to unreliable variance (i.e., error).

As stated by Linacre (2014), Rasch Test SEM value can be seen in one of the Winstepts output tables. It can also be calculated using the following formula: Rasch Test SEM = (Standard Deviation of Person Measures) * square-root (1 - Person Measure Reliability). The resulting SEM value for the test was 0.32 in logit points. This value is considered an average SEM for the sample of persons and is equivalent to Test SEM of Classical Test Theory (CTT). The SEM value in Rasch is known as the root-mean-square error (RMSE) for the sample of persons. Based on the raw-score Test Reliability Cronbach Alpha (KR-20), Rasch also reports the CTT average Raw-Score Test

SEM for the persons sample using the formula: Raw-Score Test SEM = (Standard Deviation of Person Raw Scores) * square-root (1 - Raw-Score Test Reliability). Since the raw-score Test Reliability Cronbach Alpha (KR-20) reported in Rasch outputs was 0.81 and the SD was 7.4, the resulting CTT's Test SEM value was 3.20 raw scores. The 3.20 error value in raw scores is equivalent to the 0.32 error value in logit points.

The error value of 0.32 in logits is a large measurement error as it is greater than the 0.20 acceptable error value in Rasch measurement (Akiyama, 2001). This SEM value represents the error value of the whole test rather than for each item. Rasch outputs (Table 4.1, p. 54) also show the amount of error for each item shown in the Error column. Error values shown for each item indicate that the larger these error values are, the less confidence we can have on the item difficulty measures because their logit values cannot be replicated in other testing occasions (Green, 2013).

While the SEM for the whole test was 0.32, the highest SEM value of 0.49 was recorded for the Language Use Subtest. The next highest SEM value was 0.22 for the Listening Subtest. The lowest SEM value was 0.16 for the Reading Subtest. Therefore, the SEM value of the Reading Subtest was acceptable, but the SEM values of the Language Use and Listening Subtests were indicators of large unacceptable measurement error.

Looking back at Table 4.1 (p. 54), we find that 25.0% (*n* = 15; of a possible 60) of the test items had large unacceptable error values (at or greater than 0.20) as they ranged from 0.20 to 1.83 for the 15 most difficult items. These were large error values. Three of the items with large error values (Q42LS1, Q41LS1, and Q52LS2) were from the Listening Test Subtest. These Listening items required test takers to fill in the gaps by typing in words as they listened to the audio file. The remaining 12 difficult items (Q22LU1, Q35LU2, Q26LU1, Q32LU2, Q25LU1, Q27LU1, Q37LU2, Q34LU2, Q24LU1, Q38LU2, Q23LU1, and Q40LU2) were from the Language Use Subtest. Similar to the Listening items, these Language Use items were gap-fill or constructed-response items that required test takers to type short answer responses into designated boxes without being given a word list stimulus.

Error values of 75.0% (*n* = 45) of the test items were low and acceptable (less than 0.20). Among these acceptable low error items, 38 items were well-discriminating as their fit values were within the acceptable range. Four of the low error items were overfit items that had high discrimination and three were misfit low discriminating items. Two (Q25LU1, Q34LU2) of the five misfit items do not assess test performance precisely because they had such unacceptable fit statistics indicating low discrimination and their error values were large (greater than 0.20). These results imply that items that had unacceptable error values and at the same time were less discriminating (given their fit indices) did not contribute much to reliability. This suggests that the two misfit items (Q25LU1 and Q34LU2) that had low discrimination and large unacceptable error values are the most problematic and might need to be removed from the test. The three other misfit low discriminating items (Q44LS1, Q15R2, Q59LS2) that had acceptable error values might not be as problematic and probably just need to be revised in order to raise their contribution to discrimination and reliability.

The four overfit items (Q30LU1, Q39LU2, Q33LU2, Q16R3) were very good items in terms of discrimination, given their fit indices and acceptable low error values. Since having a high level of discrimination is a desired item quality, it is worthwhile to retain the highly discriminating overfit items in the test (Green, 2013). Though the remaining nine overfit items (Q22LU1, Q42LS1, Q26LU1, Q32LU2, Q27LU1, Q38LU2, Q23LU1, Q40LU2, Q52LS2) were very good in discriminating between test takers' abilities given their fit values, their error values were largely unacceptable. Thus, these nine overfit items do not contribute independently to assessing the target construct.

In sum, through the Rasch analysis, this study reported finding 25.0% of the test items of the gap-filling type to be too difficult, indicating construct-irrelevant difficulty. The term construct-irrelevant difficulty refers to including tasks or test items that increase the difficulty of the tested construct and subsequently, some test takers can have invalidly low scores (Baghaei, 2008). Having these items increased the test difficulty and might have produced invalidly low test scores for some test takers. Instances of construct under-representation were also found by locating gaps in person ability levels not being targeted by items. Also, 30.0% (*n* = 18) of the test items had unacceptable fit indices. Five of these items were misfit and 13 were overfit. The findings imply that 30.0% of the items are not reliably testing the target ability since they do not contribute to constructing the target ability (overfitting items) nor assess student performances consistently as expected by the Rasch response model (misfitting items). On the other hand, the remaining 70.0% of the items (*n* = 42) consistently assess student performances and make independent contributions to reliably measure the target construct. It should be acknowledged here that while overfit items are not harmful, misfit items can have an impact, as stated by McNamara (1996). The conclusions made about test results are affected by this interpretation concerning misfit items.

Besides, the test had a large average SEM value of 0.32 logits and 25.0% of the test items had unacceptable error values. Among these items, two of these large error items were misfit low discriminating items and, hence, did not contribute much to reliability. Since a quarter of the test items had unacceptable error values, we should be less confident about their contribution to reliability. Furthermore, error values for the entire test and for the Language Use and Listening

Subtests were highly unacceptable. Given all of these results, taking decisions based on the scores obtained from this test that had problematic items might not be reliable and, hence, might invalidate these decisions.

4.6. Summary of Results

To summarise the findings presented above, the results reported in this chapter included Rasch analysis results on reliability and construct validity of the test. The Rasch measurement produced highly acceptable reliability estimates indicating a reliable test overall. By identifying individual item difficulty, measurement error values, and fit indices, the Rasch results also provided discrimination information indicating problematic items that need revision. A quarter of the test items (25%; *n* = 15) were overly difficult and had high unacceptable error values. These were gap-filling items in the Language Use and Listening Subtests calling upon test takers' typing ability, which indicated the presence of construct-irrelevant difficulty. It was also found that overall, 30.0% of the test items had large unacceptable fit statistics, suggesting that they did not make independent contributions to test the target language ability and did not contribute to test reliability. Furthermore, construct under-representation was identified by finding gaps between item difficulty and person ability measures along the unidimensional continuum in the Rasch item-person map, which indicated that the examinees' ability levels were not well-captured by the test. This was supported by finding that 44.9% (*n* = 93) of the test takers did not fit the Rasch-expected response model and their measurement error values were unacceptable, indicating the test might not have measured takers' true language ability reliably.

Based on the study outcomes reported in this chapter, we can address the research question RQ1 examining the extent to which the Moodle-hosted test scores can be reliable and valid indicators of the tested construct. Overall results have shown that if we consider the high reliability estimates alone, the reliability of the test as a whole can be deemed satisfactory. However, we need to also consider other evidence that pointed to reliability and construct validity concerns. Overall, the results suggested that construct-irrelevant variance and construct under-representation threatened reliability and construct validity. Based on all of these study outcomes, we infer that the Moodle-hosted test scores might not be reliable and valid indicators of the target test construct. Consequently, the test might not be valid to use for its intended pass/fail decision-making purpose.

4.7. Conclusion

Overall, by applying the AUA framework principles (Bachman, 2005; Bachman & Palmer, 2010), the validation framework (Appendix A, pp. 140-142) acted as a pragmatic tool for conducting this study. This chapter has stated the evidence in relation to Research Question One on reliability and validity of the test scores. We can state that the backing evidence warranted that the test overall had acceptable reliability estimates. Nevertheless, other evidence became the rebuttal as it refuted reliability and construct validity claims. This counter evidence reported low discrimination through finding unacceptable fit statistics, and large measurement errors for the whole test and for individual items on the test. Evidence on construct-irrelevance and construct under-representation was also reported. In sum, in the case study of administering the test in a Moodle-hosted mode, reliability and construct validity concerns existed, which suggested that using the test for its intended purpose might be unreliable and invalid.

These results were reached based on statistical item analysis conducted on one data source, test scores. Further investigation is needed to look into the rebuttals to reliability and construct validity claims using other data types and analyses to provide supporting evidence for the validity argument. This further investigation is to cater for potential sources of measurement error in the context of the case study of administering the test in a Moodle-hosted testing environment. This leads us to the second research question of this study to be addressed in the next results chapter.

Chapter 5. Results for Research Question 2

5.1. Introduction

The overall aim of the study was to provide a validity argument about using a Moodle-hosted test for its intended purpose by empirically establishing evidence on reliability and construct validity. Chapter 4 has already reported results (evidence) in relation to the first research question covering the reliability and validity of test scores. This chapter explicitly states the evidence that will be used in the validity argument in relation to the second research question (RQ2) that is, *the extent to which technology-related factors associated with the testing mode effect can interfere with test results*.

As outlined in the methodology in Chapter 3 (Section 3.5.2, pp. 42-43), a questionnaire survey was undertaken seeking evidence in the form of test takers' perceptions of the impact of various technology factors on their test performance. The questionnaire included 36 items that were a combination of selected-response (closed) and open-ended questions. Descriptive statistics were used to report on test takers' responses to each selected-response question. A statistical comparison between test takers' questionnaire responses (Q3 to Q36) with mean test scores was done on SPSS, grouping test takers according to their responses to the Likert scale items. For example, for a five-point Likert scale item examining a technology factor, mean test scores of the five respondent groups were compared to identify differences in the test performance among these groups. Grouping was done by agreement level on the Likert scale. For instance, students selecting 1 (Strongly Disagree) in responding to a questionnaire item were grouped together and their mean test scores were examined. The same grouping procedure was done for the other groups selecting 2 (Disagree), 3 (Neutral), 4 (Agree), and 5 (Strongly Agree) in responding to that item. Comparing selected-response questionnaire items (test taker's perception of a given technology issue) to respondent's test performance data (test mean scores) linked test takers' perceptions of their testing experience and any reported technology-related issues to their test performance. In this linking analysis, a first round used the Kruskal-Wallis Test to identify whether the dependent variable of test performance was statistically significantly affected by the technology-related independent variables investigated by the questionnaire items. Where significant results were found, post hoc pairwise comparisons were done using the Dunn-Bonferroni method. Effect sizes (*r*) are also reported for significant results. See Sections 3.6.2 and 3.6.3 of Chapter 3 (pp. 46-49) for the details of the statistical methods used. This chapter presents the outcomes of all of these statistical analyses showing the differences in the test performance of the respondent groups. A summary of these findings is reported in this chapter. A detailed breakdown of response frequencies are provided in

Appendix P (pp. 199-202). Boxplots illustrating the results are in Appendix Q (pp. 203-241) while detailed tables of all Kruskal-Wallis Test results are given in Appendix R (pp. 242-250).

As described in Chapter 3 (Section 3.6.4, pp. 49-50), the quantitative analysis was supported with thematic induction carried out on the open-ended responses to questionnaire items (Q18, Q19, Q34, Q35, and Q36). For Q18, participants were asked to justify their preference for a testing mode, pen and paper or online in Moodle. In Q19, participants were also asked to explain why they think they would perform best when using their preferred mode. Q34 requested respondents to explain why they liked or did not like the test on Moodle. Q35 asked participants to explain why they would like or would not like to take official exams (like mid-terms, finals, placement tests, exit tests, and so forth) on Moodle where this would be used to make decisions about the level of their language proficiency. The last open-ended question Q36 sought comments from participants on Moodle-hosted online English language testing.

The following sections lay out each emergent theme. The results of statistical analyses carried out on the selected response items, including comparative performance, are presented under each theme. Representative comments are also provided. Finally, all findings in relation to RQ2 are summarised in Table 5.11 (p. 83) in Section 5.12 of this chapter where each questionnaire item is grouped by theme.

5.2. Theme 1: Endurance

This theme of ‘endurance’ comprises three items:

- test length;
- concentration; and
- eye fatigue.

Test takers were asked to indicate agreement on a Likert item for each issue. Across all of these items, the majority of respondents perceived there to be a problem. The most problematic of these was ‘Staring at the computer screen for a long period of time causing eye fatigue’, where 72.4% of test takers broadly agreed eye fatigue was an issue for them. When asked if they felt ‘Staring at the computer screen for a long period of time made them lose concentration’, two thirds (63.8%) broadly agreed to this being a problem. Likewise, two thirds (62.6%) were in broad agreement with the statement “The test was too long as it consisted of too many sections”, indicating that they perceived test length was an issue.

To determine if these issues impacted their test performance, responses were grouped according to their agreement on the Likert item (5 groups) and each response group was compared to the mean test performance by group. The results suggested a trend towards an effect on test performance of the groups broadly agreeing to the items targeting these issues. The item on ‘concentration loss’ was an exception where test taker performance did not show a trend either way. The group broadly agreeing that ‘test length’ was an issue scored lower than the group in broad disagreement. It was observed that the group strongly disagreeing to the item on ‘eye fatigue’ scored higher than the broadly agreeing groups. However, the strongly agreeing group was a small percentage (6.9%) of those that did not experience eye fatigue. Test performance was lower for the 72.4% who did complain about the issue.

Despite a majority of test takers perceiving these variables as issues, these observed differences in test performance were not found statistically significant. This was confirmed by the results of a Kruskal-Wallis Test (Table 5.1). See Appendix Q (Figures Q1 to Q3, pp. 203-205) for boxplots and Appendix R (Tables R1 to R3, p. 242) for detailed results on these issues.

Table 5.1. Endurance: Comparison with Test Scores

Item	Kruskal-Wallis Test
Q28: The test was too long as it consisted of too many sections.	*H*(5, *n* = 174) = 4.99, *p* = .417.
Q29: Staring at the computer screen for a long period of time made me lose my concentration.	*H*(4, *n* = 174) = 3.83, *p* = .430).
Q30: Staring at the computer screen for a long period of time caused me eye fatigue.	*H*(5, *n* = 174) = 3.56, *p* = .614.

The majority of the respondents (95.4%*;* *n* = 166) provided open-ended questionnaire responses relating to the theme of ‘endurance’. Comments from some students (11.5%; *n* = 20) complained that staring constantly at the computer screen during the lengthy exam caused them loss of concentration and eye fatigue. The issue of ‘concentration loss’ was reflected in the comments where more than half of them (52.3%; *n* = 91) complained about how they struggled to maintain their concentration in the Moodle-hosted environment. Some respondents also commented that they thought their concentration level would be higher in the paper-based testing mode. Another major concern was related to the effect on the health of one’s eyes. Some respondents (31.6%; *n* = 55) stated in their comments that doing the test caused them eye fatigue because of staring at the computer screen for a long period of time. Some even mentioned how eye strain may cause them headaches. The concerns of these students over eye strain influenced their response to the Moodle-hosted testing mode, revealing a preference for paper-based testing in order to avoid the risk of eye strain.

In order to decrease the impact on their eyes and to maintain their concentration, the respondents suggested reducing the test length, decreasing the screen brightness, and using screen protectors. These reported issues are matters of concern that need further investigation in light of how they may interact with other matters indicated by students' comments.

5.3. Theme 2: Ease of Use

This theme comprised a number of technology-related issues. These included:

- ease of test navigation;
- appropriateness of the background colour;
- clarity of procedures and instructions,
- being able to successfully log onto the test;
- clarity of pictures and graphs; and
- appropriateness of font size.

Test takers were asked to indicate agreement on a Likert item for each issue. Across all of these items, only a minority of respondents perceived there to be a problem. The most problematic of these was the use of 'inappropriate font size' where 21.3% of test takers broadly agreed the font size was a problem for them. However, across the other items only 2.3% to 4.6% of test takers thought the issues were a problem for them.

The results did suggest a trend towards lower test performance of the groups broadly disagreeing to the items targeting these issues. One exception was for 'clarity of procedures and instructions' where there was no trend either way. The broadly agreeing group complaining about 'inappropriate font size' scored lower than the broadly disagreeing group. However, as shown in Table 5.2, for all six items, the differences in test performance across agreement levels were not statistically significant. See Appendix Q (Figures Q4 to Q9, pp. 206-211) for boxplots and Appendix R (Tables R4 to R9, pp. 242-243) for detailed results.

Table 5.2. Ease of Use: Comparison with Test Scores

Item	Kruskal-Wallis Test
Q6: Overall, the test was easy to navigate by moving from one page displaying a subtest to another.	$H(5, n = 174) = 3.761, p = .584$.
Q9: I think the background theme (colours) of the test was appropriate.	$H(5, n = 174) = 4.720, p = .451$.
Q13: Test procedures and instructions given were clear and easy to follow.	$H(5, n = 174) = 5.833, p = .323$.
Q25: I was able to successfully log onto Moodle and the online test.	$H(4, n = 174) = 2.65, p = .618$.
Q26: Pictures and graphs were clear.	$H(3, n = 174) = 4.35, p = .226$.
Q27: The font size was NOT appropriate.	$H(4, n = 174) = 1.96, p = .743$.

The majority of the respondents (88.5%*;* *n* = 154) provided comments relating to ease of use and identified that such matters influenced their preferred testing mode. A small number, 2.3% (*n* = 4) of the respondents commented that they thought the online testing mode was overly complex. As one student put it:

> When the test is administered online, there will be lots of instructions (do this, don't do that, etc.) but in the paper-based mode I just hold my paper and start answering with my pen (simple!) (Level 6 Commerce male student).

Another student commented that they had "enough of!" the issue of test anxiety and said that there was no need to add more to it through complex testing processes. Most comments (36.2%*;* *n* = 63) relating to pictures and graphs indicated that test takers thought they were clear. In addition, two of these students suggested the inclusion of more pictures and graphs would also help make the questions clearer to them.

Although 5.2% (*n* = 9) of the test takers' comments indicated their satisfaction with the font size and background theme, a few other students (2.9%*;* *n* = 5) suggested to increase the font size and change the background theme. Similarly, most test takers commented that it was easy to navigate through the Moodle-hosted test. A number of comments (18.4%*;* *n* = 32) indicated difficulty with navigation in the Moodle-hosted test and compared it to the ease of navigating through the test pages on paper. As one student described:

> Going through test pages on paper is easier than this test where I have to make sure that I get to all pages on the screen and I can only check by loading every page. It is a lot faster to do this on paper (Level 4 General male student).

5.4. Theme 3: Experience with Moodle Tests

Four technology-related issues came under this theme, namely:

- level of familiarity with Moodle tests;
- level of familiarity with computers;
- having enough experience with technology; and
- need for extra technical training.

Test takers specified their familiarity levels in responding to the Likert item for each of the first two issues. For each of the last two issues, test takers indicated agreement on a Likert item. Most of the respondents claimed familiarity with computers (85.6%) and Moodle tests (82.2%). When asked if they felt they had enough experience with technology to enable them to undertake a Moodle test, the majority (71.3%) agreed. However, when asked if they need ‘extra technical training’ to cope with an online exam, 45.9% of test takers broadly agreed and 35.1% said they disagreed, making for more mixed results.

The results of comparing test performance to Likert item responses on items targeting these issues suggested an impact on test results. For all issues except the ‘need for extra technical training’, the differences in test performance across agreement levels were found statistically significant on Kruskal-Wallis Tests. The effect size for ‘having enough experience with technology’ was large while it was small for ‘level of familiarity with Moodle tests’ and ‘level of familiarity with computers’. See Table 5.3 for these results.

Similar significant differences with a large effect size were found in the post hoc pairwise comparisons (using the Dunn-Bonferroni method with adjusted significance levels) between the neutral and strongly agreeing groups responding to the item on ‘having enough experience with technology’. The strongly agreeing group recorded the highest median score and mean rank (*Md* = 23.00, *mean rank* = 104.40, *n* = 45). Likewise, significant differences with a large effect size were found in the post hoc comparisons between the groups responding with very familiar and a little bit familiar to the item on ‘level of familiarity with computers’. The group very familiar with computers had the highest median score and mean rank (*Md* = 21.50, *mean rank* = 101.48, *n* = 46). No statistically significant differences were found in follow-up post hoc pairwise comparisons among the groups responding to the item on the ‘level of familiarity with Moodle tests’. The group very familiar with Moodle tests had the best median score and mean rank (*Md* = 22.00, *mean rank* =

99.29, *n* = 68). See Appendices Q for boxplots (Figures Q10 to Q13, pp. 212-215) and R (Tables R10 to R13, pp. 244-245) for detailed results.

Table 5.3. *Familiarity and Experience: Comparison with Test Scores*

Item	Kruskal-Wallis Test (Dunn-Bonferroni effect size given where significant)
Q3: Your level of familiarity with tests or quizzes on Moodle: (Very familiar; Somehow familiar; A little bit familiar; Not familiar at all)	H(3, *n* = 174) = 7.899, *p* = .048, *r* = 0.05; Not significant in post hoc comparisons.
Q4: Your level of computer-literacy or familiarity with computers: (Very familiar; Somehow familiar; A little bit familiar; Not familiar at all)	H(2, *n* = 174) = 7.58, *p* = .023, *r* = 0.04; Post hoc pairwise comparisons (Very familiar and A little bit familiar), *p* = .020, *r* = 0.49.
Q32: I have enough experience with technology to take tests on Moodle.	H(5, *n* = 174) = 18.80, *p* = .002, *r* = 0.11; Post hoc pairwise comparisons (Neutral and Strongly agree), *p* = .001, *r* = 0.62
Q33: I will need extra technical training before I am ready to take online exams.	H(5, *n* = 174) = 4.44, *p* = .487.

Notes. Effect size (*r*): Small (.01 to .05); Medium (.06 to .13); Large (*r* ≥.14).

A small percentage of respondents (3.4%*; n* = 6) provided comments relating to experience with Moodle tests. These 3.4% of the respondents commented that test takers with insufficient expertise in using the computers, internet, and Moodle tests might be at a disadvantage. In contrast, these respondents argued that other test takers might be more advantaged because “they have more experience and have the necessary keyboarding skills to do the test”, as one Level 5 Science male student commented. On computer literacy, one Level 4 General male student also commented that “using the computer and how to deal with it is in itself a test. How about then taking a test using the computer?!!!” This comment fully delineates the situation as using technology to take a test introduces heavy requirements on some students, such as keyboarding skills, in order for them to perform well in this testing mode.

5.5. Theme 4: Attitude and Resistance to Change

The following three technology-related issues were grouped under this theme:

- attitude towards the test-taking experience;
- attitude towards using new technology to take the test; and

- attitude towards taking the test on Moodle.

Test takers indicated agreement on a Likert item for the first and second issues. For the third issue, test takers specified their attitude by responding with yes or no. Two thirds of test takers broadly agreed that they “liked the test-taking experience” (61.5%) and “liked using new technology” (62.1%). However, when asked if they “liked taking the test on Moodle”, test takers’ responses were mixed since about half of them (48.9%) responded with no and the other half (50.0%) responded with yes.

The results also suggested an impact on test performance among the groups responding to the items on all issues, except the ‘attitude towards using new technology to take the test’. A prominent finding was that the group that “liked taking the test on Moodle” (*Md* = 21.00, *mean rank* = 95.39, *n* = 87) scored better on the total test than the group that did not (*Md* = 19.00, *mean rank* = 77.40, *n* = 85). However, no statistically significant differences on Kruskal-Wallis Tests were detected among the groups responding to all three items across agreement levels (Table 5.4). Boxplots and detailed tables of these results can be found in Appendices Q (Figures Q14 to Q16, pp. 216-218) and R (Tables R14 to R16, p. 245).

Table 5.4. *Attitude and Resistance to Change: Comparison with Test Scores*

Item	Kruskal-Wallis Test
Q5: Overall, I liked this test-taking experience.	*H*(5, *n* = 174) = 7.06, *p* = .216.
Q15: I liked using new technology to take this test.	*H*(5, *n* = 174) = 5.82, *p* = .324.
Q34: Did you like taking the test on Moodle? (Yes/No)	*H*(2, *n* = 174) = 5.93, *p* = .052.

The theme ‘attitude and resistance to change’ exposed test takers’ attitudes towards experiencing and using technology to take the Moodle-hosted test. A percentage of students (21.3%; *n* = 37) expressed their resistance to change from the traditional way using pen and paper to Moodle tests using the computer. The justification these students gave for their resistance was that they have been used to traditional paper-based testing since childhood in schools. As one Level 4 General male student stated, students reject the idea of computerised testing because they “spent 12 years of schooling using the traditional way so it is hard to accept this mode”. In addition, 1.1% (*n* = 2) talked about the need for more familiarity and practice with this testing mode. Students “need to practice this type of tests a lot” before having to do these tests officially, commented one Level 6

Nursing male student. As further commented by another Level 4 General male student, students' negative attitude towards these tests might change "but only after they get used to them."

5.6. Theme 5: Encountering Technical Issues

Four technology-related issues were classified under this theme including:

- technical problems during the exam;
- network efficiency;
- speed of audio file loading; and
- computer working properly during the exam.

Test takers indicated agreement on a Likert item for each issue. Across all of the items, only a minority of test takers experienced technical problems. Of these issues, 18.4% of the test takers found general 'technical problems during the exam'. When more specific issues were examined, a minority of respondents (0.6% to 6.9%) found the network, audio file load time or the computer itself as problematic.

The results showed that the test performance tending to be affected was that of the groups responding to the items on the issues of 'technical problems during the exam' and 'speed of audio file loading'. Test performance of the groups responding to the items on the issues of 'network efficiency' and 'computer working properly during the exam' did not show such a trend. However, across agreement levels on all items, Kruskal-Wallis Tests revealed no statistically significant differences in the test performance of the respondent groups (Table 5.5). See boxplots and tables of these results in Appendices Q (Figures Q17 to Q21, pp. 219-223) and R (Tables R17 to R20, pp. 245-246)

Table 5.5. Encountering Technical Issues: Comparison with Test Scores

Item	Kruskal-Wallis Test
Q20: There were technical problems during the exam.	$H(5, n = 174) = 1.49, p = .914.$
Q21: The network was efficient and did not slow down while taking the test.	$H(4, n = 174) = 1.67, p = .796.$
Q22: The audio file in the listening loaded quickly.	Overall Test: $H(4, n = 174) = 8.14, p = .087.$ Listening Test: $H(4, n = 174) = 6.50, p = .165.$
Q23: The computer worked properly during the exam.	$H(4, n = 174) = 3.09, p = .544.$

A reason given for resisting the use of technology in assessment is the fear of potential technical failures that may affect test performance. Students provided comments on the issue of 'encountering technical issues'. A small number (*n* = 14) of the respondents described technical issues they fear when being tested using Moodle, including computer malfunction, internet disconnection, and power shutdown. These students' responses towards the Moodle-hosted testing mode was that of resistance "because of worrying that the network stops or the computer malfunctions," as one Level 4 General male student commented. Part of the problem is that "the internet is sometimes slow and there are problems with the network which leads to a waste of time and the need for extra time", as commented by a Level 6 Nursing male student. Of the 14 comments, 6 focused on "the possibility of losing data due to … for example power shutdown" in the case of which "responses to test items will be lost without saving them and the test will have to be repeated", as a Level 6 Commerce male student commented. These comments indicated that students fear losing their marks in the case of not saving or not submitting their answers. As stated by a Level 4 General male student, students consider using Moodle "not to be safe as there might be a power shutdown and students' efforts might go in vain".

To help resolve the issue, in the comments provided by the 8.0% of the test takers, one male student from Level 6 Sciences suggested "strengthening the internet connection to avoid it making problems with the test". Two other male students from Level 6 Sciences recommended "keeping the computers switched on before the test" and checking "that the computer is flawless or does not have any issues to avoid wasting the test time". To ensure they do not run into technical issues during the test, they also suggested "allocating enough time in case anything goes wrong with the computer or network to avoid the negative effects of this".

5.7. Theme 6: Sound and Headphones Quality

This theme included two technology-related issues:

- sound quality of the listening tests; and
- headphones quality.

Test takers indicated agreement on a Likert item for each issue. The two issues targeted by the items were shown to be of concern for test takers. The two issues were problematic for 11.5% to 21.3% of the test takers. The 'sound quality of the listening tests' was found the most problematic, where 21.3% broadly disagreed that the sound quality was good.

The results suggested a tendency towards an impact on test performance among the groups responding to the item on the issue of 'headphones quality'. As shown in Table 5.8, while mean score differences were not statistically significant in the overall test, they were statistically significant with a large effect size in the listening test performance of the groups responding to the item on this headphones quality issue. The neutral and the agreeing groups had the same highest median score of 9.00, but the neutral group (*Md* = 9.00, *mean rank* = 111.22, *n* =25) was higher in the mean ranks than the agreeing group (*Md* = 9.00, *mean rank* = 102.56, *n* =54). However, the agreeing group had a higher number of respondents. If we exclude the neutral group, in terms of agreement and disagreement on the scale, the agreeing group can be said to have scored better on the listening test than the rest of the groups. In the post hoc pairwise comparisons, statistically significant differences with very large effect sizes in the listening test scores were identified between the neutral and strongly agreeing groups and between the agreeing and strongly agreeing groups. The neutral group (*Md* = 9.00, *mean rank* = 111.22, *n* = 25) scored higher than the strongly agreeing group (*Md* = 6.00, *mean rank* = 71.59, *n* = 75). The agreeing group (*Md* = 9.00, *mean rank* = 102.56, *n* =54) scored higher than the strongly agreeing group (*Md* = 6.00, *mean rank* = 71.59, *n* = 75).

Though the ‘sound quality of the listening tests’ was perceived problematic by 21.3% of the test takers, no impact on test performance (overall test and listening test) among the respondent groups was shown. No statistically significant differences in test performance were found (Table 5.6). Boxplots and detailed tables for the two items are provided in Appendices Q (Figures Q22 to Q25, pp. 224-227) and R (Tables R21 and R22, pp. 246-247).

Table 5.6. Sound and Headphones Quality: Comparison with Test Scores

Item	Kruskal-Wallis Test (Dunn-Bonferroni effect size given where significant)
Q11: Sound quality of the listening tests was good.	Overall Test: *H*(4, *n* = 174) = 3.48, *p* = .482. Listening Test: *H*(4, *n* = 174) = 2.72, *p* = .606.
Q24: The headphones worked properly during the exam.	Overall Test: *H*(4, *n* = 174) = 2.09, *p* = .720. Listening Test (all categories): *H*(4, *n* = 174) = 19.01, *p* = .001, *r* = 0.11. Listening (Strongly agree and Neutral) post hoc comparisons, *p* = .006, *r* = 0.40. Listening (Strongly agree and Agree): post hoc comparisons, *p* = .005, *r* = 0.31.

Notes. *r*: Small (.01 to .05); Medium (.06 to .13); Large (*r* ≥.14).

The findings presented under this theme highlight the quality of the sound and headphones in the listening test as issues of concern to some test takers. It should be noted here that due to logistical constraints, the headphones that test takers used for the listening test were not of identical make or models. A minority of test takers (8.0%; *n* = 14) provided comments in relation to the headphones and sound quality. These comments indicated that students liked the idea of doing the listening on their own using headphones given to them for the test. “Every student can listen individually to the listening test through headphones without disturbing others” was the comment given by a Level 4 General male student. Another comment was that “the listening needs to be of a better quality and the headphones too to enable us to concentrate more and listen clearly” (Level 4 Agriculture male student). As such, these comments highlighted the need to be vigilant to avoid the potential negative impact on test performance of poor sound quality.

5.8. Theme 7: Split Screen for Reading and Note-taking

There were two technology-related issues under this theme:

- split screen mode for reading tests; and
- needing to take notes during the test.

For each issue, test takers indicated agreement on a Likert item. The two issues targeted by the items were shown to be of concern for test takers. The majority of test takers (83.9%) broadly agreed that they “liked the split screen for reading tests”, and only a minority (6.3%) indicated it was a problem by their broad disagreement. ‘Needing to take notes’ was more problematic for test takers with 55.1% broadly agreeing that they needed to take notes during the test but were not able to do it on the screen when using Moodle.

The results suggested a trend of an impact on test performance among the groups responding to the item on the first issue ‘split screen mode for reading tests’. This trend was not found for the second issue. However, no statistically significant differences in test performance were detected among the groups responding to either item (Table 5.7). Boxplots and detailed tables for these items are provided in Appendices Q (Figures Q26 to Q28, pp. 228-230) and R (Tables R23 and R24, p. 247).

Table 5.7. *Split Screen for Reading and Note-Taking: Comparison with Test Scores*

Item	Kruskal-Wallis Test
Q8: I liked the split screen mode for the reading tests where the reading texts were on the left side of the screen and the questions were on the right side.	Overall Test: *H*(5, *n* = 174) = 4.169, *p* = .525. Reading Test: *H*(5, *n* = 174) = 6.783, *p* = .237.
Q31: I needed to take notes during the test.	*H*(5, *n* = 174) = 2.22, *p* = .818.

Students provided comments in relation to this theme on the split screen mode for reading and note-taking. A minority of the respondents (4.6%; *n* = 8) expressed their satisfaction with the split screen mode. Students formed a positive opinion of the split screen mode because, as a Level 5 Law male student commented, “it helped them concentrate more”. Other comments indicated that students preferred the paper-based mode since they would be able to highlight, underline, and take notes on important information and on the answers in the texts, which was not possible in the Moodle-hosted exam. Some students like to take notes, highlight, and underline important information to help them concentrate, write meanings of words, and eventually comprehend and answer the questions. Comments given by 20.7% (*n* = 36) of the respondents mentioned the use of these test-taking strategies. To resolve the struggle with having to process and comprehend questions based on reading texts and on listening test questions, one suggestion from students was to include a tool or feature that allows students to use these test-taking strategies.

5.9. Theme 8: Test Mode and Feedback

This theme included the following three technology-related issues:

- Moodle instant feedback;
- testing format preference; and
- which testing format students would perform best on.

For the first issue, test takers indicated agreement on a five-point Likert item. For the second and third issues, test takers selected among two response options in each two-point Likert item. Three quarters (74.1%) of test takers preferred pen on paper tests to that of Moodle and thought they would perform best on paper. However, when asked about ‘Moodle instant feedback’, only 12.0% broadly disagreed with Moodle showing them instant feedback at the end of the test.

The results revealed test scores were higher among the groups selecting Moodle as their ‘testing format preference’ and ‘which testing format they would perform best on’. A Kruskal-Wallis Test detected statistically significant differences with a small effect size in test performance among the groups responding to the ‘which testing format would they perform best on’ item (Table 5.8). The group perceiving their test performance to be best when using online tests on Moodle scored better on the test (*Md* = 22.50, *mean rank* = 95.19, *n* = 36). Nevertheless, in the post hoc pairwise comparisons, these differences were not found to be statistically significant. Kruskal-Wallis Tests found no statistically significant differences in test performance among the groups responding to the ‘testing format’ and ‘feedback’ items (Table 5.8). See Appendices Q (Figures Q29 to Q31, pp. 231-233) and R (Tables R25 to R27, p. 248) for detailed results.

Table 5.8. *Test Mode and Feedback: Comparison with Test Scores*

Item	Kruskal-Wallis Test (Dunn-Bonferroni effect size given where significant)
Q12: I liked that Moodle showed me instant feedback/test results at the end of the test.	*H*(4, *n* = 174) = 1.85, *p* = .763.
Q18: Which format of testing do you prefer? a) pen and paper b) online in Moodle	*H*(3, *n* = 174) = 5.98, *p* = .113.
Q19: I think I would perform best when using: a) pen and paper tests. b) online tests on Moodle.	*H*(3, *n* = 174) = 8.30, *p* = .040, *r* = 0.05; Not significant in post hoc comparisons.

Notes. r: Small (.01 to .05); Medium (.06 to .13); Large (*r* ≥.14).

Moodle marked the test right after students submitted it and showed them feedback in the form of raw scores. A percentage of students (14.4%; *n* = 25) mentioned this feature in their responses to open-ended items. These students said in their open comments that they would prefer Moodle because “it is more accurate and the result shows fast” (Level 4 General male student) and “because of the speed at which marking is done” (Level 6 Science male student). Comparing this to paper-based testing, a Level 6 Science male student noted in the comments that “the negative aspect in paper exams is that results are released late”. The comment from another Level 4 General male student was that “the Moodle test is preferred because when using pen and paper sometimes the teacher who marks the paper can’t understand what students write because the handwriting is not clear”.

On the other hand, 11.5% (*n* = 20) commented that they would prefer paper-based tests “because the teacher marking the test might be lenient and not so strict with the short answer or open-ended questions requiring some writing” unlike Moodle which “is sometimes stricter in marking than the teacher” (Level 6 Science female student). One comment from a Level 4 General male student was that “Moodle marks answers wrong if words are spelled incorrectly”. Furthermore, one of the justifications given by students in their comments for preferring the paper-based mode of testing is that their marks on the Moodle-hosted exam were “unexpectedly” “bad” or not “high”. Added to this, the Moodle testing mode did not enable students to “concentrate more” to get “high marks,” as commented by two Level 5 Science male students.

Interestingly though, with a neutral position about the testing mode in which performance would be best, one Level 6 Commerce male student wrote: “nothing works [to improve my test performance] as my mark will be the same and my effort will be the same”. To justify their view on which format they would perform best on, students indicated that “each has positives and negatives and the testing mode does not dictate the level of performance”, as commented by a Level 6 Science male student. On another note, relating to the feedback functionality on Moodle, there was a suggestion from another Level 6 Science male student to “give room to view the correct answer when seeing the test results”.

5.10. Theme 9: Appropriateness for Testing Purpose

Four technology-related issues that were addressed by the questionnaire items came under this theme, including:

- typing responses in gap-filling items;
- test reflecting true language ability;
- attitude towards taking official Moodle exams (on a Likert scale); and
- attitude towards taking official Moodle exams (Yes/No).

When asked if they ‘liked typing responses to some questions’, more than half of test takers (52.3%) broadly agreed while 31.0% were neutral and 14.4% broadly disagreed. Just half of the test takers (50.0%) broadly agreed the test reflected their true language ability, 27.6% were neutral and 20.1% broadly disagreed. Attitudes towards taking official Moodle exams were investigated via two questions; a yes/no and Likert. The yes/no item revealed that most test takers (75.3%) would not like to take official exams on Moodle where results would be used to take decisions about the level of their language proficiency. However, a lesser percentage of test takers (44.3%) broadly disagreed with the Likert statement addressing this attitude, while the remaining were split between neutral (24.7%) and broad agreement (29.9%).

The trend in the results showed a higher test performance among the groups agreeing to the items on ‘typing responses’ and ‘attitude towards taking official Moodle exams (Yes/No)’. Contrary to this trend, such a correlation on test performance was not identified among the groups responding to the items on ‘attitude towards taking official Moodle exams (on a Likert scale)’ and ‘test reflecting language ability'. Nevertheless, with the exception of the groups responding to the item on 'typing responses', no statistically significant differences in test performance were revealed (Table 5.9). See Appendices Q (Figures Q32 to Q37, pp. 234-239) and R (Tables R28 to R31, pp. 249-250) for detailed results on these issues.

Table 5.9. *Appropriateness for Testing Purpose: Comparison with Test Scores*

Item	Kruskal-Wallis Test (Dunn-Bonferroni effect size given where significant)
Q14: I liked typing my responses for some questions.	Overall Test: *H*(5, *n* = 174) = .955, *p* = .966. Listening Test: *H*(5, *n* = 174) = 4.08, *p* = .538. Language Use Test: *H*(5, *n* = 174) = 13.53, *p* = .019, *r* = 0.08; Not significant in post hoc comparisons.
Q16: I think that the test reflected my true language ability.	*H*(5, *n* = 174) = 7.87, *p* = .164.
Q17: I would like to take such online tests on Moodle as official exams (e.g. mid-terms, finals, Placement Test, Exit Test).	*H*(5, *n* = 174) = 1.17, *p* = .948.
Q35: Would you like to take official exams (like mid-terms, finals, placement tests, exit tests, and so forth) on Moodle to take decisions about the level of your language proficiency? (Yes/No)	*H*(2, *n* = 174) = 4.358, *p* = .113.

Notes. r: Small (.01 to .05); Medium (.06 to .13); Large ($r \geq$.14).

As presented in Table 5.9, for the first issue on preference for 'typing responses in gap-filling items', a Kruskal-Wallis Test showed no statistically significant differences across agreement groups for the overall test scores and the listening test section scores. However, statistically significant differences with a medium effect size were found in the language use test section. The strongly agreeing group scored the highest on the language use test (*Md* = 5, *mean rank* = 100.94, *n* = 25). In the post hoc pairwise comparisons, no statistically significant differences were found among the groups responding to this item. However, as reported in the reliability analysis results in Chapter 4, the gap-filling items were found to be the most difficult items in the Moodle-hosted test. It is interesting to note that the listening test included a lot of multiple-choice items and just a few gap-filling items, while the language use test comprised of only gap-fill questions. The mean scores for all groups on the language use test did not exceed 5 out of 20 while the mean scores of the listening test were higher at 6.9 to 8.1 out of 20. Given that significant differences in test scores were found across agreement groups in the language use test section, a higher proportion of typed response might have impacted on student test performance.

A small number of students commented (7.5%; *n* = 13) on the 'appropriateness for testing purpose'. In these comments six (3.4%) test takers suggested that Moodle be used for official testing only "for the placement and exit tests to have more accurate placement results" (Level 4 General male student). These students considered that using Moodle to do midterm and final exams was inappropriate. The justification they gave for this view was that they found such exams lengthy, "which can affect the eyes and make students distracted" and "because some students have specific strategies to answer questions which is easier in the paper mode" (Level 5 Science female student). Instead, "Moodle tests can be used for any tests during the semester [continuous assessment] but not for midterms or finals" (Level 4 General male student). Midterms and finals are high-stakes for students as they largely determine whether they pass or fail a course of study.

Furthermore, comments from seven (4.0%) respondents indicated that students found Moodle tests appropriate for multiple-choice questions. However, for open-ended test questions requiring typing of responses, the preference was for the paper-based testing mode. Besides, these students commented that they did not like the Moodle testing mode for testing reading and listening skills because of the need to concentrate more. Instead, they considered it appropriate for testing vocabulary and grammar as they do not need that much concentration.

5.11. Theme 10: Time Management

This time management theme dealt with the following two technology-related issues via Likert questions:

- the sufficiency of test timing for all test sections; and
- the presence of count-down timer.

Just over half of the test takers (55.2%) broadly agreed that "test timing was sufficient for all test sections" while a quarter (25.9%) perceived test duration to be problematic. Most test takers (89.1%) broadly agreed that they 'liked the presence of the count-down timer to help submit answers to the test questions within the given test time'. A minority of test takers (4.0%) broadly disagreed.

Comparing test performance to responses to the items revealed that the two issues tended to impact test performance. As shown in Table 5.10, the differences in the total test variable found among the groups responding to the item on the issue of 'sufficiency of test timing' were revealed to be statistically significant with a medium effect size. In the post hoc pairwise comparisons, statistically significant differences with very large effect sizes in the total test scores were identified between the disagreeing and strongly agreeing groups and between the agreeing and strongly agreeing groups. The highest in the rankings was the strongly agreeing group (*Md* = 24.00, *mean rank* = 112.83, *n* = 36) that perceived test timing sufficient for all test sections.

Furthermore, as displayed in Table 5.10, statistically significant differences with a medium effect size on the total test were detected among the groups responding to the item on the issue of 'the presence of the count-down timer'. The strongly disagreeing group, made up of only one respondent, scored the highest in the test (*Md* = 35.00, *mean rank* = 168.50, *n* =1). The strongly agreeing group had the next highest rankings in test scores (*Md* = 20.50, *mean rank* = 92.55, *n* = 110). However, when subjecting the groups to post hoc pairwise comparisons, statistically significant differences among the groups were not detected. Details of these results are in Appendices Q (Figures Q38 and O39, pp. 240-241) and R (Tables R32 and R33, p. 250).

Table 5.10. Time Management: Comparison with Test Scores

Item	Kruskal-Wallis Test (Dunn-Bonferroni effect size given where significant)
Q7: Test timing was sufficient for all test sections.	*H*(5, *n* = 174) = 15.61, *p* = .008, *r* = 0.10. Post hoc pairwise comparisons (Disagree and Strongly agree), *p* = .036, *r* = 0.51. Post hoc pairwise comparisons (Agree and Strongly agree), *p* = .048, *r* = 0.44.
Q10: I liked the presence of the count-down timer to help me submit my answers to the test questions within the given test time.	*H*(5, *n* = 174) = 11.83, *p* = .037, *r* = 0.07; Not significant in post hoc comparisons.

Notes. r: Small (.01 to .05); Medium (.06 to .13); Large (*r* ≥.14).

Where respondents made comments on timing (7.5%; *n* =13), most of these spoke of value they found in the use of the count-down timer shown on the screen. They indicated that the count-down timer was a feature that would make them favor the Moodle testing mode. These students stated that they "liked the presence of the count-down timer to help submit answers to the test questions within the given test time" (Level 5 Science female student), "which saves lots of time and keeps students from getting distracted … [and achieves]… time management for every section on the test" (Level 6 Commerce male student).

5.12. Summary

The questionnaire analyses were aimed to identify the extent to which technology-related construct-irrelevant variance factors associated with the testing mode effect can interfere with test results and, hence, impact the reliability and construct validity of the Moodle-hosted test. By presenting the themes that came up in the analyses, this chapter did identify a number of technology-related variables within these themes. All of these variables were issues of concern to test takers. Despite observing a trend pointing to some of the issues affecting test performance, the study did not detect statistically significant differences in test performance among student groups. Therefore, although students complained about such issues in their questionnaire feedback, the analyses did not find an impact on student test performance of some of these investigated variables. Generally, these issues either relate to the features of the Moodle-hosted testing mode or to the characteristics of the test takers and their personal preferences. These findings will be interpreted together with the results of RQ1 in the coming Discussion Chapter in light of the overall study aim. To summarize these results, Table 5.11 lists the technology-related issues showing which variables tended to affect test performance and which findings were statistically significant.

Table 5.11. *Summary of Results for Technology-Related Variables*

Item #	Themes and Questions	Affects Test Performance	Statistical Significance
	Theme 1: Endurance		
Q28	The test was too long as it consisted of too many sections.[a]	Y	N
Q29	Staring at the computer screen for a long period of time made me lose my concentration[a]	N	N
Q30	Staring at the computer screen for a long period of time caused me eye fatigue[a]	Y	N
	Theme 2: Ease of Use		
Q6	Overall, the test was easy to navigate by moving from one page displaying a subtest to another.[a]	Y	N
Q9	I think the background theme (colours) of the test was appropriate.[a]	Y	N
Q13	Test procedures and instructions given were clear and easy to follow.[a]	N	N

Q25	I was able to successfully log onto Moodle and the online test.[a]	Y	N
Q26	Pictures and graphs were clear.[a]	Y	N
Q27	The font size was NOT appropriate.[a]	Y	N
	Theme 3: Experience with Moodle Tests		
Q3	Your level of familiarity with tests or quizzes on Moodle: (Very familiar; Somehow familiar; A little bit familiar; Not familiar at all)	Y	Y (small effect size)[b]
Q4	Your level of computer-literacy or familiarity with computers: (Very familiar; Somehow familiar; A little bit familiar; Not familiar at all)	Y	Y*
Q32	I have enough experience with technology to take tests on Moodle.[a]	Y	Y (large effect size)
Q33	I will need extra technical training before I am ready to take online exams.[a]	Y	N
	Theme 4: Attitude and Resistance to Change		
Q5	Overall, I liked this test-taking experience.[a]	Y	N
Q15	I liked using new technology to take this test.[a]	N	N
Q34	Did you like taking the test on Moodle? (Yes/No)	Y	N
	Theme 5: Encountering Technical Issues		
Q20	There were technical problems during the exam.[a]	Y	N
Q21	The network was efficient and did not slow down while taking the test.[a]	N	N
Q22	The audio file in the listening loaded quickly.[a]	Y	N
Q23	The computer worked properly during the exam.[a]	N	N
	Theme 6: Sound Quality		
Q11	Sound quality of the listening tests was good.[a]	N	N
Q24	The headphones worked properly during the exam.[a]	Y	Y (large effect size)

	Theme 7: Split Screen for Reading and Note-Taking		
Q8	I liked the split screen mode for the reading tests where the reading texts were on the left side of the screen and the questions were on the right side.[a]	Y	N
Q31	I needed to take notes during the test.[a]	N	N
	Theme 8: Test Mode and Feedback		
Q12	I liked that Moodle showed me instant feedback/test results at the end of the test.[a]	N	N
Q18	Which format of testing do you prefer? a) pen and paper b) online in Moodle	Y	N
Q19	I think I would perform best when using: a) pen and paper tests. b) online tests on Moodle.	Y	Y (small effect size)[b]
	Theme 9: Appropriateness for Testing Purpose		
Q14	I liked typing my responses for some questions.[a]	Y	Y (medium effect size)[b]
Q16	I think that the test reflected my true language ability.[a]	N	N
Q17	I would like to take such online tests on Moodle as official exams (e.g. mid-terms, finals, Placement Test, Exit Test).[a]	N	N
Q35	Would you like to take official exams (like mid-terms, finals, placement tests, exit tests, and so forth) on Moodle to take decisions about the level of your language proficiency? (Yes/No)	Y	N
	Theme 10: Time Management		
Q7	Test timing was sufficient for all test sections.[a]	Y	Y**
Q10	I liked the presence of the count-down timer to help me submit my answers to the test questions within the given test time.[a]	Y	Y (medium effect size)[b]

Notes. Y = Yes; N = No.

[a] Five-point Likert scale: 5 Strongly agree, 4 agree, 3 neutral, 2 disagree, 1 strongly disagree;

[b] Not statistically significant in post hoc pairwise comparisons, significance is 0.05 at 95% confidence interval level;

*Initially small effect size, but very large effect size was found in post hoc;

**Initially medium effect size, but very large effect size was found in post hoc; Effect size criteria: Small (.01 to .05); Medium (.06 to .13); Large (r ≥.14).

As shown in Table 5.11, many of the technology-related variables were perceived by test takers as issues of concern. However, only eight variables were found to have a statistically significant impact on test scores using the Kruskal-Wallis Tests. Effect sizes on Kruskal-Wallis tests were small, medium, or large. These variables were:

- Q3: the level of familiarity with tests or quizzes on Moodle;
- Q4: the level of computer-literacy;
- Q7: the sufficiency of test timing for all test sections;
- Q10: the presence of the count-down timer;
- Q14: typing responses for gap-filling questions;
- Q19: which testing format students would perform best on, that is, pen and paper tests or online tests on Moodle;
- Q24: the headphones working properly during the exam; and
- Q32: having enough experience with technology to take tests on Moodle.

The Dunn-Bonferroni method with adjusted significance levels was used for further post hoc pairwise comparisons (see Section 3.6.3, pp. 47-49, Chapter 3 for details). Four items remained significant after the post hoc tests. Each had a large effect size in the post hoc tests ($r > .14$). These items were:

- Q4: the level of computer-literacy (Table 5.3, p. 71);
- Q7: the sufficiency of test timing for all test sections (Table 5.10, p. 82);
- Q24: the headphones working properly during the exam (Table 5.6, p. 75); and
- Q32: having enough experience with technology to take tests on Moodle (Table 5.3, p. 71).

The remaining four were found not statistically significant following post hoc tests. See items Q3 (Table 5.3, p. 71), Q10 (Table 5.10, p. 82), Q14 (Table 5.9, p. 80), and Q19 (Table 5.8, p. 78), marked with ‘b’ in Table 5.11).

5.13. Conclusion

This chapter has stated the evidence in relation to the sources of measurement error affecting reliability in the context of the Moodle-hosted test. Feeding into the validity argument, there is evidence that strengthens the rebuttal. Evidence has been shown that a number of technology-related issues might have affected students’ test performance. Although reliability estimates are high statistically speaking as established in the RQ1 Chapter 4 (Section 4.2, pp. 52-53), this RQ2 chapter found a high degree of perceived interference by technology-related factors. A small number of these factors were found to significantly impact test performance. Many more factors were perceived to be problematic by test takers. The avoidance of perceived and actual bias is essential in quality testing practices. Some test takers were impressed by the technology features used in the test, but they still expressed dissatisfaction with the Moodle-hosted testing mode. As such, these issues need further consideration.

Given that a number of construct-irrelevant technology-related issues tended to affect test performance, claims for reliability and construct validity of the Moodle-hosted test score-based decisions cannot be fully supported nor warranted. In short, test takers voiced their concerns about the complex process of engaging with the technology. The technology-related issues became the rebuttals to reliability and construct validity claims made in the validation study framework (as described in Section 2.8, pp. 28-29). In the following Chapter 6, these study findings will be discussed in light of the research questions and the validation framework. References to the literature will be made where relevant.

Chapter 6. Discussion

6.1. Introduction

The overall aim of the study was examining the reliability and construct validity of the Moodle-hosted test to structure a validity argument about using it for its intended purpose. The findings reported in the last two chapters will be discussed in this chapter in light of the research questions referring to the literature and the validation framework (Section 2.8, pp. 28-29). RQ1 was: *To what extent can the Moodle-hosted test scores be reliable and valid indicators of the tested construct?* RQ2 was: *To what extent can technology-related construct-irrelevant factors affect the reliability and construct validity of the Moodle-hosted test?*

6.2. RQ1: Reliability and Construct Validity

As reported in Chapter 4, evidence of highly acceptable reliability estimates was established through the Rasch analysis, suggesting that the overall reliability of the test is satisfactory. However, as will be addressed in this section, the results highlighted instances of overly difficult, misfitting and low discriminating items, and unacceptable large error values. Two threats to reliability and construct validity were reported: construct-irrelevance and construct under-representation. These reliability and construct validity concerns were counter evidence suggesting that the test scores might not be reliable and valid indicators of the tested construct.

6.2.1. Threats to reliability and construct validity.

As stated in Section 2.3.5 (Chapter 2, pp. 15-17), although acceptable reliability estimates indicate that the test is systematically testing the construct being measured, we need to identify potential sources of construct-irrelevant variance (as in problematic test items) that can threaten construct validity. The study found that 30% of the items were not reliably testing the target ability. Too many difficult items were beyond the students’ ability levels. More specifically, quarter of the test items (25%; *n* = 15) were overly difficult and had high unacceptable error values. These items were gap-filling items in the Language Use and Listening Subtests. Items found with unacceptable fit statistics and high unacceptable error values point to construct-irrelevant difficulty. These were either misfitting low discriminating items that did not assess student performances reliably as expected by the Rasch response model or overfitting items that did not contribute independently to constructing the target ability. In this study, the person reliability finding of 0.80 could be due to the relatively large number of items that are misfitting (15 items too difficult, or 25%) and probably not discriminating for the specific sample involved. Hence, construct-irrelevance was manifested through finding misfit in the data from the analysis of Rasch-generated fit statistics. As test scores get contaminated by such a threat to construct validity, that is, construct-irrelevance, the test might not have measured what it was supposed to measure. Such construct-irrelevant difficulty suggests that test takers' scores may be invalidly low due to including items that make the construct difficult (Baghaei, 2008). Finding misfit items suggested that test unidimensionality was threatened as such items might have measured construct-irrelevant sub-dimensions other than the single construct of language proficiency. Variations in item fit or item difficulty inform us that more than a single dimension or an underlying construct is being measured by the test instrument (Knoch & McNamara, 2015). This means that test performance did not only reflect language proficiency since test takers found it too difficult to respond to these gap-filling items. Responding to these gap-filling items required test takers to type in their answers, but many did not type any responses to these items. With these items, the test might have measured their typing response ability, which could be a construct-irrelevant sub-dimension that was not intended to be part of the tested construct.

To respond to these items, students required an understanding of the context in the passage, recalling words from their vocabulary, and using correct spelling. One explanation for the difficulty level of these items might be the absence of contextual clues or hints as there was no list of words nor a drop-down list for test takers to select their answers. Hence, this type of task can be cognitively more demanding even in a paper-based mode. The examinees' levels of keyboarding skills might also be a contributing factor in making the task more or less complex. In a study by Russell (1999), examinees with more keyboarding experience were found to have performed better on a computer-based test when answering open-ended test questions. Thus, student performance on constructed-response items in a computerised test might not only reflect their language abilities because it might also mirror their typing ability. This can be taken as a sign of bias towards examinees with better keyboarding skills. However, it should be acknowledged that test performance might be attributed to reasons other than technological factors. For example, as reported in Section 4.4 (pp. 58-60) and summarised in Sections 4.5 and 4.6 (pp. 60-63), the items themselves in this test were difficult as evidenced by the moderate person reliability of 0.80 that could be due to the large number of misfitting and too difficult items (25%; *n* = 15). Person fit statistics also showed that the abilities of 44.9% (*n* = 93) of test takers were not measured reliably. Where items were too difficult and probably not discriminating for students, this will impact the usefulness of test results regardless of the role of technological factors on their test performance.

These results were validated by the findings that were obtained through the comparison of test takers' performance to questionnaire responses. For example, familiarity and experience, and typing responses to the gap-filling test items were reported as construct-irrelevant technology-related variables that interfered with test performance outcomes (see upcoming Sections 6.3.1, pp. 94-96 and 6.3.4, pp. 100-103 of this chapter and Sections 5.4, pp. 70-71 and 5.10, pp. 79-81 in Chapter 5). As such, all of these results complemented each other in suggesting how these gap-filling items signal departure from unidimensionality as they introduce irrelevance in the score variance of the tested construct. Misfit items do not contribute to the measurement of the test construct, suggesting that they act as a threat to construct validity since they point to departure from test unidimensionality (Baghaei, 2008). Construct under-representation was found by identifying gaps between item difficulty and person ability measures along the unidimensional continuum in the Rasch item-person map (Figure 4.1, p. 56). The presence of such gaps indicated that examinees' ability levels were not well-captured by the test. This means that the set of items might have under-represented the test construct and the test needs items ranging in difficulty levels to address the range of ability levels. The study also reported that 44.9% (*n* = 93) of the test takers did not fit the Rasch-expected response model and their measurement error values were unacceptable. This finding indicated that the test might not have measured takers' true language ability reliably. Based on the evidence that the test had construct-irrelevance and construct under-representation issues, it was inferred that the test scores might not be reliable and valid indicators of the target test construct.

The findings of RQ1 are similar to the results reported by McNamara (1990) where overfitting items that did not make independent contributions to measure the test construct were identified, while the majority of the test items contributed to measuring the test construct. The study results also aligned with Akiyama (2001) reporting that misfit and overfit items were identified through Rasch item analysis and that all test items except one overfit item contributed to measuring the test construct. Likewise, Aryadoust and Goh (2009) also reported Rasch-supported evidence on construct-irrelevance and construct under-representation that refuted construct validity claims. This study similarly indicated that the item format (gap-filling or the limited production item types) can affect test performance by introducing variance in the observed test scores. Items requiring production in a listening test have also been reported by Coleman and Heap (1998) and Aryadoust (2012) to be difficult for students. Aryadoust (2013) presenting the results of Differential Item Functioning (DIF) analysis on IELTS suggested that high-ability test takers had an advantage over other test takers. Such students were advantaged by having better summarizing and writing skills when responding to constructed response items in the IELTS listening test besides having better listening skills as well as. Consistent with other studies, it was informative in this study to identify such difficult items through Rasch analysis. Such results imply that item type or format requiring production can produce variations in test performance that might not be attributed to the tested construct per se.

To answer RQ1 on reliability using the concepts adapted in the validity framework from the AUA of Bachman (2005) and Bachman and Palmer (1996) (see Chapter 2, Section 2.8, pp. 28-29), the backing evidence warranted that the Moodle-hosted test had highly acceptable reliability estimates. However, other backing pieces of evidence became the rebuttal as they refuted reliability and construct validity claims by identifying reliability and construct validity concerns, as reported in Chapter 4. The measurement error has been referred to as construct-irrelevant variance, especially that construct-irrelevant technology-related factors such as the familiarity and experience variable interfered with test results. Technology-related issues that examinees encountered in the Moodle-hosted testing mode (Chapter 5 RQ2 results) acted as sources of measurement error and unreliability, which consequently weakened reliability as well as construct validity claims. These reliability and construct validity concerns created bias or lack of fairness for examinees whose performance was affected by the technology-related factors experienced in the Moodle-hosted testing mode. Being considered sources of construct-irrelevant and unreliable variance in test performance, the technology-related issues will be discussed thoroughly in the RQ2 discussion section of this chapter.

6.2.2. Implications of reliability and construct validity threats.

This study responded to the need for investigating technology-enhanced assessment concerns following a validation framework that is focused on the specificities of this testing mode rather than on how it efficiently compares with the paper-based testing mode. In line with this research perspective, the study outcomes highlighting reliability and construct validity threats imply that future research is needed to further investigate the sources of unreliability and invalidity threats in a computerised testing mode. Such validation research can be guided by a sound validation framework “that is not overly preoccupied by efficiency and comparability with paper-and-pencil tests” (Chapelle, 2008, p. 131). As such, in terms of approach, the study implies that the focus of CALT researchers’ validation agenda can be shifted from the traditional comparability perspective to the specific features of CALT that can threaten reliability and construct validity. Such research can examine how particular technology-related variables pertinent to the test mode effect might contribute unreliable and construct-irrelevant variance into test scores.

The study outcomes also imply that reliability is an important element in validation research because the study highlighted important reliability as well as construct validity concerns. Other validation studies on web-based exams have reported incorporating reliability analyses such as a validation study by Chapelle, Jamieson, and Hegelheimer (2003) of a web-based English as a second language test (ESL) and another validation study of a web-based test of ESL pragmalinguistics by Roever (2006). In each of these studies, reliability evidence was gathered and incorporated as a necessary element in the validation process. Taking a similar approach, this study also reported reliability evidence. In our validation task, though obtaining reliability estimates tells us that the test is systematically testing the construct being measured, it is essential to identify potential sources of construct-irrelevant (unreliable) variance that can jeopardize construct validity. The construct represented by the test may change due to the effect of the computerised testing mode (Fulcher, 1999). However, construct-irrelevant variance should not be reflected in the test scores and the test should mirror the construct being tested only. Therefore, a "high reliability coefficient is a necessary (but not sufficient) condition to support a hypothesis of construct validity" (Roever, 2006, p. 235). Researchers need to establish other types of evidence besides reliability coefficients in order to support reliability and construct validity claims.

Variabilities in the test administration context can signal to testers that the construct is not being tested (Fulcher & Davidson, 2007). It should be noted here that contextual factors do not refer to the characteristics of the testing interface tool alone, but they are rather a function of the interaction between the examinees' characteristics and the features and requirements of the test being administered in the online testing interface environment. Like any other testing context, the Moodle-hosted testing environment in this study experienced contextual variables that created construct-irrelevant variance in the test scores. These contextual variables will be discussed in the next RQ2 discussion of results, Section 6.3, in light of the concept of technology-related construct-irrelevant factors.

Furthermore, because of variations in administration procedures, examinees might encounter technology-related issues that can affect their test performance and consequently threaten fairness in testing practices and reliability of test results. Controlling variables related to test administration has been considered as an essential element in Kunnan's (2004) fairness framework. The aim of controlling these variables is to achieve fairness or, in other words, to avoid practices of bias towards test takers of a particular gender, ability, skill, experience (such as technology experience), and so forth. The lack of bias for or against certain examinees has been called "procedural fairness" (Kane, 2010, p. 178). Achieving procedural fairness necessitates the application of standardised testing in which testers are obliged to use the same testing materials and procedures in order to avoid invalidating test score interpretations due to divergent practices. In practice, this means that all examinees should get the same test materials and resources in the form of high quality test delivery devices such as computers or tablets, good quality headphones, updated software programs, and an efficient internet connection. Added to this, test administrators should adhere to standardised test administration guidelines such as time limits, order of subtests administration, inclusive procedures for students with disabilities, and so on.

In sum, the study findings reiterate the call made in the literature (e.g., Brown, 2005; Fulcher, 1999, 2003; Taylor, Kirsch, Jamieson, & Eignor, 1999) for future studies to investigate how particular technology-related variables pertinent to the test mode effect can contribute construct-irrelevant variance into test scores. Such studies should aim to understand these variables so that practitioners can deal with them more effectively in order to eliminate the testing mode effect.

6.3. RQ2: Construct-irrelevant Factors

To address RQ2, this study investigated the extent to which technology-related construct-irrelevant variance factors associated with the testing mode effect can interfere with test results and hence impact the reliability and construct validity of the Moodle-hosted test. This question was investigated through questionnaire survey. The findings revealed that construct-irrelevant variance was present in the Moodle-hosted testing environment as test takers experienced with technology-related issues in the complex process of engaging with the technology. Since technology-related factors affected performance on the test that was delivered in this testing mode, these construct-irrelevant technology-related variables can be said to have been sources of measurement error. The amount of measurement error in the test scores can be explained by the set of technology-related factors. The tested construct is no longer just language proficiency as test performance gets affected by the technology-related encounters. Therefore, the technology-related issues are the rebuttals to reliability and construct validity claims made in the validation study framework that is described in the Methodology (Chapter 3).

More specifically, statistically significant differences in test performance were found among the examinee groups responding to questionnaire items that targeted eight technology-related variables. These factors were shown to significantly interfere with test results as they affected test performance. This part of the discussion will address the following eight variables that were reported in details in RQ2 Results (Chapter 5):

- having enough experience with technology;
- level of familiarity with Moodle tests;
- level of familiarity with computers (computer-literacy);
- headphones quality;
- attitude towards testing format (which testing format students would perform best on);
- typing responses in gap-filling items;
- presence of count-down timer; and
- sufficiency of test timing for all test sections.

Since some variables are connected to each other, they will be combined together in the next discussion sections. First of all, the ‘familiarity and experience variable’ will focus on the first three variables on the list including the levels of technology experience, familiarity with Moodle tests, and computer-literacy. The variables of the presence of the count-down timer and sufficiency of test timing for all test sections will also be combined under one variable labelled ‘test time and length’.

6.3.1. Familiarity and experience.

One of the concerns in computerised online testing is the level of familiarity or experience with technology, the testing interface, and computers. As reported in Chapter 5 (Section 5.4, pp. 70-71), this study has demonstrated that as computer literacy, familiarity with computers and experience of the Moodle-hosted testing environment increased, so did student’s performance on the test to a statistically significant degree.

These statistical findings regarding the issue of the familiarity and experience variable affecting test performance were supported with open comments that were supplied by students on the questionnaires. This issue was mentioned by a small number (3%) of the respondents in their open comments. Students’ comments with regards to this issue highlighted that it can make some students more advantaged than others. As such, the students’ open comments triangulated the statistical evidence that the familiarity variable is a technology-related issue that can have a significant impact on student test performance. Such findings also indicate that the bias resulting from the varying levels of familiarity impacting test performance can threaten test reliability and construct validity interpretations.

As mentioned in the literature review (Chapter 2, Section 2.5.1, pp. 18-19), the importance of the familiarity and experience variable has been addressed by researchers (Eignor, Taylor, Kirsch, & Jamieson, 1998; Fulcher, 1999, 2003; Kirsch, Jamieson, Taylor, & Eignor, 1998; Maycock & Green, 2005; Russell, 1999; Taylor, Jamieson, Eignor, & Kirsch, 1998; Weir, Yan, O’Sullivan, & Bax, 2007). While this study has reported findings that agreed with results of some of these studies, the findings also contradicted results of others.

Findings of research by Fulcher (1999) and Russell (1999) have similarly suggested that the familiarity variable had a significant effect on test performance. Examining the presentation mode effects, Fulcher’s (1999) study found that mean score differences of an ESL placement test were significant on a web-based test, but were not significant on a paper-based test. Similar to this study, Fulcher (1999) contended that factors affecting test performance including computer familiarity can be considered “equity issues” and bias indicators in computerised tests (p. 292). The study findings were also consistent with Russell’s (1999) study, where test takers with more keyboarding experience performed better on open-ended test items of a computer-based test. Therefore, the results of this study have confirmed outcomes reported by other studies (Fulcher, 1999; Russell, 1999) in showing that the familiarity and experience variable can significantly impact test performance.

Unlike the findings of this study, other researchers (Maycock & Green, 2005; Taylor, et al., 1998; Weir, et al., 2007) have not found that the familiarity variable had a significant effect on test performance. Taylor et al.’s (1998) study did not suggest that the lack of prior experience with computers – computer familiarity – affected examinee performance on the computer-based TOEFL. The findings of Taylor et al.’s (1998) study disagreed with the results of this study as no meaningful differences in test scores were identified between candidates familiar and non-familiar with computers. Research by Weir, et al. (2007) also found no connection between computer familiarity and test performance in their examination of the IELTS writing paper-based and computer-based versions. Moreover, research by Maycock and Green (2005) reported that computer familiarity did not have a significant effect on computer-based IELTS scores.

To sum up, this study suggested that test performance can be impacted by the familiarity and experience variable involving the levels of familiarity or experience with technology, familiarity with tests delivered on the Moodle testing interface, and computer-literacy. While this finding is in keeping with results of previous research, it contradicts findings of other research. However, this variable can be an important concern for construct validity of computerised tests because computerised tests should reflect the construct being tested only. This is because “if the test score represents both language ability and computer familiarity,… then valid generalization of test scores across modes is no longer possible” (Sawaki, 2001, p. 42). Furthermore, researchers (O’Sullivan, 2000; Weir, 2005) have emphasized that prior experience and familiarity with tests is one of the examinees’ experiential characteristics that can affect test performance. Consequently, interpretations and decisions made from test results that were impacted by the familiarity variable can be contaminated by this reliability and construct validity threat.

The findings of this study imply that practitioners should evaluate as well as increase student familiarity and experience with the technology, the testing interface and computers that are being used for assessment purposes. These implications are in line with suggestions already made by the existing literature. Practitioners can evaluate this variable by conducting familiarity studies that determine familiarity and experience levels among the test taker population (Fulcher, 2003). They can also amplify the levels of familiarity and experience by giving students access to sample and past test materials. Accessing such test materials can exemplify test task demands (Weir, 2005) and can familiarize examinees with the test format and item and task types (Fulcher, 2003). In addition, practitioners can conduct assessment tutorials to address familiarity concerns (Al-Ani, 2008; Davis, 2015; Taylor, et al., 1999). In such tutorials, candidates can get familiarized with particular test item types and testing software conventions (Davis, 2015). An instance of this practice can be seen in computer-based IELTS where sample test materials and an introductory tutorial are provided to candidates (Maycock & Green, 2005). In short, as an equity and bias-prevention measure, the familiarity and experience issue affecting test performance should be addressed by familiarizing students with the testing interface features and all technology-related equipment prior to the testing event, and by giving them access to sample test materials and tutorials.

6.3.2. Headphones quality.

As mentioned in Chapter 5 (Section 5.7, pp. 74-76), the headphones that students used to listen to the audio files of the Listening Subtest were not identical in their make or models due to logistical limitations. The headphones of 11% of the students did not work properly during the test. The listening test performance of the student groups had statistically significant differences. These findings indicated that the quality of the headphones was a technology-related variable that had a significant effect on test performance.

We infer from these results that the headphones quality can be an issue of concern in this testing mode. This is because the variations in the quality of the headphones can become construct-irrelevant technology-related sources of bias and unfairness in a testing context. Having headphones of good quality is an example of hardware requirements for computerised listening tests. Meeting hardware requirements is an essential asset to satisfy practicality, which is considered one of the principles or qualities of a good useful language test (Bachman & Palmer, 1996; Chapelle, Jamieson, & Hegelheimer, 2003). If variables (such as headphones and other hardware requirements) create inequities between test takers, then procedural fairness (Kane, 2010) is threatened, which will introduce construct-irrelevant variance in the test scores (Fulcher, 2003). In this study the headphones used for the Moodle-hosted test were not standard for all test takers. It was evident in test outcomes, participant feedback and researcher observation that there was an impact on individual test taker’s performance. The finding implies that standardizing and checking hardware such as headphones will enhance procedural fairness.

As mentioned in the literature review Chapter 2 (Section 2.5.3, pp. 20-21), encountering technical glitches during CALT administration, as is the case with the headphones issue in listening tests, is a technology-related variable of concern. The study findings are consistent with Choi, Kim, and

Boo's (2003) testing mode effect research where they found that, compared to the paper-based testing mode, the computerised testing mode significantly affected listening test performance. The findings in this study that hardware quality impacts test takers are also consistent with research by Davis, Janiszewska, Schwartz, and Holland (2016). Their study similarly reported technical issues in the use of headphones in listening to the audio part of the spelling test in the National Assessment Program for Literacy and Numeracy (NAPLAN) in Australian schools.

Furthermore, Hamouda (2013) found that poor quality equipment resulting in distorted sound is one of the physical setting problems that can interfere with students' listening comprehension. Arnold (2000) also noted that "acoustic inadequacies" is one of the factors that lead students to develop anxiety in their processing of the listening input (p. 779). Another study by Yang (2009) reported finding statistically significant differences in test performance among students tested with headphones that had specifications of three sample rates (44 kHz, 22 kHz and 11 kHz) and two sample depths (16 bit and 8 bit). Based on such findings, Yang (2009) has recommended that standardized equipment like headphones may be made mandatory for high-stakes online English language listening comprehension test administration in an EFL context. The researcher also suggested that the standard for quality adheres to digital audio play back specifications of 22 kHz and 8 bit.

The study findings are consistent with the concerns echoed in the literature on the effect of the aural input on listening test performance, and to the researcher's knowledge, no studies have reported disagreeing research results. In other words, studies in this area have reached a consensus that testers need to provide good quality standardized audio equipment to avoid disadvantaging test takers (Geranpayeh & Taylor, 2013).

Besides the quality of the recorded audio input and play-back equipment, the listening test administration conditions such as the acoustics of the testing room are variables that may affect listening test performance (Brindley, 1998). For a better quality in the delivery of the listening test audio, testers should standardize the equipment used (such as headphones) to take advantage of its usefulness to prevent external background noise and testing room acoustic effects (Geranpayeh & Taylor, 2013). Such equipment should also be carefully tested before using them for testing purposes (Brindley, 1998). Hence, practitioners must test the standard devices and equipment provided to examinees to ensure they meet good quality hardware specifications.

To sum up this section, from the findings of this study and as established in the relevant literature, it is essential that as part of test preparations, testing programs ought to provide standardized high quality headphones and other required hardware. When designing a computer-based testing interface, as emphasized by Fulcher (2003), we need to lay out hardware specifications as part of the design process. This is intended to identify and fix potential problems following a problem-resolution approach so that hardware malfunctions do not affect test performance. Overall, ongoing quality assurance requires monitoring and maintenance of standard hardware and software tools for technology-enhanced test administration such as computer screens, keyboards, headphones, audio players and web browser software.

6.3.3. Attitude towards testing format.

Although it is not logically clear how attitudes can affect test performance, the attitude about which testing format students would perform best on, that is, pen and paper tests or online tests on Moodle, was found to have affected performance on the overall Moodle-hosted test. As reported in Chapter 5 (Section 5.9, pp. 77-79), preferring pen on paper over Moodle tests, three quarters (74%) of test takers also thought they would perform best on paper. Statistically significant differences in test performance were found among the groups. Students who thought they would perform best on Moodle tests scored slightly better than those who thought they would perform best on pen and paper tests. In their open comments, test takers highlighted positives and negatives of each testing mode that make them favor a particular testing mode over another. The results imply that test takers' attitude towards the testing mode might become a source of construct-irrelevant variance in test scores as they can affect test performance.

These attitudes can be referred to as psychological characteristics of test takers and can affect their test performance. Besides motivation, psychological characteristics also include personality, cognitive style, affective schemata, concentration, memory, and emotional state (O'Sullivan, 2000; Weir, 2005). As some of these characteristics including motivation are subject to change with time (Weir, 2005), resistance to change to the new computerised mode of testing can be addressed by testers. This means that students' negative attitudes could be addressed by raising students' motivation and acceptance of innovation (online Moodle-hosted test here) with the belief that this can have a positive result in their test performance.

Furthermore, one possible limitation in this study might be that this test was not administered as a high-stakes test since students' marks in their academic program were not affected by their Moodle-hosted test scores. As Green (2013) states, this often happens with trials or field tests as candidates' motivation might inevitably be influenced and it is not clear whether their scores are indicators of their true ability since they might not have taken the test seriously. Overall, the impact of test takers' attitudes on test performance is important to examine. As recommended by Messick (1989), such attitudes are a source of evidence on construct validity because they can be regarded as a potential source of construct-irrelevance.

As mentioned in the literature in Chapter 2 (Section 2.5.4, pp. 21-22), other studies have also investigated students' attitudes towards digital delivery of testing tasks. The results of this study were consistent with other research findings. In line with the results of this study, Singer and Alexander (2017) examined differences in reading texts across digital and print mediums. Their results indicated that 69% of students expected that their comprehension would be better when reading digital texts. Performance outcomes obtained from the comprehension task were inconsistent with students' views. When identifying the main idea of the text, students did not show any differences in their performance across mediums. However, students' recall of key points and other relevant information was better when reading in print. The researchers state that they cannot assume that students' mere preference for reading in a digital environment (attitude) means that they are well prepared to perform well in reading comprehension of digital texts. From such findings, we need to note that students whose attitude indicates preference and perception of performance to be in favor of the computerised testing mode would not necessarily perform better in a digital testing environment than in print or paper-based testing format. In another study that examined test candidates' attitudes to the Fudan English Test, Fan and Ji (2014) also found that attitudinal factors explained a significantly small percentage of test score variance, supporting that test performance can be influenced by personal characteristics including attitudinal factors.

Maycock and Green (2005) similarly found that test takers varied in their attitudes as the preference of 41% of respondents was for the computer-based version of IELTS writing, compared to 35% favoring the paper-based version and 24% indicating no preference. However, no statistically significant effect was found on the test performance for the item probing whether they preferred taking the computer-based test to the paper-based test. In Fulcher's (1999) study, test takers were asked to indicate if they preferred the paper-based testing format or the Internet-based testing format. They were also requested to indicate on which test they would perform best and to nominate which they would choose if given the choice. The findings of this study contradict some of Fulcher's (1999) findings where test taker attitudes had no significant effect on their computer-based test scores. Furthermore, in a study by Stricker, Wilder, and Rock (2004) that assessed acceptance of the computer-based TOEFL among test takers, findings of positive attitudes towards the computer-based TOEFL were reported. These attitudes were also found to moderately correlate with test performance, suggesting they were not an important source of construct irrelevance in test scores. The researchers further argue that acceptance of computer-based tests will increase with technology becoming more common.

Since students in this study answered the questionnaire item asking about this attitudinal aspect after they sat the Moodle-hosted test, their experience of the Moodle-hosted testing conditions might have influenced their views about their test results. Such findings direct researchers' attention to the need to examine test taker experiences and elicit their views to determine the test impact as part of test validation. Impact of tests on test users, especially examinees, is one of the qualities to look for in the constant development of a good language test (Bachman and Palmer, 1996). Consequential validity, test power, and critical language testing (Shohamy, 2007) are the terms that the testing literature puts at the forefront when examining test impact. The voices of the test takers are central to the investigation of test impact issues. Therefore, the finding pertinent to test takers' attitudes towards the testing mode effect sheds light on the need to facilitate the transition of new assessment initiatives like the Moodle-hosted test so that resistance from its affected users including examinees can be properly addressed.

Test impact research in particular should look at students' negative attitude and resistance to educational innovations and changes including new types of assessment delivery. In addition, to better understand the influence of attitudes on language test performance, Fan and Ji (2014) suggest adopting theoretical frameworks as in the Expectancy-value theory (Jacobs & Eccles, 2000) to explore attitudinal factors such as test-taking motivation and success expectation. As they further argue, due to the importance of examinees' attitudes towards computerised tests in construct validation research, practitioners can study patterns of and the reasons behind such attitudes in order to provide intervention measures. Hence, test providers need to allow test takers equal access to test information in order to promote more positive attitudes and acceptance of the computerised testing mode.

6.3.4. Typing responses for gap-filling items.

This study has demonstrated that student test performance can be affected by the task of typing responses for gap-filling items. Statistically significant differences were detected between student groups. Examinees who strongly agreed that they liked typing responses for some questions scored the highest in the Language Use Subtest where all 20 items were gap-filling and required them to type in responses. The reliability analysis results (Chapter 4, Section 4.3, pp. 53-58) also revealed that the gap-filling items were the most difficult items in the Listening and Language Subtests. As discussed in Section 6.2.1 (pp. 88-91), these overly difficult items that constituted a quarter of the test items also had high unacceptable error values. These items were instances of construct-irrelevant variance and introduced construct-irrelevant difficulty, which suggested that including tasks or items that make the construct difficult can produce invalidly low scores. These results were supported with students' questionnaire comments that indicated Moodle tests to be more appropriate for multiple-choice type questions. For open-ended questions requiring typing of responses, students' preference was for the paper-based testing mode. In sum, constructed-response items were found difficult and inappropriate in the Moodle testing mode and introduced construct-irrelevant variance in test scores, which threatened reliability and construct validity interpretations.

As mentioned in Chapter 2 (Section 2.5.2, pp. 19-20), typing is one of the variables that has been under investigation by researchers following the inclusion of test tasks or items that demand typing of responses in computerised testing modes, and hence call upon keyboarding skills or keyboarding proficiency. Test score differences become a source of construct-irrelevant variance if they are attributed to the lack of keyboarding skills among test takers rather than their lack of the tested language construct (Wolfe & Manalo, 2005). This validity threat gets introduced since the ability to use a computer might confound our interpretation of test scores (Taylor et al., 1998). The typing variable is related to the familiarity and experience variable discussed in this chapter (Section 6.3.1, pp. 94-96). Keyboarding skills are also referred to here as analogous to typing skills.

Findings of other studies aligned with the results of this study regarding the issue of typing responses in computerised exams. Hillier (2015) reported students' views through surveys conducted prior, during, and after mid-semester trials on an e-exam system. Among the participating students' views, there was a range of positive and negative perspectives. One of the concerns that were voiced was "typing proficiency" (p. 582) as students who typed their exams in the trials reported that typing would be more time efficient for them and their good typing skills would put them at an advantage. On the other hand, students who hand-wrote their exams in the trials reported they had poor typing skills. This means that students with poor typing skills might just opt out of sitting computerised exams if given the choice and would just sit their exams in the traditional paper-based mode. Coniam (1999) also reported similar findings, where test takers' preference was for a paper-based version of the test when the testing task required them to type in words or phrases. Compared to this, their attitude towards taking a computer-based test was positive when the testing task was limited to just selecting an answer in a multiple-choice type test. Coniam (2006) further argues that examinees' negative views towards taking computer-based tests might not be attributed to computer familiarity and accessibility only but test type (multiple-choice or constructed-response) is also of importance in shaping these views.

Furthermore, Roever (2001) argues that typing speed can be a serious source of measurement error variance when examinees have to type in responses to constructed-response item types. With 60 seconds per item, examinees were able to complete 99% of each of the two multiple-choice sections of the test. On the other hand, although they were given 90 seconds per item, they could only complete 83% of the section in which they had to type in brief responses. Roever's (2001) research also raises concerns about what impact the varying levels of keyboarding skills including typing speed can have on test performance. In comparison, similar concerns about examinees' varying levels of handwriting speed and handwriting readability might exist as well. Assessing writing with paper-based tests can introduce bias nowadays as students do more word-processing than handwriting in the academic language use domain (Chapelle & Douglas, 2006). Therefore, one view would be that "differences in handwriting skills may now be a bigger barrier to fair and valid assessments than differences in word-processing skills, and word-processing skills are probably more construct-relevant as most university writing will use a word processor" (Barkaoui, 2014).

On the other hand, the findings regarding the typing issue disagreed with results of other studies addressing typing, keyboarding skills, and familiarity and experience with computers and technology. As discussed in this chapter (Section 6.3.1, pp. 94-96), researchers (Maycock & Green, 2005; Taylor, et al., 1998; Weir, et al., 2007) found that the familiarity variable had no significant effect on test performance. Studies have focused on comparing test performance across the two modes of test delivery, paper-based and computer-based. For example, in the context of writing tests delivered in the two testing modes, Weir, et al. (2007) found no significant differences in IELTS writing test performance across modes. Also, as Barkaoui's (2014) research showed that the keyboarding skill had a significant but a small effect on TOEFL-iBT writing task scores, it was concluded that test performance on these writing tasks mainly depends on the test taker's English language proficiency and writing ability. However, as students nowadays engage in language uses through computers in their academic contexts, Barkaoui (2014) argues that the language test constructs of the TOEFL-iBT writing tasks may need to be redefined to reflect keyboarding skills as part of the construct, rather than considering them as construct-irrelevant. The results reported by these studies indicate that keyboarding skills might not be as interfering as other studies have argued, and thus, performance on tasks requiring typing of responses in a computerised testing mode can be indicative of the tested construct. Hence, as argued by researchers (Barkaoui, 2014; Chapelle and Douglas, 2006), keyboarding skills might need to be part of the tested construct and might not be considered construct-irrelevant since students are required to employ such skills to perform well at university study.

We can sum up here that as examinees come to the testing session with varying typing ability levels, there could be fluctuations in their test performance that are attributable to the varying typing ability levels rather than their language proficiency. In this case, construct-irrelevance threatening reliability and construct validity gets introduced. Thus, we cannot take it for granted that our test score interpretations and decisions are reliable and valid indicators of the tested language construct.

The only exception to this would be when such typing or keyboarding skills are considered as part of the tested construct, which makes them construct-relevant.

In light of these results on the typing issue affecting test performance, practitioners need to ensure that test takers develop their keyboarding skills before they are tested via computerised testing modes. Keyboarding skills are mandatory for writing efficiently, leading to academic and professional success (Chapelle & Douglas, 2006). To eliminate the effect of weak keyboarding ability levels, we reiterate the same recommendations made about raising students' familiarity with technology (Section 6.3.1, pp. 94-96). By providing students with sample materials (Weir, 2005) and assessment tutorials (Davis, 2015; Taylor, et al., 1999) before the testing event, they can be familiarized with demands of the testing interface features and all technology-related equipment as well as test item types. From a fairness perspective, if feasible, as recommended by researchers (Russell, 1999; Wolfe & Manalo, 2005), testers might also consider allowing examinees to choose between handwriting and typing their writing test responses.

6.3.5. Test time and length.

As part of the limitations in estimating reliability, a number of factors including test length would contribute to a test reliability estimate. A longer test generates more pieces of information about the tested construct and therefore may yield a higher reliability estimate (Green, 2013). Given that the test reliability estimate was high and the test was lengthy (60 items), the questionnaire analyses results addressed whether student test performance had been affected by the test time and length. Two variables were examined: sufficiency of test timing and presence of the count-down timer feature. Test timing is considered here to be dictated by and connected to the test length variable.

As reported in Chapter 5 (Section 5.11, pp. 81-83), statistically significant differences were found in test performance when comparing levels of test takers' agreement on the sufficiency of test timing. Those who agreed performed better on the test. The findings indicated that test performance was significantly affected by the construct-irrelevant variable of the sufficiency of test timing, particularly for poorly performing students. The presence of the count-down timer feature that displayed the time remaining on the computer screen (see Chapter 5, Section 5.11, pp. 81-83) was intended to help students manage their time and submit their answers to the test questions within the allocated test time. When comparing test performance to test takers' agreement on the presence of the count-down timer, statistically significant differences were found. These results were supported by students' comments in the questionnaire. These comments (7%; *n* = 13) indicated that the count-down timer was a feature that made them favor the Moodle-hosted testing mode over the paper-based testing mode. On the whole, the results reported in Chapters 4 and 5 suggest that test performance might have been affected by the construct-irrelevant variable of test time and length.

As mentioned in Chapter 2 (Section 2.5.5, pp. 22-24), a number of other studies have addressed test time and length. The results of this research agreed with the findings of Yamamoto's (1995) study reporting the effect of TOEFL test time and length, where it was found that a small number of test takers were affected by test speediness as they became confounded by the test time limit. A small change in test duration from 55 or 60 minutes to 50 minutes made a difference, where 20% resorted to a random guessing response strategy with the lack of time to respond. In addition, Yamamoto (1995) found that the last 20% of the test did not reflect test takers' true language abilities since they were affected by test speediness after finishing 80% of the test. Similarly, research by Hale (1992) on the *Test of Written English* found that student test performance was significantly higher by about 1/4 to 1/3 point (on a 6-point scale) under the 45-minute test condition than the 30-minutes test condition. Likewise, comparing 15 and 30 minutes testing time conditions, Crone, Wright, and Baron's (1993) research found that giving students more time on the SAT II writing task resulted in significantly better test scores. Consistent results were also reported by Powers and Fowles's (1996) research, where it was found that allowing more test time positively affected test performance, where examinees performed significantly better on a 60-minutes GRE writing essay test than on a 40-minutes version. To wrap up, the findings of these studies agreed with this study's results suggesting that examinee test performance can be affected by the test time and length variable as test performance differences can occur due to test time and length limits.

On the other hand, Knoch and Elder's (2010) research disagreed with the results of this study, showing that examinees' scores on a writing test were not significantly different under short (30 minutes) and long duration (55 minutes) conditions. Their research suggested that the time variable had no significant effect on student test performance. Likewise, Ghanbari, Karampourchangi, and Shamsaddini (2015) concluded that the time pressure variable was a non-linguistic factor and had no effect on writing test performance. Kroll's (1990) research also looked at performance on 60-minute timed essays versus take-home essays written over an extended period of 10-14 days and found a small but insignificant difference in the scores. In other research by Livingston (1987), the test time limit tended to affect the test scores of the more proficient students by around half a point (on the 2 to 12 scale) and essay test scores slightly increased (with a small effect) by raising the time limit from 20 to 30 minutes. Unlike this study, these researchers provided counter evidence reporting small or insignificant effects of the test time and length variable on test performance.

To sum up our discussion of these findings, while the acceptable reliability estimate of the Moodle-hosted test might have been elevated by its test length, the presence of the count-down timer and sufficiency of test timing were identified as two construct-irrelevant variables that significantly impacted test performance. This finding means that the test might not reflect students' true ability levels if the allocated time is insufficient for such a lengthy exam. Given the conflicting findings reported by this study and other relevant studies on the impact of the test time and length variable on test performance, the effects of timing tests and providing count-down timers need further investigation in the context of validation research of computerised exams. It is likely that duration is not a determining factor alone. It needs to be set in light of other factors such as difficulty of the test and familiarity of the testing interface.

The findings of this study imply that practitioners should address the effect of the construct-irrelevant variable, test time and length, by ensuring the test time allotments are sufficient for the test population in their computerised testing context. In practice, answering test questions within the allotted time limits is a test time management skill that students need to master to successfully get satisfying results on exams including international language proficiency exams such as the IELTS (https://www.ielts.org/about-the-test/sample-test-questions) and TOEFL (https://www.ets.org/toefl/ibt/prepare). Because timed testing is endemic in the testing world, preparing for language proficiency exams means that candidates must be familiar with working within time limits. Even with the presence of the count-down timer to help students manage their test time, practitioners should bear in mind that allocating a sufficient amount of time for the test in the first place is an essential factor that can affect student test performance.

These were the technology-related variables that were found to have significantly affected test performance in the Moodle-hosted test. The study also found that a number of technology-related variables did not affect test performance significantly. These non-significant factors were highlighted as issues of concern including layout and scrolling features, note-taking and text highlighting features, and eye fatigue. For brevity's sake, these variables cannot be discussed any further in this chapter. However, it should be acknowledged that regardless of statistical significance, variables may interact in the testing environment. A valid testing environment would involve a number of factors working individually and in combination such as good quality headphones, interface features such as the split screen mode, count-down timer, clear text and layout, and so on.

In sum, technology-related issues that significantly affected test performance were highlighted in the study. Table 6.1 lists the literature sources that were in agreement and disagreement with each of the study findings, as discussed in this section. The suggestions or advice made in this chapter for practitioners to address these issues will be reiterated as implications and recommendations in the Conclusion Chapter.

Table 6.1. *Literature Agreeing and Disagreeing with Study Findings*

Finding/issue affecting test performance	Citations in agreement	Citations in disagreement
Familiarity and experience	Fulcher, 1999; O'Sullivan, 2000; Russell, 1999; Sawaki, 2001; Weir, 2005	Maycock & Green, 2005; Taylor, Jamieson, Eignor, & Kirsch, 1998; Weir, Yan, O'Sullivan, & Bax, 2007
Headphones quality	Arnold, 2000; Brindley, 1998; Choi, Kim, & Boo, 2003; Davis, Janiszewska, Schwartz, & Holland, 2016; Fulcher, 2003; Geranpayeh & Taylor, 2013; Hamouda, 2013; Yang, 2009	No disagreeing studies have been found
Attitude towards testing format (which testing format students would perform best on)	Fan & Ji, 2014; Messick, 1989; O'Sullivan, 2000; Weir, 2005; Singer & Alexander, 2017	Fulcher, 1999; Maycock & Green, 2005; Stricker, Wilder, & Rock, 2004
Typing responses for gap-filling items	Coniam, 1999; 2006; Hillier, 2015; Roever, 2001; Wolfe & Manalo, 2005	Barkaoui, 2014; Maycock & Green, 2005; Taylor, et al., 1998; Weir, et al. 2007
Test time and length	Yamamoto, 1995; Hale, 1992; Crone, Wright, & Baron, 1993; Powers & Fowles, 1996	Knoch & Elder, 2010; Ghanbari, Karampourchangi, & Shamsaddini, 2015; Kroll, 1990; Livingston, 1987

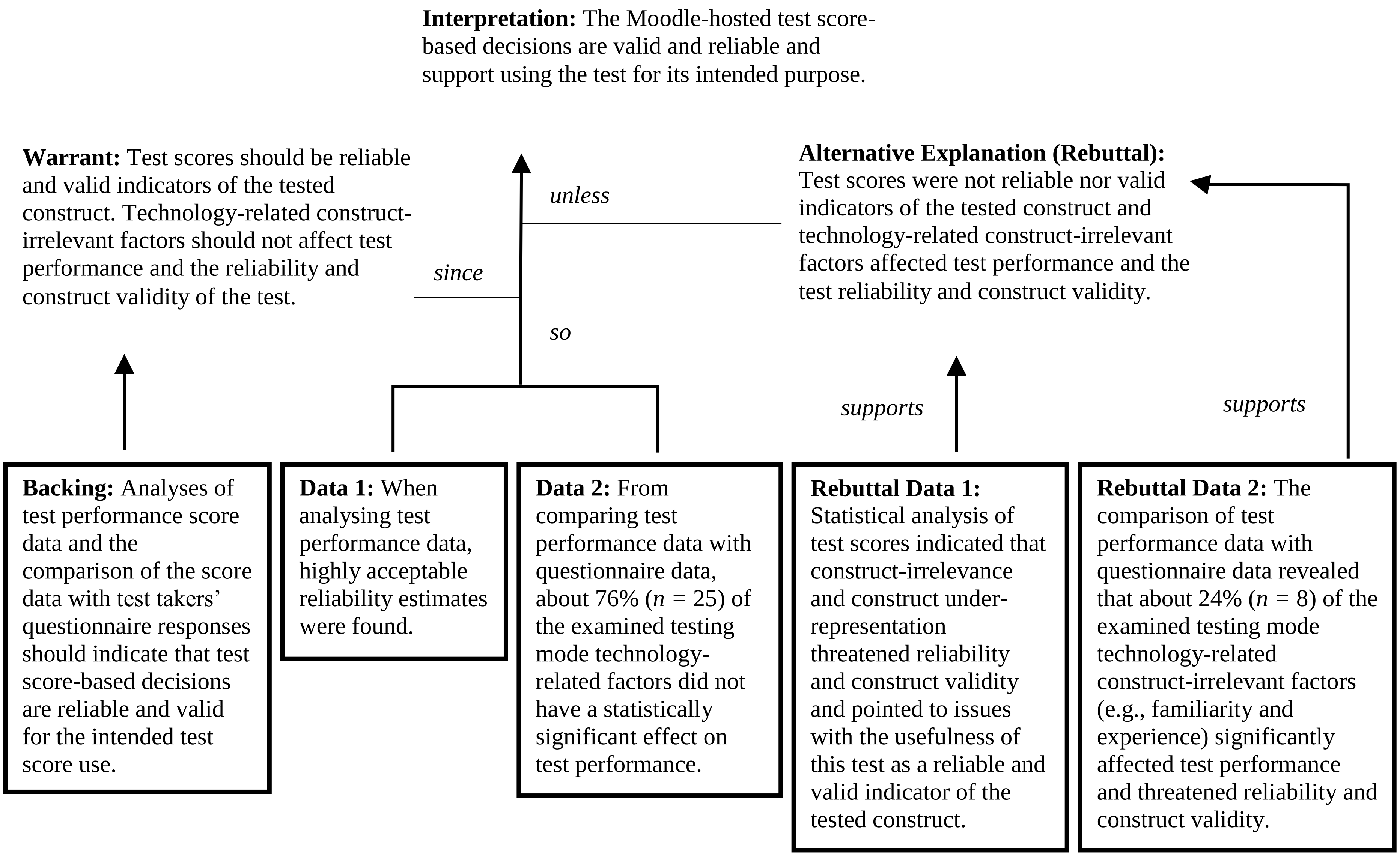
The next section brings the findings that are discussed in the RQ1 and RQ2 discussion sections under one umbrella to discuss the results in light of the validity argument.

6.4. Discussion of Validity Argument

The results reported in Chapters 4 and 5 and discussed under the RQ1 and RQ2 discussion sections in this chapter can be combined to formulate the validity argument about the reliability and construct validity of the Moodle-hosted test. Applying the AUA concepts and principles (Bachman, 2005; Bachman & Palmer, 2010), the structure of the validity argument for the Moodle-hosted test is illustrated in Figure 6.1. To process Figure 6.1, it is recommended to start by reading the Interpretation statement on the top. The rest of the information displayed in the figure is based on whether the interpretation is held true or not. The interpretation is held true since Data 1 and Data 2 provide Backing evidence as a Warrant, unless Rebuttal Data 1 and Rebuttal Data 2 provide an

Alternative Explanation. This validity argument is structured based on Bachman’s (2005) explanation of how to employ Toulmin’s (2003) argument structure in validity arguments. As outlined in the AUA (Bachman, 2005), rebuttals or counterclaims to the test intended interpretation can be considered potential alternative explanations for test performance. These rebuttals or alternative explanations for test performance are viewed in the validity argument as sources of measurement error affecting reliability. Variations in test takers’ attributes or in test characteristics can lead to such alternative explanations or rebuttals and negatively affect the validity of the intended test interpretation.

As Figure 6.1 shows, warrants of reliability and construct validity claimed in the validation study framework (Section 2.8, pp. 28-29) should be refuted because reliability and construct validity issues found in the Moodle-hosted testing environment became the rebuttals or alternative explanations for test performance. Evidence of construct-irrelevance and construct under-representation threatening reliability and construct validity indicated issues with the test usefulness as a reliable and valid indicator of the tested construct. These test characteristics were alternative explanations for test performance. The evidence also suggested that technology-related variables significantly affected examinees’ test performance. These variables can be considered construct-irrelevant factors because they interfered with test results although they were not intended to be components of the tested construct. The technology-related variables significantly affecting test performance include: familiarity and experience; typing responses for gap-filling items; headphones quality; attitude towards testing format (which testing format students would perform best on); and test time and length (including test time sufficiency and count-down timer). There were other technology-related factors insignificantly affecting test performance including: layout and scrolling features; note-taking and text highlighting features; and eye fatigue. Each individual significant factor has been discussed in-depth separately in this chapter. All of these factors still need to be researched in future studies to provide further evidence.



Interpretation: The Moodle-hosted test score-based decisions are valid and reliable and support using the test for its intended purpose.

Warrant: Test scores should be reliable and valid indicators of the tested construct. Technology-related construct-irrelevant factors should not affect test performance and the reliability and construct validity of the test.

Alternative Explanation (Rebuttal): Test scores were not reliable nor valid indicators of the tested construct and technology-related construct-irrelevant factors affected test performance and the test reliability and construct validity.

Backing: Analyses of test performance score data and the comparison of the score data with test takers' questionnaire responses should indicate that test score-based decisions are reliable and valid for the intended test score use.

Data 1: When analysing test performance data, highly acceptable reliability estimates were found.

Data 2: From comparing test performance data with questionnaire data, about 76% (*n* = 25) of the examined testing mode technology-related factors did not have a statistically significant effect on test performance.

Rebuttal Data 1: Statistical analysis of test scores indicated that construct-irrelevance and construct under-representation threatened reliability and construct validity and pointed to issues with the usefulness of this test as a reliable and valid indicator of the tested construct.

Rebuttal Data 2: The comparison of test performance data with questionnaire data revealed that about 24% (*n* = 8) of the examined testing mode technology-related construct-irrelevant factors (e.g., familiarity and experience) significantly affected test performance and threatened reliability and construct validity.

Figure 6.1. Structure of the validity argument about the Moodle-hosted test.

These factors related either to test takers' attributes (e.g., variations in familiarity and experience) or to the test characteristics (e.g., variations in administration procedures as in headphones quality, and variations in task/item difficulty as in gap-filling items requiring typing response). With these variations becoming alternative explanations or rebuttals and sources of measurement error affecting reliability, the validity of the intended test interpretation gets negatively affected (Bachman, 2005). In this study, the identification of the technology-related construct-irrelevant factors as potential sources of measurement error through established evidence has led to answering the research questions set in the framework. Therefore, in light of the established evidence, we argue that the technology-related construct-irrelevant measurement error variance found in the Moodle-hosted test scores can result in unreliable and invalid score-based interpretations and decisions about student language proficiency.

In our overall discussion of the validity argument, we note that there were warrants backed up with evidence of the highly acceptable reliability estimates and the 76% of the examined technology-related factors not significantly affecting test performance. However, these warrants were outweighed by the rebuttals of reliability and construct validity threats attribtuable to the testing mode effect technology-related construct-irrelevant factors. As such, the validity argument in this study was supported with "negative evidence" as well as "positive evidence" (Chapelle, Jamieson, & Hegelheimer, 2003, p. 411; Wang, Choi, Schmidgall, & Bachman, 2012, p. 603). As argued by Kane (2012), on the grounds of obtaining positive and negative evidence, we cannot fully justify test use. However, this does not mean that the test should not be researched nor developed further. On the contrary, validation research is an ongoing process that always sets directions for future research because tests can be seen as "provisional, work-in-progress, … experimental ... [and even as] research tools whose outcomes will help enrich our understanding of the nature of language proficiency so we can develop better tests in the future" (Taylor & Geranpayeh, 2011, p. 94).

This study has presented evidence of a high degree of perceived interference by technology-related factors as many of them were perceived to be problematic by test takers. Only a small number of these factors was discussed in this chapter as they were found to significantly impact test performance. It should be acknowledged that the discussed findings were based on perceived rather than actual bias, meaning that the bias was indicated by the perceptions of the test takers compared to their test results. The study findings highlight issues of fairness or rather bias and lack of fairness in testing practices since unreliable and invalid test results can impact students' lives. By conducting validation research like this study that sought reliability and construct validity evidence, such impact and testing mode effect issues can be identified to ensure that bias does not get introduced. Measures aimed to enhance reliability and construct validity can be put into action in order to make the most of online testing capabilities. As recommended in this study, such measures can involve provision of standardized hardware such as headphones of the same model and make, especially that the headphones quality was one of the variables significantly affecting test performance. Consistency in online delivery mechanisms should be ensured because the lack of standardization in testing conditions can potentially turn to be a potential source of error. Hence, agreeing with Kunnan's (2004) fairness framework, the study results support the view that test administration procedures should be standard so that bias does not get introduced. As already mentioned, full implications and recommendations for practitioners made from this study will be addressed in the Conclusion Chapter.

As mentioned in Chapter 2, the bulk of the literature on the testing mode effect has focused on conducting comparative studies that compared the two testing modes, paper-based and computer-based, by looking at a number of relevant variables. For example, to address validity threats of computerised testing tasks, such comparative studies investigated the relationship between the computer familiarity variable and test performance across modes (Eignor, et al., 1998; Kirsch, et al., 1998; et al., 2007; Taylor, et al., 1998). Contrary to the focus of the cross-mode comparative studies, this study endeavored to fill in the gap in the literature, which is the need to investigate validity aspects that are idiosyncratic to the features of the testing mode (Chapelle, 2008), that is, the Moodle-hosted computer-assisted web-based testing mode. Conducting validation research on features idiosyncratic to the computerised testing mode can enhance our understanding of the testing mode effect and can highlight important validity concerns for the testing community, as found in this study. Finally, applying the validation framework has successfully led to answering the research questions on reliability and construct validity and to ultimately structure the intended validity argument about the Moodle-hosted test.

In the next Conclusion Chapter, an overall summary of the study findings will be presented in light of the validity argument. The chapter will also lay out the study significance and contribution to knowledge, and the implications and recommendations for practice and future research. Study limitations will also be acknowledged.

Chapter 7. Conclusion

7.1. Introduction

The overall aim of this study was to present a validity argument about using a Moodle-hosted test for its intended purpose by examining reliability and construct validity. Chapter 4 reported findings in relation to the first research question examining the extent to which test scores can be reliable and valid indicators of the tested construct. Chapter 5 presented results in relation to the second research question investigating the extent to which technology-related construct-irrelevant factors can affect the reliability and construct validity of the test. In Chapter 6, all of these findings were discussed in light of the relevant literature and served as evidence in the validity argument. In this chapter, an overall summary of the study findings will be presented in light of the validity argument. This chapter will also highlight the significance of the study as well as the implications and recommendations for practice and future research. The study limitations will be acknowledged.

7.2. Overall Summary of the Findings

The study examined the extent to which technology-related construct-irrelevant factors can interfere with examinees’ performance and consequently pose a threat to test reliability and construct validity. This testing mode effect was investigated using a case study of administering and validating a technology-enhanced English Language Proficiency Exit Test. The test was administered on Moodle to a group of EFL students (*N* = 207) at Sultan Qaboos University in Oman. The validity argument was backed up with empirically established evidence on reliability and construct validity from the score data of this test and from post-test examinees’ questionnaires.

As explained earlier in this thesis, the study followed an argument-based evidence-supported validation research framework, which was formulated based on the principles of the Assessment Use Argument (AUA) framework of Bachman (2005) and Bachman and Palmer (2010). Using a mixed-method study design, multiple sources of evidence (Kane, 1992) were sought to support the conclusions reached in the study. To achieve the overall study aim, the study was guided by two research questions (RQ1 and RQ2) as follows:

- RQ1: To what extent can the Moodle-hosted test scores be reliable and valid indicators of the tested construct?
- RQ2: To what extent can technology-related construct-irrelevant factors affect the reliability and construct validity of the Moodle-hosted test?

As reported in Chapter 4, to address RQ1, the test score data were analysed employing Rasch statistical item analysis. To address RQ2, quantitative and qualitative types of evidence were established from statistically and thematically analysing the test taker's questionnaire responses and from comparing these responses to test performance data (see Chapter 5). Pieces of evidence established in this study were used to structure the validity argument that is intended to be disseminated to stakeholders at the study context.

Based on the conclusions drawn from the multiple sources of evidence, the validity argument presented in details in Chapter 6 (Section 6.4, pp. 107-111) can be phrased by addressing the research questions as follows. Since negative evidence pinpointed reliability and construct validity concerns, the Moodle-hosted score-based decisions cannot be reliable nor valid. As reported in Chapter 4, although there were warrants of highly acceptable reliability estimates, strong rebuttals refuted reliability and construct validity claims stated in the validation framework. These rebuttals were identified by finding two threats to reliability and construct validity: construct-irrelevance and construct under-representation. Such reliability and construct validity concerns suggested that the test scores might not be reliable and valid indicators of the tested construct.

Furthermore, student test performance did not only reflect language abilities being measured but it also echoed the testing mode effect as test performance was affected by technology-related factors that were irrelevant to the tested construct (Chapter 5). This argument mirrors what Brown (2005) articulated about the variance in test performance being, to a great extent, a measurement error variance. As the technology-related construct-irrelevant variables act as sources of measurement error, the error variance can be attributable to the testing mode effect. Supporting backing evidence established from this empirical study implied that reliability and construct validity were threatened by the testing mode effect represented by the construct-irrelevant technology-related issues. With this evidence, we can argue against making decisions about student language abilities using the scores of this test. This is because when test results are affected by such issues, decisions might be unreliable and invalid interpretations of student language proficiency. Identifying a testing mode effect highlights bias and lack of fairness issues (Sections 6.2.2, pp. 91-93 and 6.4, pp. 107-111). In sum, the study findings did not support the use of the test for its intended purpose. Despite reaching these negative outcomes in the form of a validity argument against test use, this study is of significance as it contributes to knowledge in a number of ways.

7.3. Significance of the Study

Overall, this study contributes knowledge about the testing mode effect in the Moodle-hosted testing environment in the form of a validity argument about using the test for its intended purpose.

It presents significant, argument-based and evidence-supported implications on computerised testing practices and relevant validation research. The study responds to the concerns raised in the literature about the potential effect of technology-related issues on test performance (Chapelle & Douglas, 2006; Fulcher, 2003). A set of implications and recommendations were put forward for testing practitioners and other researchers, as given in the next section. The optimal outcome is to contribute some guidelines that can be useful for creating, developing, implementing, and researching large-scale high-stakes tests on Moodle, other course management systems, or any other test delivery technologies. Such guidelines are intended to achieve reliable and valid decisions based on test scores. Since the guidelines were reached based on the findings of an empirical investigation, the study addresses the need for such guidelines, which was identified as a gap in the literature (Fulcher, 2003). Furthermore, this study contributes to the limited literature on computerised assessment in educational contexts in Oman and at Sultan Qaboos University in particular (Al-Ani, 2008; Al-Hajri, 2011; Najwani, 2013). It also addresses the need to evaluate the increasing use of the Moodle platform for assessment purposes in the study context language programs and the role it plays on student performance.

Added to this, the argument-based evidence-supported validation framework (Section 2.8, pp. 28-29) that was employed in the study functioned as a pragmatic tool used to articulate a validity argument. From study design to data collection and analysis procedures, the framework proved to be useful in providing backing evidence as warrants and rebuttals in support of the validity argument about test use. The successful application of this validation framework contributes to the rising body of validation research focusing on the use of technology for language testing and assessment. For policy-makers in the study context, this study is significant in that it has implications as a first research attempt to examine the effect of administering a web-based Moodle-hosted test intended to be used for high-stakes purposes using specific technology-enhanced testing interface features. The validity argument should prove to be useful in articulating concerns pertinent to using other Moodle-hosted tests that could lead to detrimental decisions about students' study paths. Future studies can further identify issues with this testing mode in large-scale high-stakes settings. Through such studies, policy-makers become better informed about delivering computerised tests that are justifiably fair to students. In light of this significance, the following section presents study implications and recommendations for practice and future research.

7.4. Implications

Based on the study findings (Chapters 4 and 5) that are discussed in Chapter 6, the following implications and recommendations for testing practitioners and researchers can be made.

7.4.1. Familiarity and experience.

When introducing a new digital medium for test delivery, practitioners need to evaluate the extent of their students' familiarity and experience with the particular technology to be used. This should include familiarity with all of its test delivery features such as its navigation system, layout features, and tools. Being technology-savvy in general does not necessarily mean that students would perform better on computerised exams. However, being familiar with the particular features of the computerised testing system can enhance students' acceptance and reduce their resistance towards the new testing mode. The present study reported that experience with Moodle tests interfered with test performance (Section 5.4, pp. 70-71 and Table 5.3, p. 71). As outlined in Section 6.3.1 (pp. 94-96), it is worth providing sample test materials and tutorials to train test takers in the use of the interface so that any interference of the familiarity variable on test results can be eliminated. In an introductory phase of a new technology, it is preferable that students are given the choice of a testing mode (paper-based or technology-based) in order to accommodate for their preferred exam-taking styles and preferences. Future research still needs to look into the effect of improved keyboarding skills on test performance in any particular context.

7.4.2. Typing responses for gap-filling items.

The inclusion of constructed-response items in a computerised testing mode should not be assumed to be equivalent to the traditional paper-based mode. Such items should be carefully planned as the response format in the computerised testing mode involves a typing rather than handwriting activity, which can be challenging for some examinees. As reported in Section 5.10 (pp. 79-81) and Table 5.9 (p. 80), typing responses for gap-filling items was shown as a factor interfering with test performance. Again as recommended in Section 6.3.4 (pp. 100-103), prior to taking such exams, test takers need to be exposed to items of this type to increase their familiarity and to get them used to the new response format. Test writers should follow the common standards for providing blanks of the same length for missing information in a passage, for example, so that varied lengths do not give examinees hints of the answer length. Alternative answers and acceptable variations in spelling (or even acceptable misspellings) should be keyed into the scoring algorithms of the testing interface so that all examinees are treated in the same way when the system marks their entries. Researchers should further examine the effect of including gap-filling items in computerised exams and how keyboarding skills can contribute to test performance on such items.

7.4.3. Headphones quality.

Bearing in mind the practicality and impact aspects in test evaluation (Bachman & Palmer, 1996), careful consideration of available resources is essential when planning to have high-stakes large-scale testing using technology. As recommended in Section 6.3.2 (pp. 96-98), practitioners should consider the provision of standardized testing hardware and software tools. For instance, headphones of the same make and model need to be provided so that examinees use headphones of standard features in listening exams. This recommendation is supported by finding that the quality of the headphones used in the study interfered with test performance (Section 5.7, pp. 74-76 and Table 5.6, p. 75). In addition, in the case of technical failures during exam sessions, an action plan involving trained personnel such as test administrators and technicians should be a high priority since technical issues can interfere with test performance outcomes. Overall, providing standardized hardware and software tools and following consistent exam procedures come in the interest of achieving fairness.

At the study context in particular, we need to carefully outline the computer specifications suitable for testing purposes. As described in Chapter 3 (Section 3.5.1, pp. 38-42), in order to enhance test security, the Safe Exam Browser application was used in addition to limiting test attempts and setting passwords to access tests (Al Nadabi, 2015). However, to enhance test security further and to minimize cheating, it is also important to redesign the computer laboratory layout. Updates to software and maintenance of computer hardware should constantly be made to keep computers in a good condition for testing purposes. Finally, if practically resources permit, perhaps computer laboratories should be allocated for such purposes. It is unreasonable to argue against using Moodle for assessment because it requires a more efficient technical infrastructure. Therefore, technical issues and proper facilities and resources should be on the agenda (Al-Ani, 2008; Fulcher, 2003; Hinkelman & Grose, 2004) in order to implement Moodle assessments and other technology-enhanced assessments at the study context. Future research should examine the effect of providing standardized hardware and software tools and following consistent exam procedures in high-stakes large-scale testing.

7.4.4. Test time and length.

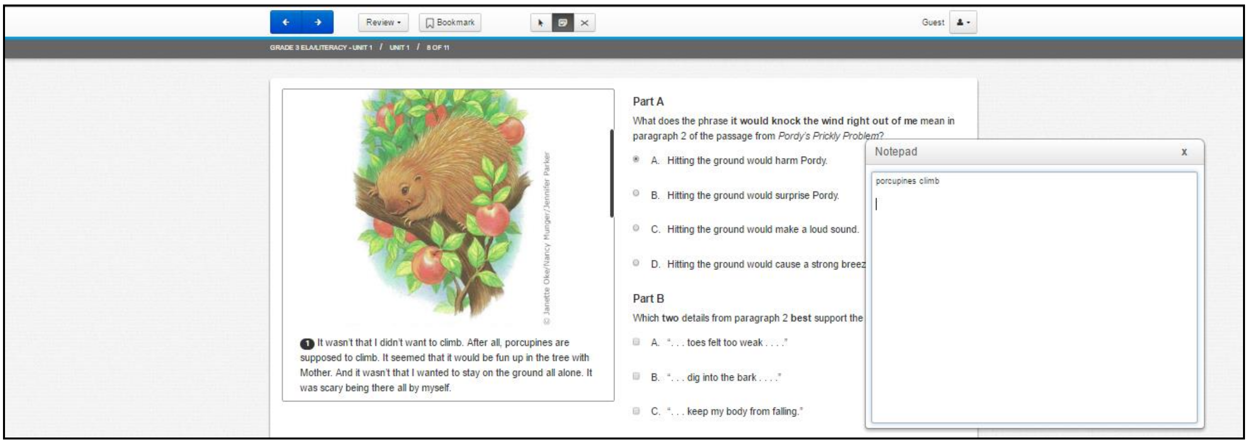
As reported in Section 5.11 (pp. 81-83) and Table 5.10 (p. 82), test time and length were shown to interfere with test performance. The sufficiency of test timing for all test sections and the presence of a count-down timer were found as factors interfering with test performance. Having a long test with more items might provide testers with more information on examinee ability, but the negative effects of increased test time and length might outweigh the benefits. Therefore, as outlined in Section 6.3.5 (pp. 103-107), further research should examine the effects of the amount of testing time required, and the use of features like count-down timers on screen to aid examinees in managing their time.

7.4.5. Using new features.

The features of the testing interface should reflect the tested construct and if equivalence to paper-based testing mode is a main concern for practitioners, such features should be carefully designed and backed up with a rationale for their inclusion into the technology-enhanced interface. One example of this is the creation of the split screen mode for reading tests in the Moodle-hosted study (Chapter 3, Section 3.5.1, pp. 38-42). This layout and scrolling feature had a trend of an impact on test performance but was not found significant (Section 5.8, pp. 76-77 and Table 5.7, p. 77). This feature came into existence because the study participants (language teachers in the pilot study) believed that it was easier to navigate through the digital reading texts and relevant test items when they can be located on the same page side by side with minimal scrolling required. As stated in Al Nadabi (2015), this can be explained by the split attention principle which is part of Sweller's (1994) cognitive load theory (Ayres & Sweller, 2005). The assumption is that examinees' concentration during the test should be increased by presenting the reading test materials in a split screen interface. This is similar to the paper-based testing mode in which the reading text is put on one page and the items are put on the opposite page. Layout and scrolling features, note-taking and text highlighting features were shown as non-significant technology-related variables but were highlighted as issues of concern that need to be examined in light of their interaction with other factors in the testing environment (Section 6.3.5, pp. 103-107). Practitioners need to articulate their rationale for such features with practical and theory-based evidence such as a reduction of examinee split attention and cognitive load when processing information.

In another example, the number of times examinees can hear the listening materials in a paper-based testing mode is controlled. This was also controlled in the Moodle-hosted test listening test using Matburry's embedded MP3 player. The rationale behind including this feature is to ensure fair testing practices since, as is the case with paper-based exams, variations in the number of times examinees access the listening materials might put some students at an advantage (Al Nadabi, 2015). Although technical features such as these can serve unique and useful purposes, they still need to reflect the construct tested in the paper-based format and avoid becoming construct-irrelevant sources of measurement error. Following this principle, it is also important to set out specifications for layout and scrolling features so that distractions caused by such features are minimised.

Moreover, tests should remain subject to improvements and additions because "testing is always the Current Best Shot" (Brown, 2008, p. 302). From the outcomes of this study, new features that can be added to the Moodle-hosted testing interface and other similar interfaces include electronic note-taking and text highlighting features. Figure 7.1 shows an example of a Notepad that test takers can use to make notes during a Partnership for Assessment of Readiness for College and Careers (PARCC) reading exam. Figure 7.2 provides a screenshot of an Answer Eliminator Tool from PARCC. Examinees can use this tool to highlight and select or cross out options when answering test items. As recommended by Care, Luo, Awwal, and Yasotha (2015), the potential benefits of using these online reading tools features can be studied in future applications of web-based testing interfaces.


Review
Bookmark
Guest
GRADE 3 ELA/LITERACY - UNIT 1 / UNIT 1 / 8 OF 11
© Janette Oke/Nancy Munger/Jennifer Parker
Notepad
x

1 It wasn't that I didn't want to climb. After all, porcupines are supposed to climb. It seemed that it would be fun up in the tree with Mother. And it wasn't that I wanted to stay on the ground all alone. It was scary being there all by myself.

Part A

What does the phrase **it would knock the wind right out of me** mean in paragraph 2 of the passage from *Pordy's Prickly Problem*?

● A. Hitting the ground would harm Pordy.

○ B. Hitting the ground would surprise Pordy.

○ C. Hitting the ground would make a loud sound.

○ D. Hitting the ground would cause a strong breez

Part B

Which **two** details from paragraph 2 **best** support the

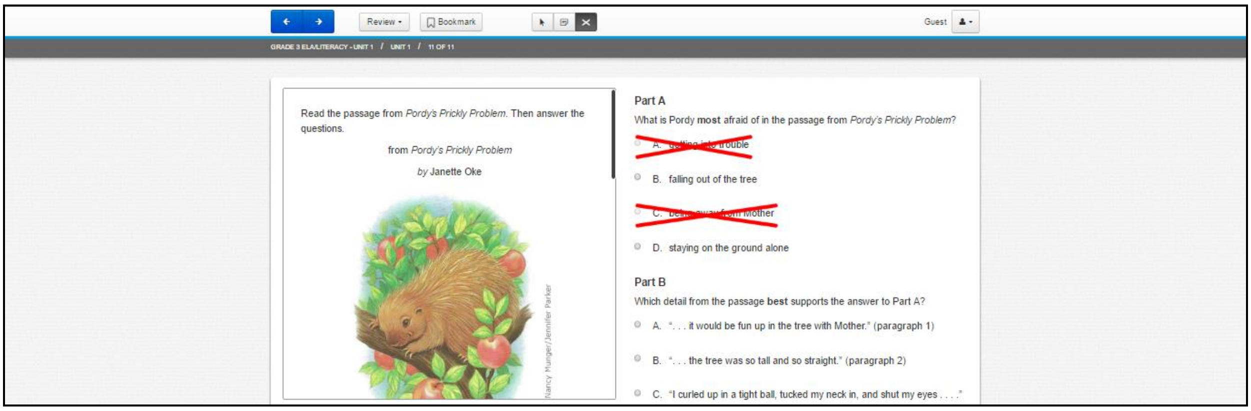
□ A. ". . . toes felt too weak . . . ."

□ B. ". . . dig into the bark . . . ."

□ C. ". . . keep my body from falling."

porcupines climb

Figure 7.1. Screenshot from PARCC reading exam using Notepad to make notes, taken from online practice tests available at https://parcc.pearson.com/practice-tests/english/


Review
Bookmark
Guest
GRADE 3 ELA/LITERACY - UNIT 1 / UNIT 1 / 11 OF 11
Nancy Munger/Jennifer Parker

Read the passage from *Pordy's Prickly Problem*. Then answer the questions.

from *Pordy's Prickly Problem*

by Janette Oke

Part A

What is Pordy **most** afraid of in the passage from *Pordy's Prickly Problem*?

○ ~~A. getting into trouble~~

○ B. falling out of the tree

○ ~~C. being away from Mother~~

○ D. staying on the ground alone

Part B

Which detail from the passage **best** supports the answer to Part A?

○ A. ". . . it would be fun up in the tree with Mother." (paragraph 1)

○ B. ". . . the tree was so tall and so straight." (paragraph 2)

○ C. "I curled up in a tight ball, tucked my neck in, and shut my eyes . . . ."

Figure 7.2. Screenshot from PARCC reading exam using Answer Eliminator Tool, taken from online practice tests available at https://parcc.pearson.com/practice-tests/english/

7.4.6. Eye fatigue.

Constant and unprotected exposure to visual activity in lengthy testing sessions might trigger eye fatigue. Hence, test length effect should be revisited and researched extensively to determine the amount of time that can be considered reasonably tolerable in a certain testing context. Eye fatigue was one of the variables that tended to affect test performance but was not found statistically significant (Section 5.2, pp. 66-68 and Table 5.1, p. 67). To examine the effect of computerised testing on eye fatigue, future research on visual ergonomics in the field of language testing in particular is needed. This issue is at large connected to the amount of time test takers spend on the test, which is mainly dictated by the test length or number of test items and sections. Therefore, future research should investigate the effects of eyestrain in relation to test length in exam conditions. Objective measures such as eye tracking technology as well as subjective measures such as surveys and think-aloud techniques can be employed to arrive at more conclusive results.

7.4.7. Attitude towards testing format.

In terms of research methodology, test takers constituted an essential source of information on how reliable and valid a testing environment is. Test takers' voices were heard through this study. For instance, this study depicted their 'attitude and resistance to change' (Section 5.5, pp. 71-73) and their views on Moodle's 'appropriateness for testing purposes' (Section 5.10, pp. 79-81). It was reported in this study that three quarters (74.1%) of test takers preferred pen on paper over Moodle tests and thought they would perform best on paper (Section 5.9, pp. 77-79). The attitude towards testing format was found to interfere with test performance (Section 6.3.3, pp. 98-100). Hence, researchers should consider putting test takers' perspectives at the heart of the research, examining the effect of the testing mode on their test performance. As recommended in Section 6.3.3 (pp. 98-100), to further identify the testing mode effect, research should focus on students' negative attitude and resistance to educational innovations and changes including new types of assessment delivery. Such research can explore attitudinal factors such as test-taking motivation and success expectation adopting theoretical frameworks as in the Expectancy-value theory (Fan and Ji, 2014; Jacobs & Eccles, 2000).

These were some of the study implications and recommendations for practice and future research. Table 7.1 summarizes the recommendations or advice made for practitioners to address the technology-related issues that the study found to affect test performance.

Table 7.1. *Practitioner Advice on Addressing Technology-Related Issues*

Issue affecting test performance	Practitioner advice
Familiarity and experience	As an equity and bias-prevention measure, evaluate and increase student familiarity and experience with technology, the testing interface and computers used for assessment purposes, by conducting familiarity studies and by giving students access to sample test materials and tutorials.
Headphones quality	To achieve procedural fairness and as a measure to prevent bias introduced by variabilities in test administration conditions, provide standard devices and equipment (including headphones) to all examinees. Test that such equipment meet good quality hardware specifications and are in good working condition prior to the testing event.
Attitude towards testing format (which testing format students would perform best on)	To promote more positive attitudes and acceptance of the computerised testing mode as well as positive test impacts, explore test takers' attitudes, provide intervention measures, and allow test takers access to test information.
Typing responses for gap-filling items	To raise test taker familiarity with test typing requirements and to develop their keyboarding skills, provide them with sample materials and assessment tutorials that help eliminate the effect of weak keyboarding ability levels. From a fairness perspective, consider offering examinees the options between handwriting and typing their test responses.
Test time and length	To determine the time and test length provision that can be at a fair level of sufficiency and to avoid advantaging or disadvantaging students in timed exam conditions, examine and allocate sufficient test time allotments and test time management features (such as count-down timer) through research and practice.

To arrive at more reliable and valid score-based decisions, practitioners and researchers may wish to consider such implications and recommendations as guidelines for creating, developing, implementing, and researching large-scale high-stakes tests delivered on Moodle, other course management systems, or any other test delivery technologies. To achieve more reliable and valid testing outcomes, these implications and recommendations should be considered keeping in mind the study limitations, as outlined next.

7.5. Limitations

The limitations of the study need to be acknowledged to inform practitioners of all of the study aspects that may have contributed to the study results. This study had limitations for a number of reasons.

7.5.1. Study design.

First of all, the study design involved the collection of score data from a single administration of the Moodle-hosted test per test taker rather over two or more administrations which would have provided longitudinal or comparative data of test taker experiences. Logistically speaking, it was not feasible to have test takers undertake the same test in two different modes, paper-based and Moodle-hosted mode. The statistical score data analysis aimed to examine RQ1 had to fit this design by applying Rasch analysis for internal test reliability of a single administration score data. Other techniques for establishing reliability such as test-retest were not possible given the way the study was designed. Task difficulty and reliability coefficients are sample dependent in the sense that they get affected by the interaction of examinees and the test task (Weir, O'Sullivan, Jin, & Bax, 2007). It must be highlighted that establishing paper-equivalence was not the focus of this study. Instead the study was looking at technology-related factors that may interfere with test validity – given technology enables new question types to be possible. Technology should go beyond paper capabilities. Therefore, the study sought to provide an in-depth study of reliability concerns experienced by examinees when interacting with the task of taking the test in the Moodle-hosted context. It is essential in future research designs to attempt to unfold the effects of specific features and potential issues in taking the test in a course management system environment like Moodle, without being distracted by the comparison to paper-based test versions. Nevertheless, this study might not provide a straightforward answer to the question that may be asked by stakeholders on whether the new testing mode is any better than the traditional paper-based format. Perhaps another follow-up study that takes a comparative design approach can be conducted for the sake of answering such a question.

7.5.2. Study data.

The study design also allowed the collection of rich data, but due to time limitations, not all the data were incorporated in the analyses. The data that were utilized in the study included the test score data and examinees' questionnaires only. The invigilators' questionnaires, interviews with examinees and invigilators, and the researcher's reflective journals had to be excluded from the analyses. This being said, leaving out a large proportion of value-laden data might have led to missing important research outputs that could have enriched the thick description of the case study.

The data that were excluded were qualitative in type and they represented the participants' views through their verbal prose as well as the researcher's account of the test experience as an active study participant. Although the pieces of evidence provided by the main data sources used in the study answered the research questions, perhaps the rest of the unanalyzed data can provide a richer and deeper picture about the case study of administering the Moodle-hosted test. This study can later be extended to include the remainder of the data in order to achieve more triangulation to support the validity argument claims with evidence utilizing the voices of the participants who went through this test experience.

7.5.3. Data analyses.

Although evidence in the study was established through quantitative and qualitative data analysis techniques in a mixed-method approach, the study was limited by dominance of quantitative techniques due to the exclusion of textual data. The study findings, therefore, reflect a lack of qualitative types of evidence. Having both quantitative and qualitative types of evidence should speak well to the mixed-method approach, but it was not feasible to fully explore in this study. In addition, due to the scope of the study, a single item analysis method was chosen. The majority of the analysis was therefore based on single items from the test takers' questionnaires. However, it is acknowledged that single items may not be reliable. Further research would be advised in taking into consideration combinations of items through techniques such as factor analysis.

7.5.4. Language use in instruments.

Another limitation relevant to the use of examinees' open comments from the questionnaires is that some of the data were translated by the researcher from the participants' first language, Arabic, to English. The questionnaires were given to examinees in both languages and they were allowed to respond in Arabic in order to let them express themselves freely without being limited by their English language ability. However, it should be acknowledged here that it is possible that translating these comments might not have fully depicted the meanings intended by the writers of these comments as meanings can be lost in translation. On the other hand, students who attempted to express their views in English might have lost the meanings that they intended to deliver due to their limited language proficiency.

7.5.5. Examined issues.

The Moodle platform was updated in the study context from the old Moodle version 1.9 to Moodle 2.7 version after this study took place. Although the study results presented in this thesis represent the experience of taking the test in the older Moodle version, such research outcomes might be held true for whatever Moodle version as no issues specific to that Moodle version were reported.

Nevertheless, it is still worth exploring the same types of issues and any others in whatever technology is used to deliver an online test. Similarly, data gathering tools including questionnaires and interviews can also be adapted and developed further to be used in future research in the area. The questionnaires and interviews that were used in the main study had already been validated through the pilot study, so their questions reflect some of the issues that had been experienced. For instance, a question about eye fatigue was added to the questionnaire only after it emerged as a recurring theme in the pilot study. Although this was done to assess what issues have been experienced in the exam context, the specific issues examined through the questionnaire items might have limited participants' responses and their reflection of their experiences in the exam context. The door was open for comments in some of the questionnaire items for examinees to reflect on their experience more freely. As said earlier, the interviews were excluded from the analyses for logistical reasons. Therefore, this study was bounded in that the participants' free expression and reflection could not be incorporated.

7.5.6. Test taker characteristics.

When it comes to the study sample, characteristics of the test takers such as the levels of their familiarity and experience with technology and their language ability levels could not be controlled given the nature of voluntary participation upon invitation. The study sample represented the three language levels of the test population intended to sit the original paper-based version of the test. However, the test takers participating in the study might not be as representative of the test population taking the test in the first semester of the academic year. Limited by time constraints in data collection, the study took place in the second semester. By that time, some students might have already gone through the English language program and other Foundation Program courses in mathematics and information technology and might have gained some skills. Other students were probably newly admitted to university in the second round of admission. Therefore, there might have been variations in the test taker characteristics that were not accounted for in the study as information on prior study at the university was not collected. Added to this, the study findings were based on the comparison of examinees' views with their test results, but their test performance might not have reflected their true language ability levels. This is because they might not have taken the test seriously since it was not conducted as an official high-stakes exam and their results did not contribute to their course grade.

7.5.7. Documenting details.

The main features of the Moodle-hosted testing interface were described in this thesis (Chapter 3, Section 3.5.1, pp. 38-42) and disseminated in Al Nadabi (2015), but many other details related to the computer and other hardware and software specifications and computer laboratories layout were not well-documented. Added to this, logistical arrangements with test takers and their invigilating teachers mandated the administration of the Moodle-hosted test in separate testing sessions for different classes using different computer laboratories. Although every examinee took the test only once, variations in exam conditions that can have implications for reliability and construct validity investigations could have occurred, but were not well-documented. This limits the thick description of the context and the study procedures, which is essential for transferability or generalizability and applicability to other contexts (Brown, 2008).

These were the limitations encountered in the study. Acknowledging these limitations is important for generalizability considerations of the findings as well as for future research planning. Practitioners and other researchers need to take these study limitations into account when embarking on similar test development projects and research studies and when adapting the study recommendations.

Based on the study implications and limitations, Table 7.2 provides an agenda for future research by summarizing the technology-related issues that should receive researchers' attention. It should be noted here that researchers may examine how technology-related factors contribute individually to test performance. However, as noted in Section 6.3.5 (pp. 103-107), such factors also need to be investigated in light of how they can interact with each other to make for a valid testing environment. For instance, in examining gap-filling items, researchers may also need to examine the effect of keyboarding skills on the test performance on such items. The effect of providing standardized technical resources and consistent test procedures can also be investigated to address potential bias issues resulting from variabilities in the testing context. To stay informed about fair testing practices, the examination of the effect of the test time and the use of test time management features like count-down timers should also be a future research agenda. As future applications of web-based testing interfaces introduce new features (such as electronic note-taking and text highlighting features), their potential effect on test performance should be examined. Future research should also look into the effect of eye fatigue in a computerised testing environment and how this may interact with the test time variable. Due to the impact of test consequences on students' lives, attitudinal factors that may interfere with test performance should also be at the forefront of future research. A bias analysis of such issues affecting test performance could be undertaken following other approaches as in Differential Item Functioning (DIF) analysis (Zumbo, 2007) that can be used to examine test fairness by identifying test score differences among particular test taker groups (Kunnan, 2010). More advanced statistical techniques can be used such as confirmatory factor analysis models and the logistic regression procedure.

Table 7.2. Agenda for Future Research

No.	Research Issue
1.	Effect of improved keyboarding skills on test performance
2.	Effect of including gap-filling items in computerised exams and how keyboarding skills can contribute to test performance on such items requiring typing responses
3.	Effect of providing standardized hardware and software tools (such as headphones) and following consistent exam procedures in high-stakes large-scale testing
4.	Effect of the amount of test time and the use of features like count-down timers in computerised exams
5.	Potential benefits of using new features in future applications of web-based testing interfaces (such as electronic note-taking and text highlighting features)
6.	Effect of computerised testing on eye fatigue employing visual ergonomics research in the field of language testing that uses objective measures (such as eye tracking technology) and subjective measures (such as surveys and think-aloud techniques)
7.	Effect of attitudinal factors on test performance voicing test takers' concerns

7.6. Concluding Remarks

Through the case study of administering the Moodle-hosted test, the articulated validity argument indicated that the use of the computerised test could not be supported in this instance because of threats to reliability and construct validity. In a nutshell, technology-related construct-irrelevant factors contributed to the testing mode effect that interfered with test results. This argument is also backed by the potential risks of creating bias or unfairly advantaging or disadvantaging examinees by test administration variations and technology-related issues interfering with test performance in the testing context. As test takers come with their own technology-related skills and attitudes, their characteristics (such as technology familiarity and experience) cannot be neglected. Hence, as discussed in Chapter 6 (Sections 6.2.2, pp. 91-93 and 6.4, pp. 107-111), this study highlights the importance of standardized testing practices that can enhance procedural fairness and limit the impact of contextual testing variables on test performance. Providing technical training and familiarizing examinees with the test format using tutorials and sample materials is also seen as one way to reduce bias that can be created by lack of familiarity with the testing format and item types (such as constructed-response items) that are introduced in the computerised testing environment.

The negative study outcomes here should not be taken as a discouragement from utilizing technology in testing and assessment. Rather, with the extensive use of technologies (such as course management systems) in education including language testing and assessment purposes, this study has highlighted the sorts of issues that can be encountered and recommends how such issues can be resolved. Only by addressing technology-related issues can the testing mode effect be reduced or eliminated. As with most case study based approaches, it should be recognized that the specific study findings are unlikely to be generalizable to other contexts in which tests are delivered via computer. This is due to the interaction of the unique features of the study sample and the specific features of the Moodle-hosted testing interface used in the study. However, the study delivers value for testing practitioners and researchers working on technology-based test development projects and research studies. Researchers can draw upon the methodological approach and validation framework to conduct similar research studies and construct evidence-supported validity arguments. Practitioners will find the validation framework and findings of this study useful in the process of designing and developing their own tests and testing interfaces for their unique test populations. Considering the validity argument made in this study, policy-makers and practitioners at the study context can find the study implications and recommendations useful in their ambitious future planning for large-scale high-stakes computerised testing practices.

References

Akiyama, T. (2001). The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context. *Melbourne Papers in Language Testing, 10*(1), 1-22.

Al-Amri, S. (2007). Computer-based vs. paper-based testing: Does the test administration mode matter? *Proceedings of the BAAL Conference 2007,*101-110. Retrieved from http://www.baal.org.uk/proc07/33_saad_al_amri.pdf

Al-Ani, W. (2008). English as foreign language student teachers' perception of the use of Moodle in foundations of education course. *Malaysian Journal of Learning and Instruction, 5*, 63-78. Retrieved from http://core.kmi.open.ac.uk/download/pdf/12116774.pdf

Al-Ani, W. T. (2013). Blended learning approach using Moodle and student's achievement at Sultan Qaboos University in Oman. *Journal of Education and Learning, 2*(3), pp. 96-110. doi:10.5539/jel.v2n3p96.

Al-Busaidi, S., & Tuzlukova, V. (2013a). Some reflections on Moodle-based learning in the English Foundation Programme of Sultan Qaboos University. *Asian Journal of Social Sciences & Humanities, 2*(3), 166-173. Retrieved from http://www.ajssh.leena-luna.co.jp/AJSSHPDFs/Vol.2(3)/AJSSH2013(2.3-19).pdf

Al-Busaidi, S. & Tuzlukova, V. (2013b). Learner autonomy support in EFL classroom: Students' perspective. In Al-Busaidi, S. & Tuzlukova, V. (2013). *General Foundation Programmes in Higher Education in the Sultanate of Oman: Experiences, Challenges and Considerations for the Future* (pp. 138-157). Muscat, Oman: Mazoon Press & Publishing.

Al-Hajri, A. O. (2011). *Computer assisted assessment in Oman: Factors affecting student performance* (Doctoral dissertation, University of Plymouth). Retrieved from http://pearl.plymouth.ac.uk/handle/10026.1/318

Al Naddabi, Z. (2007). A Moodle course: Design and implementation in English for academic purposes instruction. In T. Bastiaens & S. Carliner (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007,* 1371-1376. Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/p/26540/

Al Naddabi, Z. (2012). *The English language placement test at Sultan Qaboos University: Level representation and internal validity* (Unpublished master's thesis). Lancaster University, UK.

Al Nadabi, Z. (2015). Features of an online English language testing interface. In T. Reiners, B. R. von Konsky, D. Gibson, V. Chang, L. Irving, & K. Clarke (Eds.), *Globally connected,* *digitally enabled. Proceedings ASCILITE 2015 in Perth* (pp. CP:17-CP:21). http://www.2015conference.ascilite.org/wp-content/uploads/2015/11/ascilite-2015-proceedings.pdf

American Psychological Association. Committee on Professional Standards, American Psychological Association. Board of Scientific Affairs. Committee on Psychological Tests, & Assessment. (1989). *Guidelines for computer-based tests and interpretations*. Washington, DC: The Association.

Arnold, J. (2000). Seeing through listening comprehension exam anxiety. *TESOL Quarterly, 34*(4), 777-786.

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the International English Language Testing System (IELTS) listening module. *International Journal of Listening, 26*(1), 40-60.

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Newcastle Upon Tyne, UK: Cambridge Scholars Publishing.

Aryadoust, V. S. & Goh, C. (2009). A Rasch analysis of an international English language testing system listening sample test. Paper presented at *the 3rd Redesigning Pedagogy International Conference*, June 2009, Singapore.

Australian Bureau of Statistics (ABS) (2016, 18 February). *Household use of information technology, Australia, 2014-15*. Retrieved from http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/8146.02014-15?OpenDocument

Australian Curriculum, Assessment and Reporting Authority. (2016). *NAPLAN online technical requirements*. Retrieved from http://www.nap.edu.au/_resources/NAPLAN_online_technical_requirements_updated_October_2015.pdf

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R.E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 134-146). New York, NY: Cambridge University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34. doi: 10.1207/s15434311laq0201_1

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions,* 22(1). 1145-1146. Retrieved from https://www.rasch.org/rmt/rmt221.pdf.

Bazeley, P. (2010). NVivo. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 945-949). doi:http://dx.doi.org.ezproxy.library.uq.edu.au/10.4135/9781412961288.n28

Bazeley, P., & Jackson, K. (2013). *Qualitative data analysis with NVivo*. Los Angeles: SAGE.

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241-259.

Benedetto, S., Drai-Zerbib, V., Pedrotti, M., Tissier, G., & Baccino, T. (2013). E-readers and visual fatigue. *PLoS ONE, 8*(12): e83676. doi:10.1371/ journal.pone.0083676

Bond, T.G. (2003). Validity and assessment: A Rasch measurement Perspective. *Methodologia de las Ciencias del Comportamiento, 5*(2), 179-194.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34. doi: 10.1111/j.1745-3992.1992.tb00260.x

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics, 18,* 171-91.

Brown, J. D. (1999). Statistics corner: Questions and answers about language testing statistics: Standard error vs. Standard error of measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter, 3*(1), 20-25. Retrieved March 12, 2016 from http://jalt.org/test/bro_4.htm

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, UK: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw Hill.

Brown, J. D. (2008). Testing-context analysis: Assessment is just another part of language curriculum development. *Language Assessment Quarterly, 5*(4), 275-312. doi: 10.1080/15434300802457455

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, UK: Cambridge University Press.

Care, E., Luo, R., Awwal, N., & Yasotha, V. (2015). *NAPLAN Online readability and layout study: Literature review*. Assessment Research Centre, University of Melbourne. Retrieved from http://www.nap.edu.au/docs/default-source/default-document-library/naplan-online-readability-and-layout-study.pdf?sfvrsn=2

Castle, R. (2016). What is measurement error and what is its relationship to reliability? *The Professional Testing Blog*. Retrieved August 5, 2017 from http://www.proftesting.com/blog/2016/10/13/measurement-error-relationship-reliability/

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research (Cambridge applied linguistics series).* Cambridge, UK: Cambridge University Press.

Chapelle, C. A. (2008). Utilizing technology in language assessment. In E. Shohamy (Ed.), *Encyclopedia of Language Education (Vol. 7). Language Testing and Assessment* (2nd ed., pp. 123-134). Heidelberg, Germany: Springer. doi: 10.1007/978-0-387-30424-3

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Germany: Ernst Klett Sprachen. doi: 10.1017/CBO9780511733116

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409-439. doi: 10.1191/0265532203lt266oa

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. doi: 10.1191/0265532203lt258oa

Clark, J. L. D. (Ed.) (1978). *Direct testing of speaking proficiency: Theory and application*. Princeton, NJ: Educational Testing Service.

Cohen, A. (2012). Test-taking strategies and task design. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 262-278). Milton Park, Abingdon, Oxon; New York: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coleman, G., & Heap, S. (1998). The misinterpretation of directions for the questions in the academic reading and listening sub-tests of the IELTS test. *IEL TS Research Reports*, *1*, 38-71.

Coniam, D. (1999). Subjects' reactions to computer-based tests. *Journal of Educational Technology Systems, 23*(3), 195-206. doi: http://dx.doi.org/10.2190/HL2N-V10Q-PAGA-6MRH

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL, 18*(2), 193-211. doi: 10.1017/S0958344006000425

Coy, J. (2013). *Instant Moodle quiz module how-to*. Birmingham, UK: Packt Publishing Ltd. Retrieved from http://dl.e-book-free.com/2013/07/instant_moodle_quiz_module_how-to.pdf

Crone, C., Wright, D., & Baron, P. (1993). *Performance of examinees for whom English is their second language on the spring 1992 SAT II: Writing Test*. Unpublished manuscript prepared for ETS. Princeton, NJ: Educational Testing Service.

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: SAGE Publications.

Davidson, F. (2000). The language tester's statistical toolbox. *System*, *28*(4), 605-617. doi:10.1016/S0346-251X(00)00041-5

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, UK: UCLES and Cambridge University Press.

Davis, L. L. (2015). *Device effects in online assessment: A literature review for acara*. Retrieved from http://www.nap.edu.au/docs/default-source/default-document-library/naplan-online-device-effect-study.pdf?sfvrsn=2

Davis, L. L., Janiszewska, I., Schwartz, R., & Holland, L. (2016). *NAPLAN device effects study*. Melbourne, Australia: Pearson. Retrieved from http://www.nap.edu.au/docs/default-source/default-document-library/naplan-online-device-effect-study.pdf?sfvrsn=2

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35*(10), 1297-1326. doi: 10.1080/00140139208967394

Douglas, D., & Hegelheimer, V. (2007). Assessing language through computer technology. *Annual Review of Applied Linguistics, 27,* 115-132. doi: 10.1017/S0267190508070062

Dyson, M. C., & Kipping, G. J. (1998). The effects of line length and method of movement on patterns of reading from screen. *Visible Language*, *32*(2), 150-81.

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). Development of a scale for assessing the level of computer familiarity of TOEFL examinees. *ETS Research Report Series*, *60*, i-32. Retrieved from https://www.ets.org/Media/Research/pdf/RR-98-07.pdf

Elliot, B. (2007). *Assessment 2.0: Assessment in the age of Web 2.0*. Scottish Qualifications Authority. Retrieved from http://wiki.cetis.ac.uk/images/d/de/Assessment_2_v2.pdf

Fan, J., & Ji, P. (2014). Test candidates' attitudes and their test performance: The case of the Fudan English Test. *University of Sydney Papers in TESOL, 9*, 1-35.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London; Thousand Oaks, CA: Sage Publications.

Geranpayeh, A., & Taylor, L. (2013). *Examining listening: Research and practice in assessing second language listening. Studies in language testing, 35.*

Foundation programme English language curriculum document. (2012-2013). Retrieved from https://www.squ.edu.om/Portals/31/LC%20Documents/FPEL%20Curriculum%20Document%202012-2013%20FINAL.pdf

Fulcher, G. (1999). Computerizing an English language placement test. *ELT journal, 53*(4), 289-299.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing, 20*(4), 384-408. doi: 10.1191/0265532203lt265oa

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Abingdon, [England]; New York, NY: Routledge.

Ghanbari, N., Karampourchangi, A., & Shamsaddini, M.Z. (2015). An exploration of the effect of time pressure and peer feedback on the Iranian EFL students' writing performance. *Theory and Practice in Language Studies, 5*(11), 2251-2261.

Green, R. (2013). *Statistical analyses for language testers*. New York, NY: Palgrave Macmillan.

Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*. Dordrecht, Germany: Springer. doi: 10.1007/978-94-007-2324-5

Gruba, P. & Hinkleman, D. (2012). *Blending technologies in second language classrooms*. UK: Palgrave Macmillan.

Hamouda, A. (2013). An investigation of listening comprehension problems encountered by Saudi students in the el listening classroom. *International Journal of Academic Research in Progressive Education and Development, 2*(2), 113-155.

Heigham, J., & Croker, R. A. (2009). *Qualitative research in applied linguistics: A practical introduction*. Houndmills, Basingstoke, Hampshire [England]; New York; NY: Palgrave Macmillan.

Hale, G. (1992). Effects of amount of time allowed on the Test of Written English. *ETS Research Report Series, 1992*(1), I-35.

Hillier, M. (2015). *E-exams with student owned devices: Student voices*. Presented at the International Mobile Learning Festival Conference: Mobile Learning, MOOCs and 21st Century learning (pp. 582-608), Hong Kong. 22-23 May. Retrieved from *http://www.transformingexams.com/files/Hillier_IMLF2015_full_paper_formatting_fixed.pdf*

Hinkelman, D., & Grose, T. (2004). Placement testing and audio quiz-making with open source software. *Proceedings of CLaSIC. CLS International Conference* (pp. 974-981). Retrieved from http://englishforum.sgu.ac.jp/downloads/Old-EnglishPlacementTests/SingaporeCALL-2004.12.3.pdf

Information Technology Authority. (December, 2012). *Information and communication technology (ICT) survey results*. Retrieved from http://www.oman.om/wps/wcm/connect/6114f831-434a-418a-a970-6fd16386b358/ICT+Surveys+Results+2012+English+Final.pdf?MOD=AJPERES

ITU (2016a). *ICT development index 2016: Oman*. Retrieved from http://www.itu.int/net4/ITU-D/idi/2016/#idi2016countrycard-tab&OMN

ITU (2016b). *ICT development index 2016: Australia*. Retrieved from http://www.itu.int/net4/ITU-D/idi/2016/#idi2016countrycard-tab&AUS

Jacobs, J. E., & Eccles, J. S. (2000). Parents, task values, and real-life achievement-related choices. In C. Sansone & J. M. Harackiewicz (Eds), *Intrinsic Motivation* (pp. 405-439). San Diego, CA: Academic Press.

Jordan, S. (2008). Online interactive assessment with short free-text questions and tailored feedback. *New Directions*, (4), 17-20. doi: 10.11120/ndir.2008.00040017

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535. doi: 10.1037/0033-2909.112.3.527

Kane, M. T. (2010). Validity and fairness. *Language Testing*, *27*(2), 177-182. doi:10.1177/0265532209349467

Kane, M. T. (2011). Book review: Language assessment in practice: Developing language assessments and justifying their use in the real world. *Language Testing*, *28*(4), 581-587. doi: 10.1177/0265532211400870

Kane, M. T. (2012). Articulating a validity argument. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language Testing* (pp. 34-47). Milton Park, Abingdon, Oxon, UK; New York, NY: Routledge.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17. doi: 10.1111/j.1745-3992.1999.tb00010.x

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees*. TOEFL Research Report 59. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-98-06.pdf

Knoch, U., & Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency. *System: An International Journal of Educational Technology and Applied Linguistics, 38*(1), 63-74.

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275-304). New York, NY: Routledge.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.)*, Second language writing: Research insights for the classroom* (pp. 140-154). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing*, *9*(1), 30-49. doi: 10.1177/026553229200900104

Kunnan, A. J. (1998). *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium*. Long Beach, Mahwah, NJ: Lawrence Erlbaum.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona conference* (pp. 27-48). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2010). Statistical analyses for test fairness, *Revue française de linguistique appliquée, 1*(XV), 39-48. Retrieved from https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2010-1-page-39.htm

Lado, R., 1961. *Language testing: The construction and use of foreign language tests*. London, UK: Longman.

Linacre, J. (2012). *A user's guide to Winsteps Ministep Rasch-Model computer programs Program Manual 3.75. 0*. Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf

Linacre, J. (2014). *Standard error of measurement of a test*. Rasch Measurement Forum. Retrieved from http://raschforum.boards.net/thread/10/standard-error-measurement-test

Livingston, S. A. (1987). The effects of time limits on the quality of student-written essays. *Paper presented at the Annual Meeting of the American Educational Research Association*, April 1987, Washington.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351-372. doi: 10.1177/026553220101800403

Mangen, A., Walgermo, B. R., & Bronnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research, 58*, 61-68. doi: 10.1016/j.ijer.2012.12.002

Matbury (2010a). *How to use: How to deploy the player in web pages*. Retrieved from https://code.google.com/p/moodle-mp3-player-for-tests/wiki/HowToUse

Matbury (2010b). Moodle-mp3-player-for-tests*: An MP3 player for embedding in tests.* Retrieved from https://code.google.com/p/moodle-mp3-player-for-tests/downloads/detail?name=mp3_player_for_tests_2010_04_08_11.zip

Maycock, L. & Green, T. (2005). The effects on performance of computer familiarity and attitudes towards CB IELTS. *Research Notes, 20,* 3-8. Retrieved from http://www.cambridgeenglish.org/images/23138-research-notes-20.pdf

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-76.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8*(2), 139-159.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK; New York, NY: Longman.

McNamara, T. F. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31-51.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension* (Language learning. Monograph series). Malden, Mass.: Blackwell Pub.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series, 1993*(2).

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477-496. doi:10.1191/0265532202lt241oa

Moodle statistics (2017, August 17). Retrieved from https://moodle.net/stats/

Myrick, J. (2010). *Moodle 1.9 testing and assessment*: Birmingham, UK: Packt Publishing Ltd. Retrieved from http://my.safaribooksonline.com/book/web-applications-and-services/9781849512343

Najwani, Z. A. S. (2013). *Foundation programme students' attitudes to Moodle Quizzes.* (Unpublished master's thesis). University of Leeds, UK.

National Council on Measurement in Education, American Psychological Association, and American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Noijons, J. (1994). Testing computer assisted language testing: Towards a checklist for CALT. *CALICO Journal*, *12*(1), 37-58. Retrieved from https://calico.org/html/article_580.pdf

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing*. Unpublished PhD dissertation: University of Reading, UK.

O'Sullivan, B. (2012). The assessment development process. In C. Coombe, B. O'Sullivan, & S. Stoynoff. (Eds.), *The Cambridge guide to second language assessment* (pp. 47 - 58). Cambridge, UK: Cambridge University Press.

Pallant, J. (2010). *SPSS survival manual: A step by step guide to data analysis using SPSS* (4th ed). Maidenhead, UK: Open University Press/McGraw-Hill.

Pallant, J. (2013). *A step by step guide to data analysis using IBM SPSS: Survival manual*. Maidenhead, UK: McGraw Hill.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. Statistics for Social and Behavioral Sciences Series. doi: 10.1007/978-1-4613-0083-0

Pinto, R. M. (2010). Mixed methods design. In N. Salkind (Ed.), *Encyclopedia of research design qualitative research* (pp. 813-819). doi: http://dx.doi.org/10.4135/9781412961288.n245

Powers, D., & Fowles, M. (1996). Effects of applying different time limits to a proposed GRE Writing Test. *Journal of Educational Measurement, 33*(4), 433-452.

Putney, L. (2010). Case study. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 116-120). doi: http://dx.doi.org.ezproxy.library.uq.edu.au/10.4135/9781412961288.n39

Questions (2013, October 16). Retrieved from http://docs.moodle.org/25/en/Questions

Rassekh, S. (2004). *Education as a motor for development: Recent education reforms in Oman with particular reference to the status of women and girls*. UNESCO: IBE. Retrieved from http://www.ibe.unesco.org/fileadmin/user_upload/archive/Publications/innodata/inno15.pdf

Roever, C. (2001). *A Web-based test of interlanguage pragmatic knowledge: Implicatures, speech acts, and routines*. Unpublished manuscript, University of Hawai'i at Manoa.

Roever, C. (2006). Validation of a Web-Based Test of ESL Pragmalinguistics. *Language Testing, 23*(2), 229-256. doi: 10.1191/0265532206lt329oa

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20).

Safe exam browser. (2015). Retrieved from http://sourceforge.net/projects/seb/?source=typ_redirect and http://www.safeexambrowser.org/news_en.html

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology: A Refereed Journal for Second and Foreign Language Educators, 5*(2), 38-59.

Scully, J. (2006). Developing synergies in blended e-learning for language in higher education. *Malaysian Journal of Distance Education*, *8*(1), 89-101.

Scully, J. (2008). Blending Moodle for language learning: Thinking differently about language learning. *Moodle Majlis in the Middle East*, held at Sultan Qaboos University in Oman on 18/10/08. Retrieved from http://www.slideshare.net/burscu/blending-moodle-for-language-learning-presentation

Scully, J. (2013). A Moodlereader based extensive reading programme in SQU. *Moodle Majlis 2013*. Retrieved from http://www.slideshare.net/burscu/moodlereader-at-squ-for-moodlemajlis-28325493

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*(4), 331-345. doi: 10.1016/S0191-491X(98)00020-0

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. New York, NY: Longman.

Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education: Principles, Policy & Practice*, *14*(1), 117-130. doi: 10.1080/09695940701272948

Singer, L. M., & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education, 85*(1), 155-172. doi: 10.1080/00220973.2016.1143794

Staller, N. J. (2010). Qualitative research. In N. J. Salkind (Ed.), *Encyclopedia of research design qualitative research* (pp. 1159-1164). doi: http://dx.doi.org/10.4135/9781412961288.n350

Stoynoff, S. (2012). Research agenda: Priorities for future research in second language assessment. *Language Teaching, 45*(2), 234-249. doi: 10.1017/S026144481100053X

Stricker, L., Wilder, G., & Rock, D. (2004). Attitudes about the Computer-Based Test of English as a Foreign Language. *Computers in Human Behavior, 20*(1), 37-54.

Sunuodula, M., Feng, A., & Adamson, B. (2015). Trilingualism and Uyghur identity in the People's Republic of China. In D. Evans (Ed.), *Language and Identity: Discourse in the World* (pp. 81-104). UK: Bloomsbury Publishing. Retrieved from http://books.google.com.au/books

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction, 4*, 295-312. doi:10.1016/0959-4752(94)90003-5

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Pearson/Allyn & Bacon.

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. Retrieved from https://www.ets.org/Media/Research/pdf/RR-98-08.pdf

Taylor, C., Kirsch, I., Jamieson, J., & Eignor, D. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219-274. doi: 10.1111/0023-8333.00088

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes, 10*(2), 89-101. doi: 10.1016/j.jeap.2011.03.002

Toulmin, S. (2003). *The uses of argument*. Cambridge, UK: Cambridge University Press. Retrieved from http://books.google.com.au/books

Trusiewicz, D., Niesluchowska, M., & Makszewska-Chetnik, Z. (1995). Eye-strain symptoms after work with a computer screen. *Klinika Oczna, 97*(11-12), 343-345.

Uddin, A., Ahmar, F., & Al Raja, M. (2016). E-examinations for management students in Oman. *I J A B E R,* 14(1), 87-95. Retrieved from http://serialsjournals.com/serialjournalmanager/pdf/1459747283.pdf

United Nations Conference on Trade and Development. (2014). *The science, technology and innovation policy reviews: Oman*. United Nations. Retrieved from http://unctad.org/en/PublicationsLibrary/dtlstict2014d1_en.pdf

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27*(4), 493-513. doi: 10.1177/0265532209355668

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing, 29*(4), 603-619. doi: 10.1177/0265532212448619

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave Macmillan.

Weir, C. J., O'Sullivan, B., Jin, Y., & Bax, S. (2007). Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS Writing component: Effects and impact. *IEL TS Research Reports*, *7*, 311-347. Retrieved from https://www.ielts.org/~/media/research-reports/ielts_rr_volume07_report6.ashx

Wolfe, E. W., & Manalo, J. R. (2005). An investigation of the impact of composition medium on the quality of TOEFL writing scores (*TOEFL Research Report No. RR-72*). Princeton, NJ: ETS.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. Wilmington, DE: Wide Range Inc.

Xi, X. (2008). Methods of test validation. In E. Shohamy and N. H. Hornberger (Eds), *Encyclopedia of Language and Education, Vol.7* (pp. 177-196). Springer US.

Yamamoto, K. (1990). *HYBIL: A computer program to estimate HYBRID model parameters*. Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. *ETS Research Report Series*, i-39. Retrieved from https://www.ets.org/Media/Research/pdf/RR-95-02.pdf

Yan, Z., Hu, L., Chen, H., & Lu. F. (2008). Computer Vision Syndrome: A widely spreading but largely unknown epidemic among computer users. *Computers in Human Behavior,24*(5), 2026-2042.

Yang, X. (2009). *Effects of digital audio quality on students' performance in LAN delivered English listening comprehension tests,* ProQuest Dissertations and Theses.

Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly, 7*(2), 119-136. doi: 10.1080/15434300903452355

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233. doi: 10.1080/15434300701375832

Appendices

Appendix A: Study validation framework			
Overall study aim and research questions	**Overall study aim:** To provide a validity argument about using a Moodle-hosted test for its intended purpose by empirically examining reliability and construct validity evidence.	**RQ1:** To what extent can the Moodle-hosted test scores be reliable and valid indicators of the tested construct?	**RQ2:** To what extent can technology-related construct-irrelevant factors affect the reliability and construct validity of the Moodle-hosted test? Examined technology factors included: technology familiarity, typing ability, eye fatigue, test time and length, technical issues, attitude, equipment quality, reading tools, and layout and scrolling features.
Data collection	A variety of research and data collection instruments used in a mixed-method research paradigm to collect data to examine RQ1 and RQ2.	Test takers' overall scores on the Moodle-hosted test, subtests scores, and responses to individual items; and the item statistics report on Moodle	1) Moodle-hosted test score data was transferred from the Moodle Excel spreadsheet to SPSS; 2) Perceptions of stakeholders (test takers & invigilators) were elicited using questionnaires and audio-recorded semi-structured interviews (Appendices H to K, pp. 178-186). Retrospective verbal protocols of the test taking experience (for test takers) and the test invigilation experience (for invigilators) provided data on the effect of particular construct-irrelevant factors on test performance in the testing context; and 3) The researchers' reflective journals/reports recorded observations and field notes about the study.

Data analysis	A set of psychometric and non-psychometric data analysis procedures	Rasch item response theory item analysis on Winsteps software provided detailed reliability estimates and standard error of measurement (SEM) values for the whole test and for every item measure. This analysis also provided fit statistics results that gave discrimination information on how each item contributed to the tested construct and identified whether most items measure the targeted ability. Problematic items could also be identified. Through variable maps showing the match between item difficulty and person ability, instances of construct under-representation could be observed.	1) Testees’ selected-response questionnaire items were analysed statistically using SPSS software descriptive statistics and frequencies. 2) Testees’ test score data were linked to responses to selected-response questionnaire items on SPSS and reported in frequency, descriptive statistics, and boxplots to link test performance to feedback on the testing experience. 3) Open-ended constructed-response items from test takers’ questionnaires were analysed using thematic induction to look for patterns and common themes in the data. 4) Thematic induction was planned for other data types including invigilators’ questionnaires, interviews with test takers and invigilators, and the researchers' verbal protocols on the reflective journals. However, these data were not incorporated in the study reported in this thesis.
Validity backing evidence supporting validity assumptions	Multiple sources of evidence established from a variety of research instruments (data collection and analysis procedures) to support validity assumptions	Given high calculated reliability estimates, fit and highly discriminating items, and acceptable low SEM (evidence), the Moodle-hosted test score-based decisions will be reliable for the intended use (claim/assumption from scores).	Evidence should support the claim that the testing mode does not introduce construct-irrelevant variance or measurement error that may affect the reliability and construct validity of the Moodle-hosted test score-based decisions. Any potential issues pertinent to the use of Moodle to assess language can be resolved and should not act as sources of construct-irrelevant variance.
Validity concepts	overall validity argument	reliability; construct validity; construct-irrelevance; construct under-representation; measurement error	testing mode effect; technology-related issues; construct-irrelevant variance; test taker characteristics; impact; consequential validity; fairness; test bias; critical language testing; test power; feasibility or practicality

Warrant (if-then rule)	A set of warrants with supporting backing evidence	If test scores indicate high reliability estimates, fit and highly discriminating items, and acceptable low SEM, then it is warranted that the Moodle-hosted test scores will be reliable and valid indicators of the tested construct.	If evidence supports the claim that the testing mode does not introduce construct-irrelevant variance or measurement error, then reliability and construct validity of the Moodle-hosted test score-based decisions can be established. If the test is proved to measure test takers' English language ability only and examinees' test performance is not affected by construct-irrelevant technology-related variables in the Moodle-hosted testing environment, then test use can be justifiably supported. This type of warranted evidence should refute the rebuttal and support the use of the test based on the justified test score-based decisions.
Rebuttal (if-then rule)	A set of rebuttals with supporting refuting evidence to support assumptions	If test scores indicate low reliability estimates, misfit and low discriminating items, and unacceptable large SEM, then the Moodle-hosted test scores will not be as reliable and valid.	If quantitative and qualitative types of evidence support that construct-irrelevant factors related to the testing mode effect contribute to measurement error variance, then reliability and construct validity can be threatened. Therefore, the use of the test based on test score-based decisions will not be justified nor supported.

Appendix B: Ethical considerations and relevant forms (information sheets and consent forms for participants and ethics approval letters)

Information sheet for participants (UQ master students)

- **Title of the research:** A validation framework for an online English language Exit Test: A case study using Moodle as an assessment management system
- **Researcher:** Ms. Zakiya Al Nadabi, a teacher from the Language Centre at Sultan Qaboos University in Oman, who is currently doing this study for her PhD at the University of Queensland in Australia
- **Contact detail/s**: email: zakiya.alnadabi@uq.net.au

1. Project's purpose:

The study aims to trial an online English language test on Moodle and would like you to volunteer to participate in it.

2. Do I have to take part/ what are my rights as a participant?

Your participation is entirely voluntary and you have the right to withdraw from this study at any time provided that you inform the researcher. Please do this via the contact point shown above. Should you decide to withdraw from the study, the data collected from you after commencing the study will be destroyed and none of it will be used in this study.

3. What should I do, if I would like to take part?

If you agree to take part in this study, you will be involved in a 30 minutes judgmental validation session with others from your course. If you participate in this study, you will undertake a prototype online English Language Exit Test (ET) created in Moodle interface. This will be carried out in your class and you will be provided with a laptop and a USB stick to access the test. You will then be asked to provide feedback on the prototype by completing in a questionnaire. Your feedback will assist the researcher in transferring a working paper-based version of the ET to the Moodle platform for the main study. These procedures will help resolve usability issues in the online test interface. Your input will also help refine and validate the questionnaires for the later pilot-testing phase of the study.

4. Are there any disadvantages/ risks for me, if I take part in this study?

There are no foreseeable risks beyond that of everyday life due to your involvement in the study. On the contrary, your involvement will bring you benefits by being exposed to an online language test development which will be a valuable learning experience.

CONTINUE ON NEXT PAGE

5. How confidentially will my information be treated?

Please note that at the time you are sitting the exam and filling-in the questionnaire, your confidentiality and privacy cannot be maintained because you can be easily identified by your colleagues and teachers. However, the data collected and stored will be securely maintained and kept using pseudonyms or fake names to conceal your identity. Data will also be kept confidential by the use of password-protection for soft copies on the computer hard drive, flash disks, and CDs. Your data is confidential and no individual will be identified in the dissemination of the data. Data relating to you may be shared outside the immediate project team with staff members in Sultan Qaboos University, The University of Queensland and with research partners or related projects for the purposes of carrying out further research, research management or administration and quality control. Any sharing of data beyond the project team will be in aggregate or in a de-identified form to protect your privacy. De-identified data will be potentially made available to related projects to enable further analysis to be carried out. Selected de-identified segments of your data may be quoted in reports and publications.

6. How will the research results be revealed?

The research results will appear in a PhD dissertation and other publications and public presentations.

7. Who has ethically reviewed this research?

"This study has been cleared in accordance with the ethical review guidelines and processes of The University of Queensland. These guidelines are endorsed by the University's principal human ethics committee, the Human Experimentation Ethical Review Committee, and registered with the Australian Health Ethics Committee as complying with the National Statement. You are free to discuss your participation in this study with Principal Supervisor, Dr. Mathew Hillier (contactable on m.hillier@uq.edu.au). If you would like to speak to an officer of the University not involved in the study, you may contact the School Ethics Officer on 3365 6502."

This study has also been cleared in accordance with the ethical review guidelines and processes of the Language Centre at Sultan Qaboos University. If you have any ethical concerns about this study or your participation in it, please feel free to contact the Chair of the Language Centre Research Committee at the following address:

Faisal Said Al-Maamari, PhD
Language Centre, Sultan Qaboos University
Office # 1056; Extension # 2131; Email: faisalf@squ.edu.om

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

School of Education

CRICOS PROVIDER NUMBER 00025B

Participant consent form (UQ master students)

- Full title of research project: A validation framework for an online English language Exit Test: A case study using Moodle as an assessment management system

- Name, position, and contact address of Researcher establishing contact with participants: Ms. Zakiya Al Nadabi, a teacher from the Language Centre at Sultan Qaboos University in Oman, who is currently doing this study for her PhD at the University of Queensland in Australia; email: zakiya.alnadabi@uq.net.au

- Please tick (√) the following statements as appropriate.

□ 1. I confirm that I have read and understood the participant information sheet explaining the above research project and I have had the opportunity to ask questions about the project.

□ 2. I understand that my participation is voluntary and that I am free to withdraw at any time provided that I inform the researcher. I do not have to give any reason for my withdrawal and there will not be any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline. If I withdraw from this study, my data will be destroyed and will not be used in the study.

□ 3. I understand that my responses will be kept strictly confidential. I give permission to the researcher and others to have access to my anonymised responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.

□ 4. I agree for the data collected from me to be used in future research.

□ 5. I agree to take part in the above research project by participating in the judgmental validation session for the Moodle-hosted sample test prototype by sitting the test and filling in a questionnaire.

Name of Participant: ____ Date: ______________ Signature: _________________ Course/Section: _______________________________ Mobile: __________________ Email: ___________________________	Name of Researcher: **ZAKIYA AL NADABI** Date: _________________________ Signature: ____________________

Information sheet for participants in judgmental validation session

- **Title of the research:** An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences
- **Researcher:** Ms. Zakiya Al Nadabi, a teacher from the Language Centre at Sultan Qaboos University in Oman, who is currently doing this study for her PhD at the University of Queensland in Australia
- **Contact detail/s**: email: zakiyasa@squ.edu.om

1. Project's purpose:

The study aims to trial/pilot an online English language test hosted on Moodle and would like you to volunteer to participate in it.

2. Do I have to take part/ what are my rights as a participant?

Your participation is entirely voluntary and you have the right to withdraw from this study at any time provided that you inform the researcher. Please do this via the contact point shown above. Should you decide to withdraw from the study, the data collected from you after commencing the study will be destroyed and none of it will be used in this study.

3. What should I do, if I would like to take part?

If you agree to take part in this study, you will be involved in a judgmental validation session as part of a pilot study. If you participate in this pilot study, you will sit a 90 minutes online Moodle-hosted English language test in a computer laboratory with other fellow teachers. You will then be asked to provide feedback on the test by filling in a questionnaire and taking part in a follow-up audio-recorded focus group semi-structured interview that may last for 30-40 minutes. Your input will help resolve usability issues in the online test interface. Your feedback will also assist the researcher in validating and preparing the online test instrument and other study instruments for a larger main study to follow this pilot study.

4. Are there any disadvantages/ risks for me, if I take part in this study?

There are no foreseeable risks beyond that of everyday life due to your involvement in the study. On the contrary, your involvement will be an avenue for professional development and your contribution to the study will be valuable to the workplace.

5. How confidentially will my information be treated?

Please note that at the time you are sitting the exam, filling-in the questionnaire and taking part in interviews, your confidentiality and privacy cannot be maintained because you can be easily identified by your colleagues. However, the data collected and stored will be securely maintained and kept using pseudonyms or fake names to conceal your identity. Data will also be kept confidential by the use of password-protection for soft copies on the computer hard drive, flash disks, and CDs. Your data is confidential and no individual will be identified in the dissemination of the data. Data relating to you may be shared outside the immediate project team with staff members in Sultan Qaboos University, The University of Queensland and with research partners or related projects for the purposes of carrying out further research, research management or administration and quality control. Any sharing of data beyond the project team will be in aggregate or in a de-identified form to protect your privacy. De-identified data will be potentially made available to related projects to enable further analysis to be carried out. Selected de-identified segments of your data may be quoted in reports and publications.

6. How will the research results be revealed?

The research results will appear in a PhD dissertation and other publications and public presentations.

7. Who has ethically reviewed this research?

"This study has been cleared in accordance with the ethical review guidelines and processes of The University of Queensland. These guidelines are endorsed by the University's principal human ethics committee, the Human Experimentation Ethical Review Committee, and registered with the Australian Health Ethics Committee as complying with the National Statement. You are free to discuss your participation in this study with Principal Supervisor, Dr. Mathew Hillier (contactable on m.hillier@uq.edu.au*). If you would like to speak to an officer of the University not involved in the study, you may contact the School Ethics Officer on 3365 6502."*

This study has also been cleared in accordance with the ethical review guidelines and processes of the Language Centre at Sultan Qaboos University. If you have any ethical concerns about this study or your participation in it, please feel free to contact the Deputy Director for Professional Development and Research at the following address:

Faisal Said Al-Maamari, PhD
Language Centre, Sultan Qaboos University
Office # 1031; Extension # 1625; Email: faisalf@squ.edu.om

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

School of Education

CRICOS PROVIDER NUMBER 00025B

Consent form for participants in judgmental validation session

- **Title of the research:** An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences
- **Researcher:** Ms. Zakiya Al Nadabi, a teacher from the Language Centre at Sultan Qaboos University in Oman, who is currently doing this study for her PhD at the University of Queensland in Australia
- **Contact detail/s**: email: zakiyasa@squ.edu.om

- Please tick (√) the following statements as appropriate.

□ 1. I confirm that I have read and understood the participant information sheet explaining the above research project and I have had the opportunity to ask questions about the project.

□ 2. I understand that my participation is voluntary and that I am free to withdraw at any time provided that I inform the researcher. I do not have to give any reason for my withdrawal and there will not be any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline. If I withdraw from this study, my data will be destroyed and will not be used in the study.

□ 3. I understand that my responses will be kept strictly confidential. I give permission to the researcher and others to have access to my anonymised responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.

□ 4. I agree for the data collected from me to be used in future research.

□ 5. I agree to take part in the above research project by participating in the judgmental validation session in which I sit the online Moodle-hosted English language test and give feedback on a questionnaire and an audio-recorded focus group semi-structured interview.

Name of Participant: _____________________________ Date: ______________ Signature: _________________	Name of Researcher: **ZAKIYA AL NADABI** Date: ________________________ Signature: ____________________

Information sheet for usability study test takers

- **Title of the research:** An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences
- **Researcher:** Ms. Zakiya Al Nadabi, a teacher from the Language Centre at Sultan Qaboos University in Oman, who is currently doing this study for her PhD at the University of Queensland in Australia
- **Contact detail/s**: email: zakiyasa@squ.edu.om

1. Project's purpose:

The study aims to trial a web-based English language test on Moodle and would like you to volunteer to participate in it.

2. Do I have to take part/ what are my rights as a participant?

Your participation is entirely voluntary and you have the right to withdraw from this study at any time provided that you inform the researcher. Please do this via the contact point shown above. Should you decide to withdraw from the study, the data collected from you after commencing the study will be destroyed and none of it will be used in this study.

3. What should I do, if I would like to take part?

If you agree to take part in this study, you will be asked to sit a web-based English language test that may take about 90 minutes. The test will be conducted in a computer laboratory under supervised conditions with the presence of the researcher. After taking the exam, you will be asked to fill in a questionnaire reflecting your test-taking experience. You will also participate in a follow-up audio-recorded focus group interview for about 30-40 minutes in which you will be asked to talk about your test-taking experience.

4. Are there any disadvantages/ risks for me, if I take part in this study?

There are no foreseeable risks beyond that of everyday life due to your involvement in the study. On the contrary, you will benefit from your involvement by practicing test-taking when participating in the trial of the web-based English language test, which will be a valuable learning experience for you. Your course grades and academic performance will not be affected by your performance on the test as no decisions will be made based on this test score.

5. How confidentially will my information be treated?

Please note that at the time you are sitting the exam, filling-in the questionnaire, and taking part in interviews, your confidentiality and privacy cannot be maintained because you can be easily identified by others in the study context. However, the data collected and stored will be securely maintained and kept using pseudonyms or fake names to conceal your identity. Data will also be kept confidential by the use of password-protection for soft copies on the computer hard drive, flash disks, and CDs.Your data is confidential and no individual will be identified in the dissemination of the data. Data relating to you may be shared outside the immediate project team with staff members in Sultan Qaboos University, The University of Queensland and with research partners or related projects for the purposes of carrying out further research, research management or administration and quality control. Any sharing of data beyond the project team will be in aggregate or in a de-identified form to protect your privacy. De-identified data will be potentially made available to related projects to enable further analysis to be carried out. Selected de-identified segments of your data may be quoted in reports and publications.

6. How will the research results be revealed?

The research results will appear in a PhD dissertation and other publications and public presentations.

7. Who has ethically reviewed this research?

"This study has been cleared in accordance with the ethical review guidelines and processes of The University of Queensland. These guidelines are endorsed by the University's principal human ethics committee, the Human Experimentation Ethical Review Committee, and registered with the Australian Health Ethics Committee as complying with the National Statement. You are free to discuss your participation in this study with Principal Supervisor, Dr. Mathew Hillier (contactable on m.hillier@uq.edu.au). If you would like to speak to an officer of the University not involved in the study, you may contact the School Ethics Officer on 3365 6502."

This study has also been cleared in accordance with the ethical review guidelines and processes of the Language Centre at Sultan Qaboos University. If you have any ethical concerns about this study or your participation in it, please feel free to contact the Deputy Director for Professional Development and Research at the following address:
Faisal Said Al-Maamari, PhD
Language Centre, Sultan Qaboos University
Office # 1031; Extension # 1625; Email: faisalf@squ.edu.om

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

School of Education

CRICOS PROVIDER NUMBER 00025B

STUDENT CONSENT FORM (USABILITY STUDY) **Researcher: Sections A & D ONLY**	**استمارة طالب للموافقة على المشاركة في بحث علمي** **الطالب: الأجزاء (ب) و (ج، إذا توفر)**

A) Full title of research project:
An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences
Name, position, and contact address of Researcher establishing contact with students:
Ms. Zakiya Al Nadabi, a teacher from the Language Centre who is currently doing this study for her PhD at the University of Queensland in Australia; email: zakiyasa@squ.edu.om

B) Tick (√) as appropriate.			**أ) ضع علامة (√) كما يناسب.**
I confirm that I have read and understood the information sheet for the above research project and have had the opportunity to ask questions.	**No لا** □	**Yes نعم** □	أؤكد أنني قد قرأت وفهمت ورقة المعلومات المعدة للبحث أعلاه وتمت إتاحة الفرصة الكاملة لي لطرح الأسئلة.
I understand that my participation is voluntary and that I am free to withdraw at any time provided that I inform the researcher. There will be no penalty and I do not need to give a reason. In addition, should I not wish to answer any particular question or questions, I am free to decline. If I withdraw from this study, my data will be destroyed and will not be used in the study.	**No لا** □	**Yes نعم** □	أعي بأن مشاركتي في البحث اختيارية وأنه يمكنني الانسحاب في أي وقت بعد إعلام الباحث بذلك بدون الحاجة الى إعطاء أي تفسير وهذا يعني بأن البيانات التي تم تجميعها مني لن تستخدم في هذا البحث.
I understand that anonymized quotes of any data I provide for this research project may be used in future publications and that my de-identified data may be accessed by other researchers.	**No لا** □	**Yes نعم** □	أعي أن البيانات التي سأدلي بها في هذا البحث ستستخدم لأغراض النشر العلمي وسيقرؤها الآخرون وسيطلع على هذه البيانات باحثون آخرون شريطة الحفاظ على السرية التامة وعدم الإدلاء باسمي علنا.
I understand that if I participate/not participate in the research project that my marks will NOT be affected in any way.	**No لا** □	**Yes نعم** □	أعي أن مشاركتي أو عدم مشاركتي في البحث لن تؤثر البتة على علاماتي الدراسية بالجامعة أو المركز سلبا كان أم ايجابا.
I agree to take part in the research study by sitting the web-based English language test, filling in a questionnaire, and participating in a follow-up focus group interview. I agree to the interview being audio recorded.	**No لا** □	**Yes نعم** □	أقر بموافقتي على المشاركة في هذا البحث العلمي وذلك بأداء اختبار اللغة الانجليزية المذكور وملئ الإستبانة المرافقة. أقرأيضا بموافقتي على المشاركة في مقابلة البحث الجماعية وعلى التسجيل الصوتي لهذه المقابلة.

C) Name of Participant: اسم المشترك: _____________________________________ Course/Section: ________________________ Mobile: ______________________________ Email: ________________________________	Date: التاريخ: __________________________	Signature: ب) التوقيع: _____________________
D) Name of Researcher: اسم الباحثة: **Zakiya Al Nadabi**	Date: التاريخ:	Signature: التوقيع:

Information sheet for examinees in main study

- **Title of the research:** An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences
- **Researcher:** Ms. Zakiya Al Nadabi, a teacher from the Language Centre at Sultan Qaboos University in Oman, who is currently doing this study for her PhD at the University of Queensland in Australia
- **Contact detail/s**: email: zakiyasa@squ.edu.om

1. Project's purpose:

The study aims to trial a web-based English language test on Moodle and would like you to volunteer to participate in it.

2. Do I have to take part/ what are my rights as a participant?

Your participation is entirely voluntary and you have the right to withdraw from this study at any time provided that you inform the researcher. Please do this via the contact point shown above. Should you decide to withdraw from the study, the data collected from you after commencing the study will be destroyed and none of it will be used in this study.

3. What should I do, if I would like to take part?

If you agree to take part in this study, you will be asked to sit a web-based English language test that may take about 90 minutes. The test will be conducted in a computer laboratory under supervised conditions with the presence of one to two invigilators. After taking the exam, you will be asked to fill in a questionnaire reflecting your test-taking experience. You will also be invited to participate in an audio-recorded follow-up interview for about 30 minutes in which you will be asked to talk about your test-taking experience. The interview will either be conducted in a group or individually depending on your arrangement with the researcher.

4. Are there any disadvantages/ risks for me, if I take part in this study?

There are no foreseeable risks beyond that of everyday life due to your involvement in the study. On the contrary, you will benefit from your involvement by practicing test-taking when participating in the trial of the web-based English language test, which will be a valuable learning experience for you. Your course grades and academic performance will not be affected by your performance on the test as no decisions will be made based on this test score.

5. How confidentially will my information be treated?

Please note that at the time you are sitting the exam, filling-in the questionnaire, and taking part in interviews, your confidentiality and privacy cannot be maintained because you can be easily identified by others in the study context. However, the data collected and stored will be securely maintained and kept using pseudonyms or fake names to conceal your identity. Data will also be kept confidential by the use of password-protection for soft copies on the computer hard drive, flash disks, and CDs.Your data is confidential and no individual will be identified in the dissemination of the data. Data relating to you may be shared outside the immediate project team with staff members in Sultan Qaboos University, The University of Queensland and with research partners or related projects for the purposes of carrying out further research, research management or administration and quality control. Any sharing of data beyond the project team will be in aggregate or in a de-identified form to protect your privacy. De-identified data will be potentially made available to related projects to enable further analysis to be carried out. Selected de-identified segments of your data may be quoted in reports and publications.

6. How will the research results be revealed?

The research results will appear in a PhD dissertation and other publications and public presentations.

7. Who has ethically reviewed this research?

"This study has been cleared in accordance with the ethical review guidelines and processes of The University of Queensland. These guidelines are endorsed by the University's principal human ethics committee, the Human Experimentation Ethical Review Committee, and registered with the Australian Health Ethics Committee as complying with the National Statement. You are free to discuss your participation in this study with Principal Supervisor, Dr. Mathew Hillier (contactable on m.hillier@uq.edu.au). If you would like to speak to an officer of the University not involved in the study, you may contact the School Ethics Officer on 3365 6502."

This study has also been cleared in accordance with the ethical review guidelines and processes of the Language Centre at Sultan Qaboos University. If you have any ethical concerns about this study or your participation in it, please feel free to contact the Deputy Director for Professional Development and Research at the following address:

Faisal Said Al-Maamari, PhD
Language Centre, Sultan Qaboos University
Office # 1031; Extension # 1625; Email: faisalf@squ.edu.om

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

School of Education

CRICOS PROVIDER NUMBER 00025B

STUDENT CONSENT FORM **Researcher: Sections A & D ONLY**			**استمارة طالب للموافقة على المشاركة في بحث علمي** **الطالب: الأجزاء (ب) و (ج، إذا توفر)**
A) Full title of research project: An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences **Name, position, and contact address of Researcher establishing contact with students:** Ms. Zakiya Al Nadabi, a teacher from the Language Centre who is currently doing this study for her PhD at the University of Queensland in Australia; email: zakiyasa@squ.edu.om			
B) Tick (√) as appropriate.			**ب) ضع علامة (√) كما يناسب.**
I confirm that I have read and understood the information sheet for the above research project and have had the opportunity to ask questions.	**No لا** □	**Yes نعم** □	أؤكد أنني قد قرأت وفهمت ورقة المعلومات المعدة للبحث أعلاه وتمت إتاحة الفرصة الكاملة لي لطرح الأسئلة.
I understand that my participation is voluntary and that I am free to withdraw at any time provided that I inform the researcher. There will be no penalty and I do not need to give a reason. In addition, should I not wish to answer any particular question or questions, I am free to decline. If I withdraw from this study, my data will be destroyed and will not be used in the study.	**No لا** □	**Yes نعم** □	أعي بأن مشاركتي في البحث اختيارية وأنه يمكنني الانسحاب في أي وقت بعد إعلام الباحث بذلك بدون الحاجة الى إعطاء أي تفسير وهذا يعني بأن البيانات التي تم تجميعها مني لن تستخدم في هذا البحث.
I understand that anonymized quotes of any data I provide for this research project may be used in future publications and that my de-identified data may be accessed by other researchers.	**No لا** □	**Yes نعم** □	أعي أن البيانات التي سأدلي بها في هذا البحث ستستخدم لأغراض النشر العلمي وسيقرؤها الآخرون وسيطلع على هذه البيانات باحثون آخرون شريطة الحفاظ على السرية التامة وعدم الإدلاء باسمي علنا.
I understand that if I participate/not participate in the research project that my marks will NOT be affected in any way.	**No لا** □	**Yes نعم** □	أعي أن مشاركتي أو عدم مشاركتي في البحث لن تؤثر البتة على علاماتي الدراسية بالجامعة أو المركز سلبا كان أم ايجابا.
I agree to take part in the research study by sitting the web-based English language test and filling in a questionnaire on my test-taking experience.	**No لا** □	**Yes نعم** □	أقر بموافقتي على المشاركة في هذا البحث العلمي وذلك بأداء اختبار اللغة الانجليزية المذكور وملئ الإستبانة المرافقة.
I also agree to participate in a follow-up interview to talk about my test-taking experience. I agree to the interview being audio recorded.	**No لا** □	**Yes نعم** □	أقرأيضا بموافقتي على المشاركة في مقابلة البحث وعلى التسجيل الصوتي لهذه المقابلة.

C) Name of Participant: اسم المشترك: ______________________________________ Course/Section: ________________________ Mobile: ______________________________ Email: _______________________________	Date: التاريخ: ___________________________	Signature: ج) التوقيع: _____________________
D) Name of Researcher: اسم الباحثة: **Zakiya Al Nadabi**	Date: التاريخ:	Signature: التوقيع:

Information sheet for invigilators

- **Title of the research:** An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences
- **Researcher:** Ms. Zakiya Al Nadabi, a teacher from the Language Centre at Sultan Qaboos University in Oman, who is currently doing this study for her PhD at the University of Queensland in Australia
- **Contact detail/s**: email: zakiyasa@squ.edu.om

1. Project's purpose:

The study aims to trial a web-based English language test hosted on Moodle and would like you to volunteer to participate in it.

2. Do I have to take part/ what are my rights as a participant?

Your participation is entirely voluntary and you have the right to withdraw from this study at any time provided that you inform the researcher. Please do this via the contact point shown above. Should you decide to withdraw from the study, the data collected from you after commencing the study will be destroyed and none of it will be used in this study.

3. What should I do, if I would like to take part?

If you agree to take part in this study, you will supervise/invigilate a group of test takers while they sit a web-based English language test that may take approximately 90 minutes in a computer laboratory. After invigilating the exam, you will be required to fill in a questionnaire reflecting your test invigilation experience. You will also be invited to volunteer to participate in a follow-up audio-recorded semi-structured interview with the researcher for about 30 minutes in which you will be asked to talk about your test invigilation experience. The interview will either be conducted in a group (with other fellow teachers who have invigilated the test) or individually depending on your arrangement with the researcher.

4. Are there any disadvantages/ risks for me, if I take part in this study?

There are no foreseeable risks beyond that of everyday life due to your involvement in the study. On the contrary, your involvement will be an avenue for professional development and your contribution to the study will be valuable to the workplace. Through your involvement, you will be able to have your say on the testing experience and voice your concerns and observations for future applications of web-based testing of English language ability to take major high-stakes decisions about students' language proficiency.

5. How confidentially will my information be treated?

Please note that at the time you are invigilating the exam, filling-in the questionnaire, and taking part in the interview, your confidentiality and privacy cannot be maintained because you can be easily identified by students and colleagues at the study context. However, the data collected and stored will be securely maintained and kept using pseudonyms or fake names to conceal your identity. Data will also be kept confidential by the use of password-protection for soft copies on the computer hard drive, flash disks, and CDs. Your data is confidential and no individual will be identified in the dissemination of the data. Data relating to you may be shared outside the immediate project team with staff members in Sultan Qaboos University, The University of Queensland and with research partners or related projects for the purposes of carrying out further research, research management or administration and quality control. Any sharing of data beyond the project team will be in aggregate or in a de-identified form to protect your privacy. De-identified data will be potentially made available to related projects to enable further analysis to be carried out. Selected de-identified segments of your data may be quoted in reports and publications.

6. How will the research results be revealed?

The research results will appear in a PhD dissertation and other publications and public presentations.

7. Who has ethically reviewed this research?

"This study has been cleared in accordance with the ethical review guidelines and processes of The University of Queensland. These guidelines are endorsed by the University's principal human ethics committee, the Human Experimentation Ethical Review Committee, and registered with the Australian Health Ethics Committee as complying with the National Statement. You are free to discuss your participation in this study with Principal Supervisor, Dr. Mathew Hillier (contactable onm.hillier@uq.edu.au). If you would like to speak to an officer of the University not involved in the study, you may contact the School Ethics Officer on 3365 6502."

This study has also been cleared in accordance with the ethical review guidelines and processes of the Language Centre at Sultan Qaboos University. If you have any ethical concerns about this study or your participation in it, please feel free to contact the Deputy Director for Professional Development and Research at the following address:

Faisal Said Al-Maamari, PhD
Language Centre, Sultan Qaboos University
Office # 1031; Extension # 1625; Email: faisalf@squ.edu.om

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

School of Education

CRICOS PROVIDER NUMBER 00025B

Consent form for invigilators

- **Title of the research:** An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences
- **Researcher:** Ms. Zakiya Al Nadabi, a teacher from the Language Centre at Sultan Qaboos University in Oman, who is currently doing this study for her PhD at the University of Queensland in Australia
- **Contact detail/s**: email: zakiyasa@squ.edu.om

- Please tick (√) the following statements as appropriate.

□ 1. I confirm that I have read and understood the participant information sheet explaining the above research project and I have had the opportunity to ask questions about the project.

□ 2. I understand that my participation is voluntary and that I am free to withdraw at any time provided that I inform the researcher. I do not have to give any reason for my withdrawal and there will not be any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline. If I withdraw from this study, my data will be destroyed and will not be used in the study.

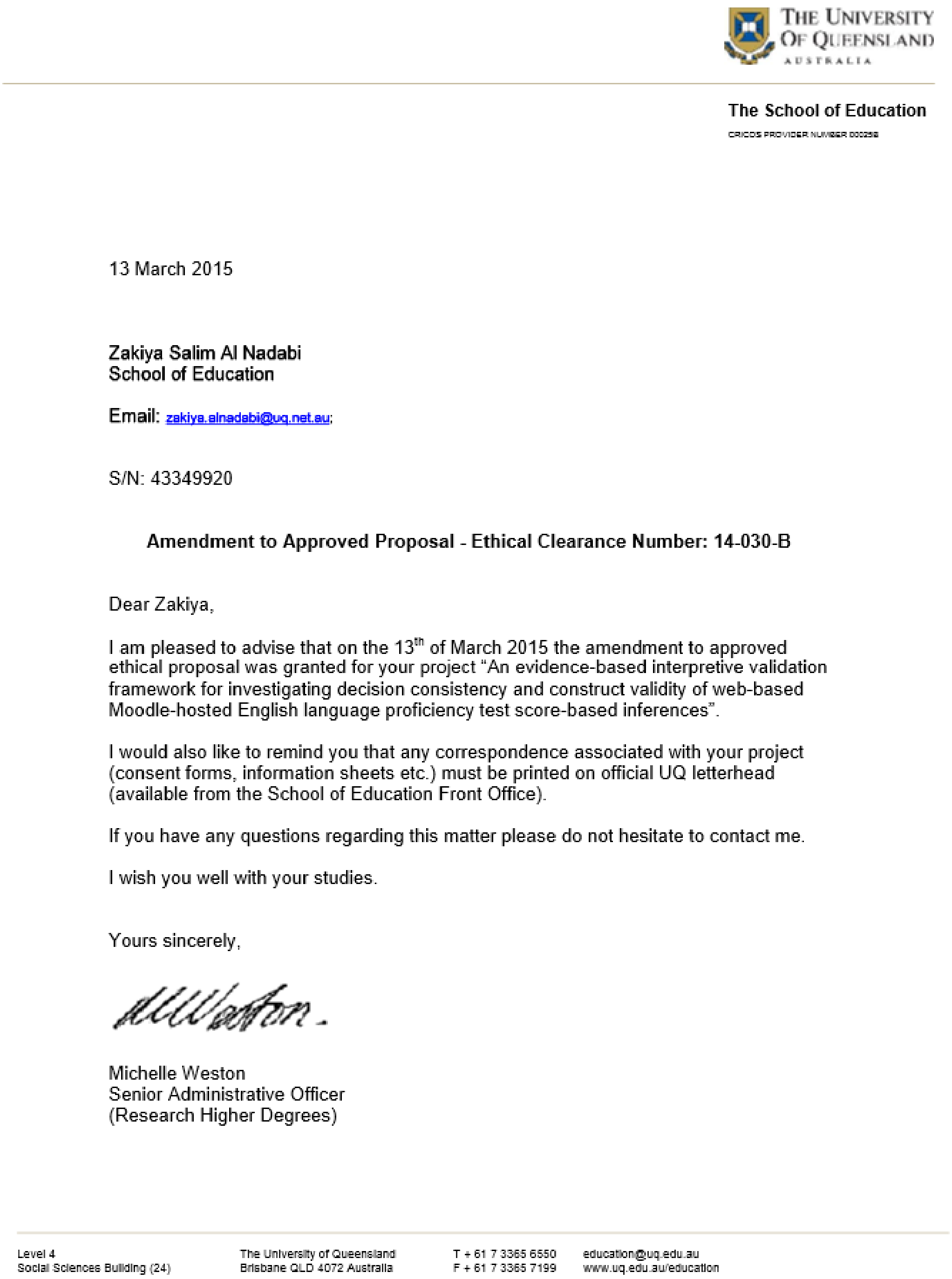
□ 3. I understand that my responses will be kept strictly confidential. I give permission to the researcher and others to have access to my anonymised responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.

□ 4. I agree for the data collected from me to be used in future research.

□ 5. I agree to take part in the above research project by invigilating the web-based English language test in a computer laboratory and filling in a follow-up questionnaire on the test invigilation experience.

□ 6. I also agree to take part in the above research project by participating in a follow-up audio-recorded interview to talk about the test invigilation experience.

Name of Participant: ____ Date: _________________________ Signature: _________________________	Name of Researcher: **ZAKIYA AL NADABI** Date: _________________________ Signature: _________________________



The University Of Queensland
Australia

The School of Education

CRICOS PROVIDER NUMBER 00025B

13 March 2015

Zakiya Salim Al Nadabi
School of Education

Email: zakiya.alnadabi@uq.net.au;

S/N: 43349920

Amendment to Approved Proposal - Ethical Clearance Number: 14-030-B

Dear Zakiya,

I am pleased to advise that on the 13th of March 2015 the amendment to approved ethical proposal was granted for your project "An evidence-based interpretive validation framework for investigating decision consistency and construct validity of web-based Moodle-hosted English language proficiency test score-based inferences".

I would also like to remind you that any correspondence associated with your project (consent forms, information sheets etc.) must be printed on official UQ letterhead (available from the School of Education Front Office).

If you have any questions regarding this matter please do not hesitate to contact me.

I wish you well with your studies.

Yours sincerely,

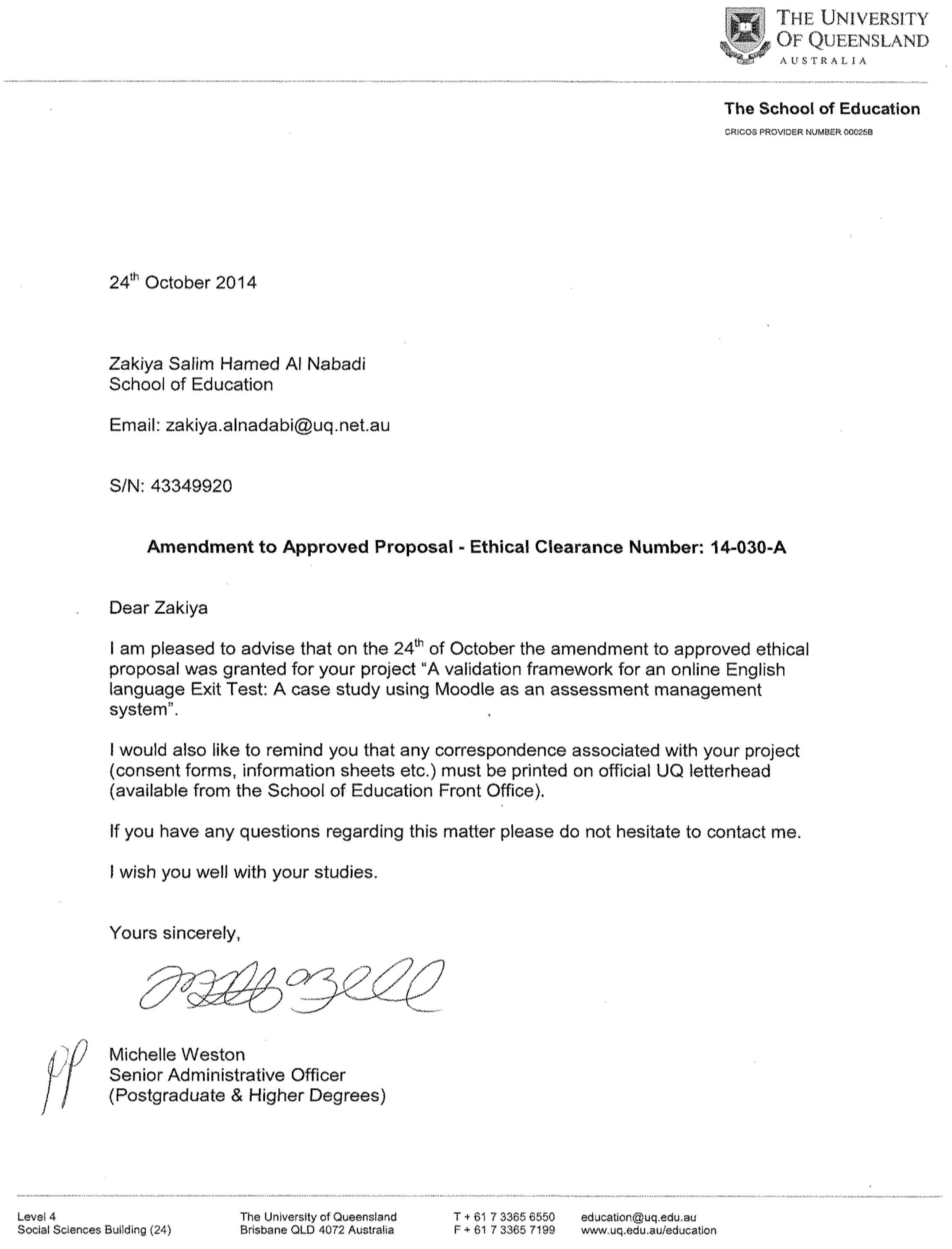
Michelle Weston
Senior Administrative Officer
(Research Higher Degrees)

Level 4
Social Sciences Building (24)

The University of Queensland
Brisbane QLD 4072 Australia

T + 61 7 3365 6550
F + 61 7 3365 7199

education@uq.edu.au
www.uq.edu.au/education



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

The School of Education
CRICOS PROVIDER NUMBER 00025B

24th October 2014

Zakiya Salim Hamed Al Nabadi
School of Education

Email: zakiya.alnadabi@uq.net.au

S/N: 43349920

Amendment to Approved Proposal - Ethical Clearance Number: 14-030-A

Dear Zakiya

I am pleased to advise that on the 24th of October the amendment to approved ethical proposal was granted for your project "A validation framework for an online English language Exit Test: A case study using Moodle as an assessment management system".

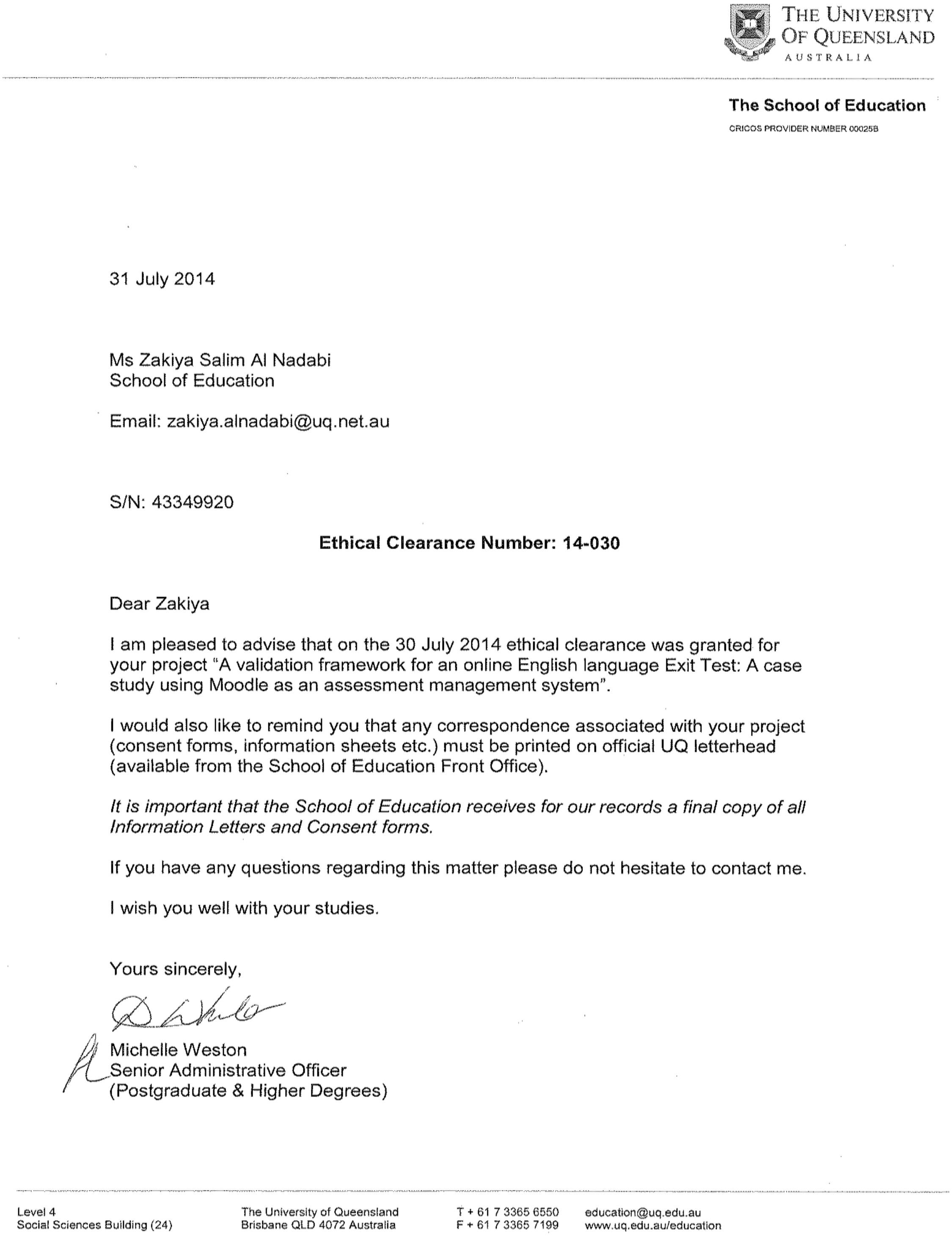
I would also like to remind you that any correspondence associated with your project (consent forms, information sheets etc.) must be printed on official UQ letterhead (available from the School of Education Front Office).

If you have any questions regarding this matter please do not hesitate to contact me.

I wish you well with your studies.

Yours sincerely,

pp
Michelle Weston
Senior Administrative Officer
(Postgraduate & Higher Degrees)



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

The School of Education
CRICOS PROVIDER NUMBER 00025B

31 July 2014

Ms Zakiya Salim Al Nadabi
School of Education

Email: zakiya.alnadabi@uq.net.au

S/N: 43349920

Ethical Clearance Number: 14-030

Dear Zakiya

I am pleased to advise that on the 30 July 2014 ethical clearance was granted for your project "A validation framework for an online English language Exit Test: A case study using Moodle as an assessment management system".

I would also like to remind you that any correspondence associated with your project (consent forms, information sheets etc.) must be printed on official UQ letterhead (available from the School of Education Front Office).

It is important that the School of Education receives for our records a final copy of all Information Letters and Consent forms.

If you have any questions regarding this matter please do not hesitate to contact me.

I wish you well with your studies.

Yours sincerely,

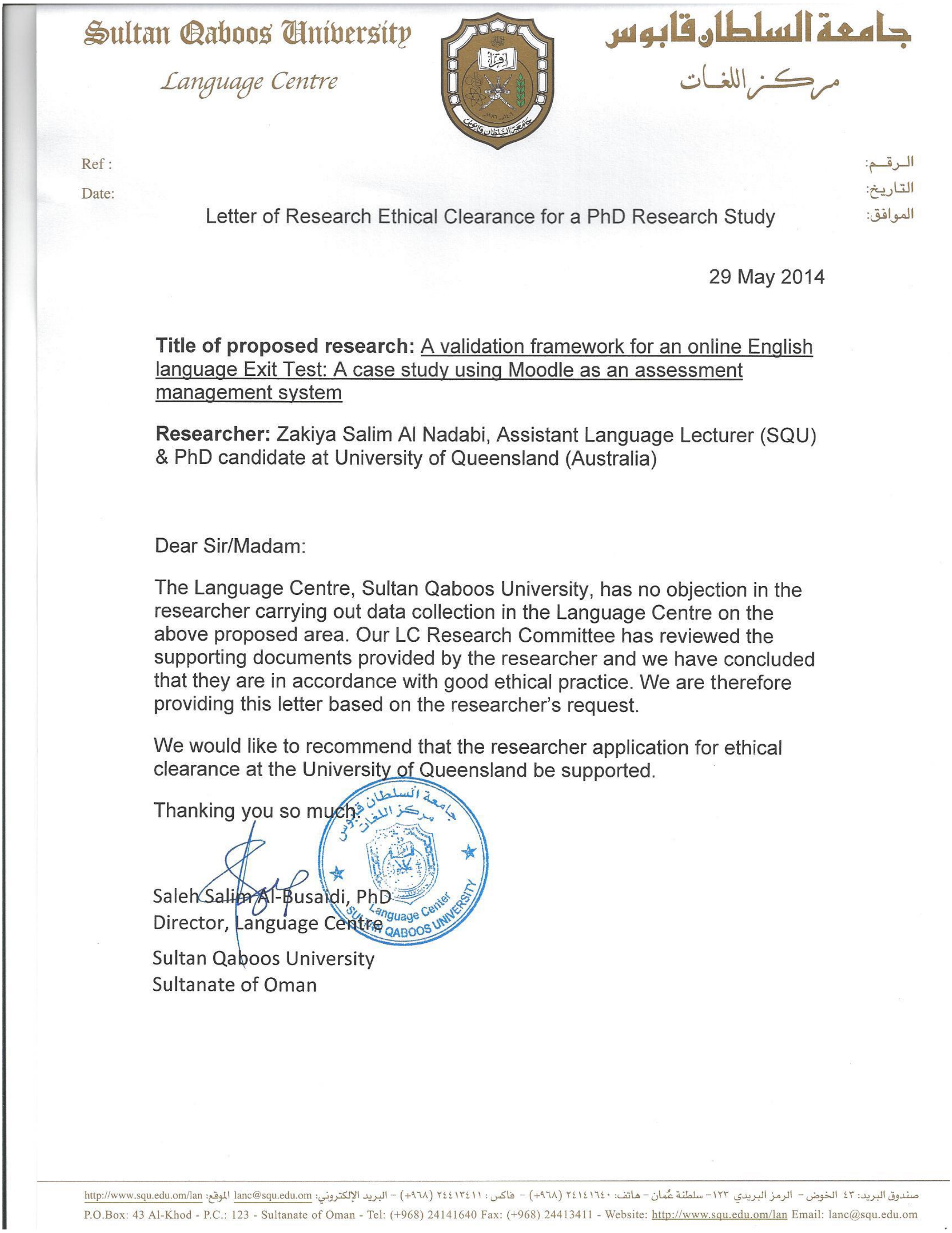
Michelle Weston
Senior Administrative Officer
(Postgraduate & Higher Degrees)

Level 4
Social Sciences Building (24)

The University of Queensland
Brisbane QLD 4072 Australia

T + 61 7 3365 6550
F + 61 7 3365 7199

education@uq.edu.au
www.uq.edu.au/education



Sultan Qaboos University

Language Centre

جامعة السلطان قابوس

مركز اللغات

Ref : الرقم:

Date: التاريخ:

الموافق:

Letter of Research Ethical Clearance for a PhD Research Study

29 May 2014

Title of proposed research: A validation framework for an online English language Exit Test: A case study using Moodle as an assessment management system

Researcher: Zakiya Salim Al Nadabi, Assistant Language Lecturer (SQU) & PhD candidate at University of Queensland (Australia)

Dear Sir/Madam:

The Language Centre, Sultan Qaboos University, has no objection in the researcher carrying out data collection in the Language Centre on the above proposed area. Our LC Research Committee has reviewed the supporting documents provided by the researcher and we have concluded that they are in accordance with good ethical practice. We are therefore providing this letter based on the researcher's request.

We would like to recommend that the researcher application for ethical clearance at the University of Queensland be supported.

Thanking you so much.

جامعة السلطان قابوس
مركز اللغات
Language Center
SULTAN QABOOS UNIVERSITY

Saleh Salim Al-Busaidi, PhD
Director, Language Centre
Sultan Qaboos University
Sultanate of Oman

صندوق البريد: ٤٣ الخوض – الرمز البريدي ١٢٣– سلطنة عُمان – هاتف: ٢٤١٤١٦٤٠ (+٩٦٨) – فاكس : ٢٤٤١٣٤١١ (+٩٦٨) – البريد الإلكتروني: lanc@squ.edu.om الموقع: http://www.squ.edu.om/lan

P.O.Box: 43 Al-Khod - P.C.: 123 - Sultanate of Oman - Tel: (+968) 24141640 Fax: (+968) 24413411 - Website: http://www.squ.edu.om/lan Email: lanc@squ.edu.om

Appendix C: Questionnaire for UQ students in the first prototype trial

Thank you for taking part in this study. Now that you have finished taking the online test on Moodle, we would like you to fill in this questionnaire about your test experience. We truly appreciate and value your feedback.

Questions 1 – 4: Background information (Bio data):

1) Gender:
   - □ Male
   - □ Female
2) Age (years):
   - □ ____________________
3) Your level of familiarity (high, average, low, none) with tests or quizzes on Moodle:
   - □ High
   - □ Average
   - □ Low
   - □ None
4) Your level of computer-literacy or familiarity with computers (high, average, low):
   - □ High
   - □ Average
   - □ Low

About the exam system (5 = **agree** strongly, 1 = strongly **disagree**)

Please indicate your level of agreement with each statement:	5	4	3	2	1
5) Overall my experience of this exam was positive.					
6) I ran out of time.					
7) I felt this particular exam suited the use of computers.					
8) I felt the e-exam system was easy to use.					
9) I felt the e-exam system was reliable against technical failures.					
10) I felt the e-exam system was secure against cheating.					
11) I would recommend the e-exam system to others.					

PLEASE CONTINUE ON NEXT PAGE

The test-taking experience (5 = **agree** strongly, 1 = strongly **disagree**)

Please indicate your level of agreement with each statement:	5	4	3	2	1
12) I liked the Moodle-hosted exam.					
13) I did not have any technical issues with the test.					
14) The test navigation system was clear and easy to follow.					
15) Test procedures and instructions given were clear and easy to follow.					
16) I liked that Moodle marked my responses right away and showed me instant feedback immediately upon submission.					
17) I liked typing my responses for some questions. (ignore if not applicable)					
18) The listening test sound was of good quality. (ignore if not applicable)					
19) I did not have any technical problems with the listening audio files. (ignore if not applicable)					

20) If you have any other specific comments on your experience taking the Moodle-hosted test, you can use the space provided below or extra paper if needed.

Opinions about Moodle and the test (5 = **agree** strongly, 1 = strongly **disagree**)

Please indicate your level of agreement with each statement:	5	4	3	2	1
21) I would support using Moodle to automatically mark objective test items.					
22) I would support using Moodle to automatically mark short answer- (20 words or less) test items.					
23) I think using Moodle for automatic marking would be more convenient than manual marking on paper.					
24) Receiving immediate feedback on Moodle test results is very useful to test takers.					
25) I would recommend the use of Moodle-hosted exams to take decisions about students' levels of language proficiency.					
26) I would recommend the use of Moodle to run computerised tests.					

27) What suggestions would you like to make in order to improve test-taking on Moodle?

End of Questionnaire – Thank you

Appendix D: Questionnaire for judgmental validation session participants

Thank you for taking part in this study. Now that you have finished taking the online test on Moodle, we would like you to fill in this questionnaire about your test experience. We truly appreciate and value your feedback.

Background information:

1) Gender:

- □ Male
- □ Female

2) Current course:

- □ ____________________

For questions 3 - 8, please answer the following open-ended questions from your experience with this Moodle-hosted test.

3) What do you think of the Moodle-hosted test overall?

____

____

____

4) Did you like/dislike your test experience? Why?

____

____

____

Please continue on next page

5) From your test experience, do you think it is practical to run the tests on Moodle? Why or why not? What technical issues did you face while taking the exam? Were there any problems with the network and loading of Moodle pages or login?

____

____

____

____

6) What do you think about the use of Moodle to run official exams? Would you recommend Moodle to be used to run official exams like mid-terms, finals, placement tests, exit tests, and so forth) to take decisions about students' levels of language proficiency? Why or why not?

____

____

____

____

7) As in the case of this Moodle-hosted exam, Moodle can be used to run objective exams and automatically mark students' responses (even short-answer), and thus relieving teachers from marking duties (marking and double-marking) they usually have. What do you think of the potential of relieving teachers from test marking duties? Would you support this testing practice? Why or why not?

____

____

____

____

8) What suggestions do you have to improve test-taking on Moodle?

____

____

____

____

End of Questionnaire – Thank you!

Appendix E: Semi-structured interview for judgmental validation participants

The following is a set of questions for a semi-structured interview conducted with the researcher being the moderator (asking questions) and the judgmental validation participants being the interviewees. The interview was conducted in a group after these participants trialled the test. Questions were rephrased to explain them to participants and were followed with other probing and follow-up questions.

1) What do you think of the Moodle-hosted test overall?
2) What did you like about your test experience? Why?
3) What did you dislike about your test experience? Why?
4) What technical issues did you face during the exam? Were there any problems with the network and loading of Moodle pages or login?
5) Do you think that the computer labs are well-equipped and efficient for taking tests on Moodle?
6) From your test experience, do you think it is practical to run the tests on Moodle? Why or why not?
7) What do you think about the use of Moodle to run official exams? Would you like Moodle to be used to run official exams like mid-terms, finals, placement tests, exit tests, and so forth) to take decisions about students' levels of language proficiency? Why or why not?
8) As in the case of this Moodle-hosted test, Moodle can be used to run objective exams and automatically mark students' responses (even short-answer), and thus relieving invigilators from the post-invigilation marking duties (marking and double-marking) they usually have. What do you think of the potential of relieving you as a teacher in the LC from marking duties of the objective tests? Would you support this testing practice? Why or why not?
9) What suggestions do you have to improve test-taking on Moodle and invigilation or supervision of such exams?

Appendix F: Usability study test takers' questionnaires

Questionnaire used on the 12th and 13th of April, 2015

Name: ________________________________ ID: __________ Section: ______

Questionnaire for Test Takers **استبانة للطلبة المؤدين للامتحان**

Thank you for taking part in this study. Now that you have finished taking the online test on the learning management system Moodle, we would like you to fill in this questionnaire about your test-taking experience. We appreciate your feedback on all questions.

نشكركم على مشاركتكم في هذا المشروع البحثي. بعدما أنهيتم أداء هذا الامتحان على المودل الرجاء تعبئة هذه الاستبانة عن تجربتكم لهذا الامتحان. شاكرين لكم إجابتكم على جميع الأسئلة وعلى تعليقاتكم وآراءكم.

For questions 1 – 4, please check (✔) all that apply to you. Please select only one answer for each question. **الرجاء اختيار ما يناسبك للإجابة على الأسئلة التالية.**

1) Your current course of study and level **المستوى أو المقرر الذي تدرسه حاليا**

□ FPEL0340 (level 4)
□ FPEL0450 (level 5)
□ FPEL0560 (level 6)
□ FPEL0604 (level 6)

2) Gender: **الجنس**

□ Male ذكر □
□ Female أنثى □

3) Your level of familiarity with tests or quizzes on Moodle:
مدى معرفتك للامتحانات المستخدمة في المودل

□ Very familiar أعرفها جيدا □
□ Somehow familiar أعرفها بعض الشي □
□ A little bit familiar أعرفها قليلا □
□ Not familiar at all لا أعرفها مطلقا □

4) Your level of computer-literacy or familiarity with computers:
مدى معرفتك أو خبرتك باستخدامات جهاز الحاسب الآلي

□ Very familiar لدي خبرة واسعة □
□ Somehow familiar لدي خبرة بعض الشي □
□ A little bit familiar خبرتي ضئيلة جدا □
□ Not familiar at all ليس لدي أية خبرة مطلقا □

Please continue on the next page. الرجاء المتابعة في الصفحة التالية.

For questions 5 – 14, circle the option (5 = strongly agree; 4= agree; 3= neutral; 2 = disagree; 1= strongly disagree) that best applies to you.

الرجاء الإجابة على الأسئلة التالية باختيار أحد الأرقام من 1 – 5 الذي يعبر عن رأيك. 5= أوافق جدا 4= أوافق 3=محايد 2=لا أوافق 1=أوافق بشدة.

No.	**Section 1: Overall test-taking experience** تجربتك لهذا الامتحان بشكل عام	**5** **Strongly agree** أوافق جدا	**4** **Agree** أوافق	**3** **Neutral** محايد	**2** **Disagree** لا أوافق	**1** **Strongly disagree** لا أوافق بشدة
5)	Overall, I liked this test-taking experience. أعجبني هذا الامتحان بشكل عام.	5	4	3	2	1
6)	Overall, the test was easy to navigate by moving from one page displaying a subtest to another. بصورة عامة كان من السهل الانتقال من صفحة إلى أخرى لأداء أجزاء الامتحان.	5	4	3	2	1
7)	Test timing was sufficient for all test sections. كان الوقت المخصص للامتحان كافيا للإجابة على جميع الأسئلة.	5	4	3	2	1
8)	Sound quality of the listening tests was good. صوت التسجيل في امتحان الاستماع كان جيدا.	5	4	3	2	1
9)	I liked that Moodle showed me instant feedback/test results at the end of the test. أعجبني أن نتائج الامتحان وإجاباتي للأسئلة ظهرت على الشاشة بعد الانتهاء من الامتحان مباشرة.	5	4	3	2	1
10)	I liked seeing my marks on all test questions as well as the overall test score. أعجبني أنه كان بامكاني رؤية الدرجات التي حصلت عليها في كل سؤال والدرجة الكلية بالامتحان.	5	4	3	2	1
11)	I liked typing my responses for some questions. أعجبني أن أقوم بطباعة إجاباتي لبعض الأسئلة.	5	4	3	2	1
12)	I liked using new technology to take this test. أعجبني أن أؤدي هذا الامتحان باستخدام التقنية الحديثة.	5	4	3	2	1
13)	I think that the test reflected my true language ability. أعتقد بأن هذا الامتحان قد عكس مستوى اللغة لدي.	5	4	3	2	1
14)	I would like to take such online tests on Moodle as official exams (e.g. mid-terms, finals, Placement Test, Exit Test). أحبذ أن أؤدي الامتحانات الرسمية (مثل امتحانات المنتصف والامتحانات النهائية وامتحانات تحديد المستوى والاجتياز) باستخدام الانترنت على المودل.	5	4	3	2	1

Please continue on the next page.

الرجاء المتابعة في الصفحة التالية.

For questions 15 - 16, select the option (a or b) that best represents your opinion and then explain your answer.

الرجاء اختيار ما يناسبك للإجابة على الأسئلة التالية وبعدها توضيح لماذا اخترت هذه الإجابة.

15) Which format of testing do you prefer?
a) pen and paper
b) online in Moodle

أي نوع من الامتحانات تفضل؟
أ) الامتحانات التقليدية بالورقة والقلم
ب) الامتحانات على الإنترنت باستخدام المودل

- Explain your answer.

وضح إجابتك.

____
____
____

16) I think I would perform best when using ___________.
a) pen and paper tests.
b) online tests on Moodle.

أعتقد بأن أدائي في الامتحانات سيكون أفضل عندما تكون هذه الامتحانات __________.
أ) تقليدية بالورقة والقلم
ب) على الإنترنت باستخدام المودل

- Explain your answer.

وضح إجابتك.

____
____
____

For questions 17 – 26, circle the option (5 = strongly agree; 4= agree; 3= neutral; 2 = disagree; 1= strongly disagree) that best applies to you.

الرجاء الإجابة على الأسئلة التالية باختيار أحد الأرقام من 1 – 5 الذي يعبر عن رأيك. 5= أوافق جدا 4= أوافق 3=محايد 2=لا أوافق 1=أوافق بشدة.

No.	**Section 2: Issues and limitations** المشاكل والتحديات التي واجهت في هذا الامتحان	**5** **Strongly agree** أوافق جدا	**4** **Agree** أوافق	**3** **Neutral** محايد	**2** **Disagree** لا أوافق	**1** **Strongly disagree** لا أوافق بشدة
17) ..	There were technical problems during the exam. واجهت مشاكل تقنية خلال أدائي لهذا الامتحان.	5	4	3	2	1
18) ..	The network was efficient and did not slow down while I was taking the test. شبكة الانترنت كانت جيدة ولم يحدث أي بطء فيها.	5	4	3	2	1
19) ..	The audio files in the listening loaded quickly. الملفات الصوتية بامتحان الاستماع فتحت بسرعة.	5	4	3	2	1
20) ..	The computer worked properly during the exam. الحاسب الآلي كان يعمل بشكل جيد خلال الامتحان.	5	4	3	2	1
21) ..	The headphones worked properly during the exam. السماعات كانت تعمل بشكل جيد خلال الامتحان.	5	4	3	2	1
22) ..	I was able to successfully log onto Moodle and the online test. الدخول إلى المودل والامتحان تم بنجاح.	5	4	3	2	1
23) ..	Pictures and graphs were clear. الصور والأشكال التوضيحية كانت واضحة.	5	4	3	2	1
24) ..	Test procedures and instructions given were clear and easy to follow. إجراءات وتعليمات الامتحان المعطاة كانت واضحة وكان من السهل اتباعها لأداء الامتحان.	5	4	3	2	1
25) ..	I have enough experience with technology to take tests on Moodle. خبرتي بتقنية المعلومات كافية لتمكنني من أداء الامتحانات على المودل.	5	4	3	2	1
26) ..	I will need extra technical training before I am ready to take online exams. سأحتاج لتدريب اضافي في تقنية المعلومات لأكون جاهزا لأداء امتحانات على الانترنت.	5	4	3	2	1

Please continue on the next page. **الرجاء المتابعة في الصفحة التالية.**

For questions 27– 28, please check (✔) Yes or No and then explain your answer.

للإجابة على الأسئلة التالية الرجاء اختيار نعم أو لا ومن ثم توضيح لماذا اخترت هذه الاجابة.

27) Did you like taking the test on Moodle? هل أعجبك أداء هذا الامتحان باستخدام المودل؟

□ Yes □ نعم

□ No □ لا

- Explain your answer. Why Yes? Or why No? وضح اجابتك. لماذا اخترت نعم/لا؟

28) Would you like to take official exams (like mid-terms, finals, placement tests, exit tests, and so forth) on Moodle to take decisions about the level of your language proficiency?

هل ترغب في أداء الامتحانات الرسمية (مثل امتحانات المنتصف والامتحانات النهائية وامتحانات تحديد المستوى والاجتياز) باستخدام الانترنت على المودل وذلك لاتخاذ قرارات رسمية بخصوص مستوى مهارات اللغة لديك.

□ Yes □ نعم

□ No □ لا

- Explain your answer. Why Yes? Or why No? وضح اجابتك. لماذا اخترت نعم/لا؟

29) What other suggestions or comments would you like to give on Moodle- hosted online English language testing?

ما تعليقاتك أو اقتراحاتك بشأن أداء امتحانات اللغة الانجليزية على الانترنت باستخدام المودل؟

End of Questionnaire – Thank You!

Questionnaire used on the 15th of April, 2015 for test takers

Name: ______________________________ ID: ___________   **Section: ______**

Questionnaire for Test Takers	**استبانة للطلبة المؤدين للامتحان**
Thank you for taking part in this study. Now that you have finished taking the online test on the learning management system Moodle, we would like you to fill in this questionnaire about your test-taking experience. We appreciate your feedback on all questions.	نشكركم على مشاركتكم في هذا المشروع البحثي. بعدما أنهيتم أداء هذا الامتحان على المودل الرجاء تعبئة هذه الاستبانة عن تجربتكم لهذا الامتحان. شاكرين لكم إجابتكم على جميع الأسئلة وعلى تعليقاتكم وآراءكم.

For questions 1 – 4, please check (✔) all that apply to you. Please select only one answer for each question. **الرجاء اختيار ما يناسبك للإجابة على الأسئلة التالية.**

1) Your current course of study and level **المستوى أو المقرر الذي تدرسه حاليا**

□ FPEL0340 (level 4)

□ FPEL0450 (level 5)

□ FPEL0560 (level 6)

□ FPEL0604 (level 6)

2) Gender:	**الجنس**
□ Male	□ ذكر
□ Female	□ أنثى

3) Your level of familiarity with tests or quizzes on Moodle:
مدى معرفتك للامتحانات المستخدمة في المودل

□ Very familiar	□ أعرفها جيدا
□ Somehow familiar	□ أعرفها بعض الشي
□ A little bit familiar	□ أعرفها قليلا
□ Not familiar at all	□ لا أعرفها مطلقا

4) Your level of computer-literacy or familiarity with computers:
مدى معرفتك أو خبرتك باستخدامات جهاز الحاسب الآلي

□ Very familiar	□ لدي خبرة واسعة
□ Somehow familiar	□ لدي خبرة بعض الشي
□ A little bit familiar	□ خبرتي ضئيلة جدا
□ Not familiar at all	□ ليس لدي أية خبرة مطلقا

Please continue on the next page. الرجاء المتابعة في الصفحة التالية.

For questions 5 – 17, circle the option (5 = strongly agree; 4= agree; 3= neutral; 2 = disagree; 1= strongly disagree) that best applies to you.

الرجاء الإجابة على الأسئلة التالية باختيار أحد الأرقام من 1 – 5 الذي يعبر عن رأيك. 5= أوافق جدا 4= أوافق 3=محايد 2=لا أوافق 1=أوافق بشدة.

No.	**Section 1: Overall test-taking experience** تجربتك لهذا الامتحان بشكل عام	**5 Strongly agree** أوافق جدا	**4 Agree** أوافق	**3 Neutral** محايد	**2 Disagree** لا أوافق	**1Strongly disagree** لا أوافق بشدة
5)....	Overall, I liked this test-taking experience. أعجبني هذا الامتحان بشكل عام.	5	4	3	2	1
6)....	Overall, the test was easy to navigate by moving from one page displaying a subtest to another. بصورة عامة كان من السهل الانتقال من صفحة إلى أخرى لأداء أجزاء الامتحان.	5	4	3	2	1
7)....	Test timing was sufficient for all test sections. كان الوقت المخصص للامتحان كافيا للإجابة على جميع الأسئلة.	5	4	3	2	1
8)....	I liked the split screen mode for the reading tests where the reading texts were on the left side of the screen and the questions were on the right side. أعجبني شكل امتحانات القراءة التي تضمنت الفقرات على الجانب الأيسر من الشاشة والأسئلة على الجانب الأيمن.	5	4	3	2	1
9)....	I think the background theme (colours) of the test was appropriate. أعتقد أن ألوان خلفية واجهة الامتحان كانت مناسبة.	5	4	3	2	1
10)..	I liked the presence of the count-down timer to help me submit my answers to the test questions within the given test time. أعجبني وجود الساعة التي تظهر الوقت المتبقي على انتهاء الامتحان وساعدني هذا على تنظيم وقتي لتسليم إجاباتي على الأسئلة قبل انتهاء مدة الامتحان.	5	4	3	2	1
11)..	Sound quality of the listening tests was good. صوت التسجيل في امتحان الاستماع كان جيدا.	5	4	3	2	1
12)..	I liked that Moodle showed me instant feedback/test results at the end of the test. أعجبني أن نتائج الامتحان وإجاباتي للأسئلة ظهرت على الشاشة بعد الانتهاء من الامتحان مباشرة.	5	4	3	2	1
13)..	Test procedures and instructions given were clear and easy to follow. إجراءات وتعليمات الامتحان المعطاة كانت واضحة وكان من السهل اتباعها لأداء الامتحان.	5	4	3	2	1
14)..	I liked typing my responses for some questions. أعجبني أن أقوم بطباعة إجاباتي لبعض الأسئلة.	5	4	3	2	1
15)..	I liked using new technology to take this test. أعجبني أن أؤدي هذا الامتحان باستخدام التقنية الحديثة.	5	4	3	2	1
16)..	I think that the test reflected my true language ability. أعتقد بأن هذا الامتحان قد عكس مستوى اللغة لدي.	5	4	3	2	1
17)..	I would like to take such online tests on Moodle as official exams (e.g. mid-terms, finals, Placement Test, Exit Test). أحبذ أن أؤدي الامتحانات الرسمية (مثل امتحانات المنتصف والامتحانات النهائية وامتحانات تحديد المستوى والاجتياز) باستخدام الانترنت على المودل.	5	4	3	2	1

Please continue on the next page. **الرجاء المتابعة في الصفحة التالية.**

For questions 18- 19, select the option (a or b) that best represents your opinion and then explain your answer.

الرجاء اختيار ما يناسبك للإجابة على الأسئلة التالية وبعدها توضيح لماذا اخترت هذه الإجابة.

18) Which format of testing do you prefer? أي نوع من الامتحانات تفضل؟

a) pen and paper أ) الامتحانات التقليدية بالورقة والقلم

b) online in Moodle ب) الامتحانات على الإنترنت باستخدام المودل

- Explain your answer. وضح إجابتك.

____

____

____

____

19) I think I would perform best when using ____________.

a) pen and paper tests.

b) online tests on Moodle.

أعتقد بأن أدائي في الامتحانات سيكون أفضل عندما تكون هذه الامتحانات ___________.

أ) تقليدية بالورقة والقلم

ب) على الإنترنت باستخدام المودل

- Explain your answer. وضح إجابتك.

____

____

____

____

Please continue on the next page. الرجاء المتابعة في الصفحة التالية.

For questions 20 – 32, circle the option (5 = strongly agree; 4= agree; 3= neutral; 2 = disagree; 1= strongly disagree) that best applies to you.

الرجاء الإجابة على الأسئلة التالية باختيار أحد الأرقام من 1 – 5 الذي يعبر عن رأيك. 5= أوافق جدا 4= أوافق 3=محايد 2=لا أوافق 1=أوافق بشدة.

No.	**Section 2: Issues and limitations** المشاكل والتحديات التي واجهتك في هذا الامتحان	**5 Strongly agree** أوافق جدا	**4 Agree** أوافق	**3 Neutral** محايد	**2 Disagree** لا أوافق	**1Strongly disagree** لا أوافق بشدة
20)...	There were technical problems during the exam. واجهت مشاكل تقنية خلال أدائي لهذا الامتحان.	5	4	3	2	1
21)...	The network was efficient and did not slow down while taking the test. شبكة الانترنت كانت جيدة ولم يحدث أي بطء فيها.	5	4	3	2	1
22)...	The audio file in the listening loaded quickly. الملف الصوتي بامتحان الاستماع فتح بسرعة.	5	4	3	2	1
23)...	The computer worked properly during the exam. الحاسب الآلي كان يعمل بشكل جيد خلال الامتحان.	5	4	3	2	1
24)...	The headphones worked properly during the exam. السماعات كانت تعمل بشكل جيد خلال الامتحان.	5	4	3	2	1
25)...	I was able to successfully log onto Moodle and the online test. الدخول إلى المودل والامتحان تم بنجاح.	5	4	3	2	1
26)...	Pictures and graphs were clear. الصور والأشكال التوضيحية كانت واضحة.	5	4	3	2	1
27)...	The font size was NOT appropriate. لم يكن حجم الخط مناسبا.	5	4	3	2	1
28)...	The test took a very long time and consisted of too many sections. أخذ الامتحان وقتا طويلا جدا وتضمن أجزاء عديدة.	5	4	3	2	1
29)...	Staring for a long period at the computer screen caused me eye fatigue that affected my concentration. كان النظر المتواصل لشاشة الحاسب الالي مرهقا لعيني ومشتتا لتركيزي.	5	4	3	2	1
30)...	I needed to take notes during the test. احتجت لكتابة ملاحظات خلال أدائي للإمتحان.	5	4	3	2	1
31)...	I have enough experience with technology to take tests on Moodle. خبرتي بتقنية المعلومات كافية لتمكنني من أداء الامتحانات على المودل.	5	4	3	2	1
32)...	I will need extra technical training before I am ready to take online exams. سأحتاج لتدريب اضافي في تقنية المعلومات لأكون جاهزا لأداء امتحانات على الانترنت.	5	4	3	2	1

Please continue on the next page. الرجاء المتابعة في الصفحة التالية.

For questions 33– 35, please check (✔) Yes or No and then explain your answer.

للإجابة على الأسئلة التالية الرجاء اختيار نعم أو لا ومن ثم توضيح لماذا اخترت هذه الاجابة.

33) Did you like taking the test on Moodle?

هل أعجبك أداء هذا الامتحان باستخدام المودل؟

□ Yes

□ No

□ نعم

□ لا

- Explain your answer. Why Yes? Or why No?

وضح اجابتك. لماذا اخترت نعم/لا؟

____

____

____

34) Would you like to take official exams (like mid-terms, finals, placement tests, exit tests, and so forth) on Moodle to take decisions about the level of your language proficiency?

هل ترغب في أداء الامتحانات الرسمية (مثل امتحانات المنتصف والامتحانات النهائية وامتحانات تحديد المستوى والاجتياز) باستخدام الانترنت على المودل وذلك لاتخاذ قرارات رسمية بخصوص مستوى مهارات اللغة لديك.

□ Yes

□ No

□ نعم

□ لا

- Explain your answer. Why Yes? Or why No?

وضح اجابتك. لماذا اخترت نعم/لا؟

____

____

____

35) What other suggestions or comments would you like to give on Moodle- hosted online English language testing?

ما تعليقاتك أو اقتراحاتك بشأن أداء امتحانات اللغة الانجليزية على الانترنت باستخدام المودل؟

____

____

____

End of Questionnaire – Thank You!

Appendix G: Usability study test takers' interview

1) How would you describe your experience of taking the Moodle-hosted test, positive or negative? and why?

كيف تصف تجربتك لهذا الامتحان باستخدام الانترنت وخاصة المودل؟ هل كانت تجربة ايجابية أم سلبية؟ مع بيان السبب.

2) What do you think about the use of Moodle to run official exams? Would you like Moodle to be used to run official exams (like mid-terms, finals, placement tests, exit tests, and so forth) to take decisions about the level of your language proficiency? Why or why not?

ما رأيك في استخدام المودل لأداء الامتحانات؟ هل تؤيد هذا الاستخدام للمودل لأداء الطلبة للامتحانات الرسمية (مثل امتحان المنتصف والنهائي وتحديد المستوى والاجتياز) وذلك لاتخاذ قرارات حول مستوى اللغة لديك؟

3) Compare the Moodle-hosted test with paper-based tests. Which test method would you prefer (paper-based or Moodle-based tests)? Why?

عند مقارنتك لهذا الامتحان باستخدام المودل للامتحانات التقليدية باستخدام الورقة والقلم أيهما تفضل ولماذا؟

4) From your experience of taking the Moodle-hosted test, do you think it is practical to take tests on Moodle? Why or why not?

من خلال تجربتك لأداء هذا الامتحان هل تعتقد أن أداء الامتحان باستخدام المودل عملي؟ وضح رأيك.

5) Do you think that the computer labs are well-equipped and efficient for taking tests on Moodle?

هل تعتقد أن مختبرات الحاسب الآلي مجهزة جيدا لتكون ذا فاعلية لأداء الامتحانات باستخدام المودل؟

6) What technical issues did you face? Were there any problems with the network and loading of Moodle pages or login?

ما المشاكل التقنية التي واجهتها عند أداءك لهذا الامتحان؟ هل صادفتك مشاكل بشبكة الانترنت أو دخول المودل أو تصفحه؟

7) What do you think of the feedback you received from Moodle on your test performance? Do you like that your responses are scored by machine? Why or why not?

ما رأيك بالتغذية الراجعة أو نتيجة أدائك في الامتحان التي ظهرت بعد الامتحان مباشرة؟ هل يعجبك أن اجاباتك تم تصحيحها مباشرة بالحاسب الآلي أو نظام المودل؟ وضح اجابتك.

8) What suggestions do you have to improve test-taking on Moodle?

ماذا تقترح لتطوير الامتحانات على المودل؟

Appendix H: Main study examinees' questionnaire

Name: ________________________________ ID: ___________	**Section: ______**

Questionnaire for Test Takers	**استبانة للطلبة المؤدين للامتحان**
Thank you for taking part in this study. Now that you have finished taking the online test on the learning management system Moodle, we would like you to fill in this questionnaire about your test-taking experience. We appreciate your feedback on all questions.	نشكركم على مشاركتكم في هذا المشروع البحثي. بعدما أنهيتم أداء هذا الامتحان على المودل الرجاء تعبئة هذه الاستبانة عن تجربتكم لهذا الامتحان. شاكرين لكم إجابتكم على جميع الأسئلة وعلى تعليقاتكم وآراءكم.
For questions 1 – 4, please check (✔) all that apply to you. Please select only one answer for each question. **الرجاء اختيار ما يناسبك للإجابة على الأسئلة التالية.**	
1) Your current course of study and level **المستوى أو المقرر الذي تدرسه حاليا**	
□ FPEL0340 (level 4)	
□ FPEL0450 (level 5)	
□ FPEL0560 (level 6)	
□ FPEL0604 (level 6)	
2) Gender:	**الجنس**
□ Male	□ ذكر
□ Female	□ أنثى
3) Your level of familiarity with tests or quizzes on Moodle: **مدى معرفتك للامتحانات المستخدمة في المودل**	
□ Very familiar	□ أعرفها جيدا
□ Somehow familiar	□ أعرفها بعض الشي
□ A little bit familiar	□ أعرفها قليلا
□ Not familiar at all	□ لا أعرفها مطلقا
Please continue on the next page. الرجاء المتابعة في الصفحة التالية.	
4) Your level of computer-literacy or familiarity with computers: **مدى معرفتك أو خبرتك باستخدامات جهاز الحاسب الآلي**	
□ Very familiar	□ لدي خبرة واسعة
□ Somehow familiar	□ لدي خبرة بعض الشي
□ A little bit familiar	□ خبرتي ضئيلة جدا
□ Not familiar at all	□ ليس لدي أية خبرة مطلقا
Please continue on the next page. الرجاء المتابعة في الصفحة التالية.	

For questions 5 – 17, circle the option (5 = strongly agree; 4= agree; 3= neutral; 2 = disagree; 1= strongly disagree) that best applies to you.

الرجاء الإجابة على الأسئلة التالية باختيار أحد الأرقام من 1 – 5 الذي يعبر عن رأيك. 5= أوافق جدا 4= أوافق 3=محايد 2=لا أوافق 1=أوافق بشدة.

No.	**Section 1: Overall test-taking experience** تجربتك لهذا الامتحان بشكل عام	**5 Strongly agree** أوافق جدا	**4 Agree** أوافق	**3 Neutral** محايد	**2 Disagree** لا أوافق	**1Strongly disagree** لا أوافق بشدة
5)….	Overall, I liked this test-taking experience. أعجبني هذا الامتحان بشكل عام.	5	4	3	2	1
6)….	Overall, the test was easy to navigate by moving from one page displaying a subtest to another. بصورة عامة كان من السهل الانتقال من صفحة إلى أخرى لأداء أجزاء الامتحان.	5	4	3	2	1
7)….	Test timing was sufficient for all test sections. كان الوقت المخصص للامتحان كافيا للإجابة على جميع الأسئلة.	5	4	3	2	1
8)….	I liked the split screen mode for the reading tests where the reading texts were on the left side of the screen and the questions were on the right side. أعجبني شكل امتحانات القراءة التي تضمنت الفقرات على الجانب الأيسر من الشاشة والأسئلة على الجانب الأيمن.	5	4	3	2	1
9)….	I think the background theme (colours) of the test was appropriate. أعتقد أن ألوان خلفية واجهة الامتحان كانت مناسبة.	5	4	3	2	1
10)…	I liked the presence of the count-down timer to help me submit my answers to the test questions within the given test time. أعجبني وجود الساعة التي تظهر الوقت المتبقي على انتهاء الامتحان وساعدني هذا على تنظيم وقتي لتسليم إجاباتي على الأسئلة قبل انتهاء مدة الامتحان.	5	4	3	2	1
11)…	Sound quality of the listening tests was good. صوت التسجيل في امتحان الاستماع كان جيدا.	5	4	3	2	1
12)…	I liked that Moodle showed me instant feedback/test results at the end of the test. أعجبني أن نتائج الامتحان وإجاباتي للأسئلة ظهرت على الشاشة بعد الانتهاء من الامتحان مباشرة.	5	4	3	2	1
13)…	Test procedures and instructions given were clear and easy to follow. إجراءات وتعليمات الامتحان المعطاة كانت واضحة وكان من السهل اتباعها لأداء الامتحان.	5	4	3	2	1
14)…	I liked typing my responses for some questions. أعجبني أن أقوم بطباعة إجاباتي لبعض الأسئلة.	5	4	3	2	1
15)…	I liked using new technology to take this test. أعجبني أن أؤدي هذا الامتحان باستخدام التقنية الحديثة.	5	4	3	2	1
16)…	I think that the test reflected my true language ability. أعتقد بأن هذا الامتحان قد عكس مستوى اللغة لدي.	5	4	3	2	1
17)…	I would like to take such online tests on Moodle as official exams (e.g. mid-terms, finals, Placement Test, Exit Test). أحبذ أن أؤدي الامتحانات الرسمية (مثل امتحانات المنتصف والامتحانات النهائية وامتحانات تحديد المستوى والاجتياز) باستخدام الانترنت على المودل.	5	4	3	2	1

Please continue on the next page. **الرجاء المتابعة في الصفحة التالية.**

For questions 18- 19, select the option (a or b) that best represents your opinion and then explain your answer.

الرجاء اختيار ما يناسبك للإجابة على الأسئلة التالية وبعدها توضيح لماذا اخترت هذه الإجابة.

18) Which format of testing do you prefer?
a) pen and paper
b) online in Moodle

أي نوع من الامتحانات تفضل؟
أ) الامتحانات التقليدية بالورقة والقلم
ب) الامتحانات على الإنترنت باستخدام المودل

- Explain your answer. وضح إجابتك

____

____

____

19) I think I would perform best when using ____________.
a) pen and paper tests.
b) online tests on Moodle.

أعتقد بأن أدائي في الامتحانات سيكون أفضل عندما تكون هذه الامتحانات __________.
أ) تقليدية بالورقة والقلم
ب) على الإنترنت باستخدام المودل

- Explain your answer. وضح إجابتك.

____

____

____

Please continue on the next page. الرجاء المتابعة في الصفحة التالية

For questions 20 – 33, circle the option (5 = strongly agree; 4= agree; 3= neutral; 2 = disagree; 1= strongly disagree) that best applies to you.

الرجاء الإجابة على الأسئلة التالية باختيار أحد الأرقام من 1 – 5 الذي يعبر عن رأيك. 5= أوافق جدا 4= أوافق 3=محايد 2=لا أوافق 1=أوافق بشدة.

No.	**Section 2: Issues and limitations** المشاكل والتحديات التي واجهتك في هذا الامتحان	**5 Strongly agree** أوافق جدا	**4 Agree** أوافق	**3 Neutral** محايد	**2 Disagree** لا أوافق	**1Strongly disagree** لا أوافق بشدة
20)…	There were technical problems during the exam. واجهت مشاكل تقنية خلال أدائي لهذا الامتحان.	5	4	3	2	1
21)…	The network was efficient and did not slow down while taking the test. شبكة الانترنت كانت جيدة ولم يحدث أي بطء فيها.	5	4	3	2	1
22)…	The audio file in the listening loaded quickly. الملف الصوتي بامتحان الاستماع فتح بسرعة.	5	4	3	2	1
23)…	The computer worked properly during the exam. الحاسب الآلي كان يعمل بشكل جيد خلال الامتحان.	5	4	3	2	1
24)…	The headphones worked properly during the exam. السماعات كانت تعمل بشكل جيد خلال الامتحان.	5	4	3	2	1
25)…	I was able to successfully log onto Moodle and the online test. الدخول إلى المودل والامتحان تم بنجاح.	5	4	3	2	1
26)…	Pictures and graphs were clear. الصور والأشكال التوضيحية كانت واضحة.	5	4	3	2	1
27)…	The font size was NOT appropriate. لم يكن حجم الخط مناسبا.	5	4	3	2	1
28)…	The test was too long as it consisted of too many sections. كان الامتحان طويلا جدا حيث أنه تضمن أجزاء عديدة.	5	4	3	2	1
29)…	Staring at the computer screen for a long period of time made me lose my concentration. كان النظر المتواصل لشاشة الحاسب الالي مشتتا لتركيزي.	5	4	3	2	1
30)…	Staring at the computer screen for a long period of time caused me eye fatigue. كان النظر المتواصل لشاشة الحاسب الالي مرهقا لعيني.	5	4	3	2	1
31)…	I needed to take notes during the test. احتجت لكتابة ملاحظات خلال أدائي للإمتحان.	5	4	3	2	1
32)…	I have enough experience with technology to take tests on Moodle. خبرتي بتقنية المعلومات كافية لتمكنني من أداء الامتحانات على المودل.	5	4	3	2	1
33)…	I will need extra technical training before I am ready to take online exams. سأحتاج لتدريب اضافي في تقنية المعلومات لأكون جاهزا لأداء امتحانات على الانترنت.	5	4	3	2	1

Please continue on the next page. الرجاء المتابعة في الصفحة التالية.

For questions 34– 36, please check (✔) Yes or No and then explain your answer.
للإجابة على الأسئلة التالية الرجاء اختيار نعم أو لا ومن ثم توضيح لماذا اخترت هذه الاجابة.

34) Did you like taking the test on Moodle?
هل أعجبك أداء هذا الامتحان باستخدام المودل؟

□ Yes　　□ نعم
□ No　　□ لا

- Explain your answer. Why Yes? Or why No?
وضح اجابتك. لماذا اخترت نعم/لا؟

____
____
____

35) Would you like to take official exams (like mid-terms, finals, placement tests, exit tests, and so forth) on Moodle to take decisions about the level of your language proficiency?
هل ترغب في أداء الامتحانات الرسمية (مثل امتحانات المنتصف والامتحانات النهائية وامتحانات تحديد المستوى والاجتياز) باستخدام الانترنت على المودل وذلك لاتخاذ قرارات رسمية بخصوص مستوى مهارات اللغة لديك.

□ Yes　　□ نعم
□ No　　□ لا

- Explain your answer. Why Yes? Or why No?
وضح اجابتك. لماذا اخترت نعم/لا؟

____
____
____

36) What other suggestions or comments would you like to give on Moodle- hosted online English language testing?

ما تعليقاتك أو اقتراحاتك بشأن أداء امتحانات اللغة الانجليزية على الانترنت باستخدام المودل؟

____
____
____

End of Questionnaire – Thank You!

Appendix I: Main study invigilators' questionnaire

Thank you for taking part in this study. Now that you have finished invigilating the test on Moodle, we would like you to fill in this questionnaire about your test invigilation experience. We truly appreciate and value your feedback.

Background information:

- ***For questions 1 – 2, please check (✔) all that apply to you. Please select only one answer for each question.***

1) Gender:

□ Male
□ Female

2) The current course of study/level/section that you teach:

□ FPEL0340 (level 4)
□ FPEL0450 (level 5)
□ FPEL0560 (level 6)
□ FPEL0604 (level 6)
□ Section: ________________

- ***For questions 3 - 10, please write your answer in the space provided. You can use extra papers if needed.***

3) What do you think of the Moodle-hosted test overall?

____
____
____
____

4) Did you like/dislike your test invigilation experience? Why?

____
____
____
____

5) From your test experience, do you think it is practical to run the tests on Moodle? Why or why not?

____
____
____
____

Please continue on the next page

6) Do you think that the computer labs are well-equipped and efficient for taking tests on Moodle?

____

____

____

____

7) What technical issues did you face during exam invigilation? For example, were there any problems with the network and loading of Moodle pages or login? Do you think students' test performance was affected by any technical issues (e.g. specific features of the testing interface)?

____

____

____

____

8) What do you think about the use of Moodle to run official exams? Would you like Moodle to be used to run official exams like mid-terms, finals, placement tests, exit tests, and so forth) to take decisions about students' levels of language proficiency? Why or why not?

____

____

____

____

9) As in the case of this Moodle-hosted test, Moodle can be used to run objective exams and automatically mark students' responses (even short-answer), and thus relieving invigilators from the post-invigilation marking duties they usually have. What do you think of the potential of relieving you as a teacher in the LC from marking duties of the objective tests? Would you support this testing practice?

____

____

____

____

10) What suggestions do you have to improve test-taking on Moodle and invigilation or supervision of such exams?

____

____

____

____

End of Questionnaire – Thank you!

Appendix J: Main study examinees' semi-structured interview

1) How would you describe your experience of taking the Moodle-hosted test, positive or negative? and why?

كيف تصف تجربتك لهذا الامتحان باستخدام الانترنت وخاصة المودل؟ هل كانت تجربة ايجابية أم سلبية؟ مع بيان السبب.

2) What do you think about the use of Moodle to run official exams? Would you like Moodle to be used to run official exams (like mid-terms, finals, placement tests, exit tests, and so forth) to take decisions about the level of your language proficiency? Why or why not?

ما رأيك في استخدام المودل لأداء الامتحانات؟ هل تؤيد هذا الاستخدام للمودل لأداء الطلبة للامتحانات الرسمية (مثل امتحان المنتصف والنهائي وتحديد المستوى والاجتياز) وذلك لاتخاذ قرارات حول مستوى اللغة لديك؟

3) Compare the Moodle-hosted test with paper-based tests. Which test method would you prefer (paper-based or Moodle-based tests)? Why?

عند مقارنتك لهذا الامتحان باستخدام المودل للامتحانات التقليدية باستخدام الورقة والقلم أيهما تفضل ولماذا؟

4) From your experience of taking the Moodle-hosted test, do you think it is practical to take tests on Moodle? Why or why not?

من خلال تجربتك لأداء هذا الامتحان هل تعتقد أن أداء الامتحان باستخدام المودل عملي؟ وضح رأيك.

5) Do you think that the computer labs are well-equipped and efficient for taking tests on Moodle?

هل تعتقد أن مختبرات الحاسب الآلي مجهزة جيدا لتكون ذا فاعلية لأداء الامتحانات باستخدام المودل؟

6) What technical issues did you face? Were there any problems with the network and loading of Moodle pages or login? Do you think your test performance was affected by any technical issues (e.g. specific features of the testing interface)?

ما المشاكل التقنية التي واجهتها عند أداءك لهذا الامتحان؟ هل صادفتك مشاكل بشبكة الانترنت أو دخول المودل أو تصفحه؟ هل تعتقد أن أداءك في هذا الامتحان تأثر بأي مشاكل تقنية (مثلا تلك التي تتعلق بخصائص معينة لواجهة الامتحان)؟

7) What do you think of the feedback you received from Moodle on your test performance? Do you like that your responses are scored by machine? Why or why not?

ما رأيك بالتغذية الراجعة أو نتيجة أدائك في الامتحان التي ظهرت بعد الامتحان مباشرة؟ هل يعجبك أن اجاباتك تم تصحيحها مباشرة بالحاسب الآلي أو نظام المودل؟ وضح اجابتك.

8) What suggestions do you have to improve test-taking on Moodle?

ماذا تقترح لتطوير الامتحانات على المودل؟

Appendix K: Main study invigilators’ semi-structured interview

The following is a set of questions for semi-structured interviews that were conducted with the researcher being the moderator (asking questions) and invigilators being the interviewees. These interviews were conducted individually depending on participants' arrangement with the researcher. Questions were rephrased to explain them to participants and were followed with other probing and follow-up questions.

1) What do you think of the Moodle-hosted test overall?

2) What did you like about your test invigilation experience? Why?

3) What did you dislike about your test invigilation experience? Why?

4) What technical issues did you face during exam invigilation? Were there any problems with the network and loading of Moodle pages or login?

5) Do you think that the computer labs are well-equipped and efficient for taking tests on Moodle?

6) From your test experience, do you think it is practical to run the tests on Moodle? Why or why not?

7) What do you think about the use of Moodle to run official exams? Would you like Moodle to be used to run official exams like mid-terms, finals, placement tests, exit tests, and so forth) to take decisions about students’ levels of language proficiency? Why or why not?

8) As in the case of this Moodle-hosted test, Moodle can be used to run objective exams and automatically mark students’ responses (even short-answer), and thus relieving invigilators from the post-invigilation marking duties (marking and double-marking) they usually have. What do you think of the potential of relieving you as a teacher in the LC from marking duties of the objective tests? Would you support this testing practice? Why or why not?

9) What suggestions do you have to improve test-taking on Moodle and invigilation or supervision of such exams?

Appendix L: Invigilation instructions

- Collect students’ phones and place them on teacher’s desk. Ensure that books or any other materials are not within students’ reach during the test.
- Ensure students sit at computer stations that have headphones set up.
- Hand in papers to students to take notes on (if needed) during the test.
- Help students follow researcher’s instructions to log into the online test.
- Be vigilant throughout the testing session.
- Assist students who experience issues during the test.
- Ensure students adhere to the given test time and submit their responses at the end by clicking “SUBMIT ALL AND FINISH”.
- Collect all papers handed in for note-taking.
- Students can collect their phones at the end of the testing session.
- Finally, report on the test invigilation experience using the invigilator’s questionnaire and (if possible) in a follow-up audio-recorded interview with the researcher

Study procedure:

- 5 minutes: Headphones set-up and Log-in process
- 60 minutes: Students sit reading and language use test.
- 30 minutes: Students sit listening test.
- 10 minutes: Invigilator and ALL students fill in relevant questionnaires.
- 5 minutes: Information sheets are passed to students to keep for their records. Students receive consent form to sign and indicate their willingness (or not) to participate in follow-up interview.

After the testing event in researcher’s office:

- About 30 minutes: Volunteering students take part in audio-recorded individual or group interview/ discussion with the researcher.
- About 30 minutes: Volunteering invigilator participates in audio-recorded interview with the researcher.

Appendix M: Pilot study information

Table M1 shows the pilot study participating student and teacher sample. In the first exam trial, 23 volunteering students in a Master program at UQ participated. The sample (*n* = 23) comprised of 4 males (17.4%) and 19 females (82.6%). The role of these pilot study participants was to trial the Moodle-hosted test prototype and to provide the researcher with feedback via questionnaires (Appendix C, pp. 162-163).

Table M1. *Pilot Study Participating Sample*

Event	Participant	Course	Male	Female	Total
First Exam Trial	UQ Master students	Language Testing Course	4	19	23
Judgmental Validation Session	Language teachers	Teaching FPEL Courses	1	3	4
Usability Testing Day 1	SQU Level 6 students	FPEL604 (SCI)		2	2
Usability Testing Day 2	SQU Level 4 students	FPEL0340 (GEN)	5		7
Usability Testing Day 3	SQU Level 6 students	FPEL604 (SCI)		2	
		FPEL604 (AGR)	4	12	16
Total			14	38	52

Notes. FPEL = Foundation Program of English Language; GEN = SCI = Sciences; General English; AGR = Agriculture.

The pilot study also involved a sample of volunteering students (*n* = 25) taking foundation English language courses at Levels 4 and 6 at the LC in SQU, Oman, who took part in the usability testing sessions over three days in April of 2015. After sitting the test, female participants formed the majority of the sample 64% (*n* = 16) in the usability testing sessions. The majority of the usability study participants were at Level 6, the highest level equivalent to IELTS Band 5, and were enrolled in FPEL0604 Agriculture program (64%; *n* = 16) and FPEL0604 Sciences program (16%; *n* = 4). Twenty percent (*n* = 5) were enrolled in Level 4 (pre-intermediate English language proficiency) of the General FPEL0340 program. The 16 students from Agriculture were tested with their teacher acting as the invigilator and assisted by the researcher, while the remainder were invigilated by the researcher alone. The SQU students provided feedback on the Moodle-hosted test user interface via questionnaires (Appendix F, pp. 167-176) and focus group semi-structured interviews (Appendix G, p. 177). All students returned questionnaires with ten opting to be interviewed (five from Level 4, four from Level 6 Sciences, and one from Level 6 Agriculture).

Language teachers from the SQU English Language Foundation Program were invited to participate via email. Those that volunteered were involved in the pilot study. As can be seen in Table M1, four language teachers participated in the judgmental validation session. There were three females and one male in this sample. These judges provided valuable feedback to the researcher via a questionnaire (Appendix D, pp. 164-165) and a focus group semi-structured interview (Appendix E, p. 166) after trialling the Moodle-hosted test.

The two sets of data (quantitative and qualitative) generated at the pilot study were analysed as appropriate. Responses to selected-response Likert-type scale questionnaire items from participants of the first exam trial at UQ and from the test takers in the usability testing sessions were analysed statistically using SPSS software v.23 descriptive statistics, frequencies, and bar charts. Thematic induction was used to analyse the following textual data:

1) responses to the open-ended questionnaire items from participants in the first exam trial, judgemental validation session, and usability testing sessions;
2) focus group semi-structured interview data from the judgemental validation session and usability testing sessions; and
3) the researcher's field notes and observations on reflective journals.

The textual data were analysed thematically to identify potential issues that can affect test performance in the testing environment. Table M2 gives a summary of some of the themes identified in the pilot study. The themes that came up in the data were considered to be technology-related construct-irrelevant issues potentially affecting test performance. The pilot study informed the main study by identifying these issues, and the researcher took action to tackle such issues using the problem resolution approach. Recording observations and field notes of such issues on the reflective journals was essential in taking these actions.

Table M2. *Technology-Related Issues Addressed In the Pilot Study*

Issues	Actions taken
• Reading text on top of the page and the questions following the reading text made test takers inconveniently scroll up and down too much.	• Created a split screen mode for the reading test in which the reading text is put on the left side of the screen and the questions are placed on the right side. • Added questionnaire and interview items that asked about this feature.
• Background theme (colours) of the testing interface was inappropriate for test takers.	• Changed the background theme and added questions in the data collection instruments that asked about it.
• The listening test needed to be separated from the other test sections because putting the listening on the last page of the entire test allowed test takers to access the listening materials more than once. This is not a fair standard practice as it can dis(advantage) test takers.	• The listening test was separated from the entire test and given its own time limit using the count-down timer function on Moodle. • The MP3 player for listening tests was embedded in the Moodle-hosted listening test to limit the number of times test takers play the listening audio file to once only and to disable the stop and pause functions for a fairer standard practice. • Questions were added to the data collection instruments that asked study participants about this feature.

Appendix N: Descriptive statistics of Moodle-hosted test

Table N1 gives descriptive statistics for the Moodle-hosted test. Green (2013) describes how to run such analyses on SPSS (p. 35) and how to interpret inferential statistics (p. 45). As can be seen in Table N1, the smallest value of the mode is 19.00 and the median is 21.00 with a minimum mark of 6 and a maximum mark of 40. When we divide the value of skewness by the standard error of skewness = .195 / .196 = 1.154.This value of 1.154 is not higher than +2, so this is a symmetrical positive skew indicating that there are more test takers at the higher end of the distribution. The negative sign in the kurtosis value -.687 indicates that the test takers are more spread out and the distribution of test scores is platykurtic distribution telling us that there is more variability in the test scores.

Table N1. *Descriptive Statistics on the Moodle-hosted Test Total*

N Valid	207
N Missing	0
Mean	20.87
Std. Error of Mean	.516
Median	21.00
Mode	19[a]
Std. Deviation	7.419
Variance	55.046
Skewness	.195
Std. Error of Skewness	.169
Kurtosis	-.687
Std. Error of Kurtosis	.337
Range	34
Minimum	6
Maximum	40

Notes. Multiple modes exist. The smallest value is shown.

To get a pictorial representation of these descriptive statistics on the test total, Figure N1 presents a histogram that has been created based on the data set.

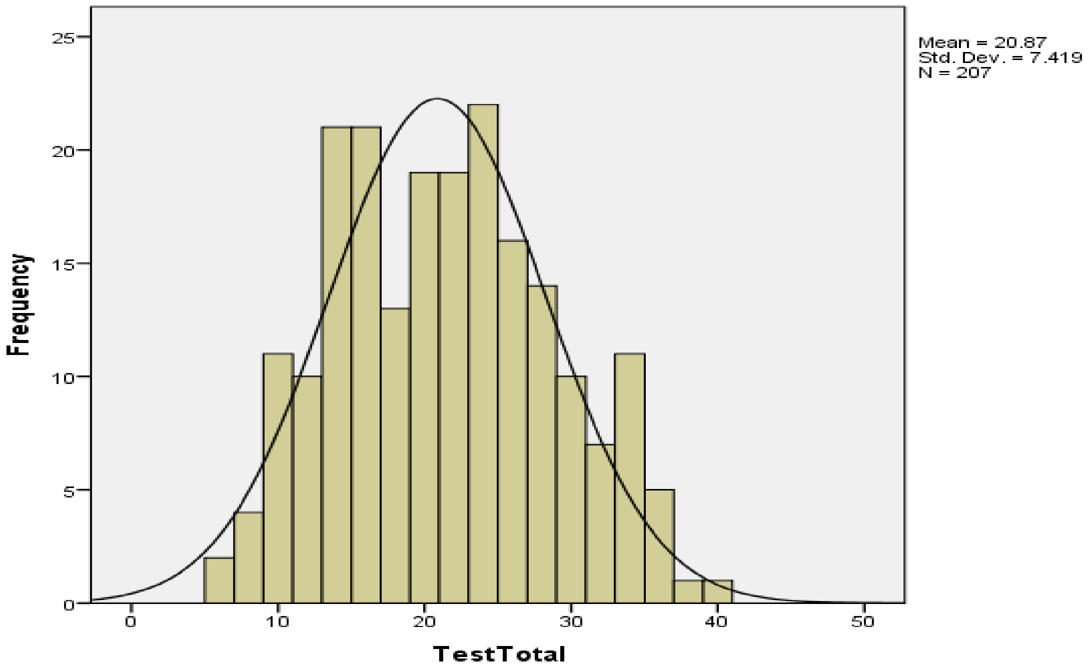


Figure N1. Histogram of the Moodle-hosted Test total.

Appendix O: Person statistics

Table O1. *Person Measures*

EN	TS	TC	ME	MSE*	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PERSON
124	40	60	.93	.33	1.19	1.1	1.19	.8	Exam124L6CAM
142	37	60	.62	.32	1.17	1.1	1.11	.5	Exam142L6EEAL
146	36	60	.52	.31	1.02	.2	.99	.0	Exam146L6EEAL
181	36	60	.52	.31	1.00	.1	.92	-.3	Exam181L6MED0560
198	36	60	.52	.31	.85	-1.0	.83	-.7	Exam198L6SCI
153	35	60	.42	.31	1.13	.9	1.02	.1	Exam153L6ENG
154	35	60	.42	.31	1.17	1.2	1.23**	1.0	Exam154L6ENG
160	34	60	.32	.31	.99	.0	.89	-.4	Exam160L6ENG
167	34	60	.32	.31	1.09	.7	1.15	.7	Exam167L6ENG
187	34	60	.32	.31	.95	-.3	.87	-.6	Exam187L6MED0560
113	33	60	.23	.31	.97	-.2	1.04	.2	Exam113L6CAM
127	33	60	.23	.31	.95	-.3	.91	-.3	Exam127L6CAM
144	33	60	.23	.31	1.01	.1	.98	.0	Exam144L6EEAL
170	33	60	.23	.31	1.02	.2	1.03	.2	Exam170L6ENG
184	33	60	.23	.31	1.08	.6	1.36**	1.6	Exam184L6MED0560
191	33	60	.23	.31	1.34	2.3	1.66 **	2.6	Exam191L6MED0560
199	33	60	.23	.31	1.14	1.0	1.00	.1	Exam199L6SCI
201	33	60	.23	.31	1.08	.6	.97	-.1	Exam201L6SCI
186	32	60	.14	.30	1.05	.4	.94	-.2	Exam186L6MED0560
204	32	60	.14	.30	1.02	.2	.92	-.3	Exam204L6SCI
123	31	60	.05	.30	1.01	.1	.90	-.4	Exam123L6CAM
131	31	60	.05	.30	1.18	1.4	1.12	.6	Exam131L6CEPS
148	31	60	.05	.30	1.19	1.4	1.13	.6	Exam148L6EEAL
161	31	60	.05	.30	1.05	.5	1.01	.1	Exam161L6ENG
166	31	60	.05	.30	1.02	.2	.93	-.2	Exam166L6ENG
162	30	60	-.05	.30	1.16	1.3	1.12	.6	Exam162L6ENG
165	30	60	-.05	.30	1.07	.6	.97	-.1	Exam165L6ENG
173	30	60	-.05	.30	1.00	.1	1.17	.8	Exam173L6MED0560
116	29	60	-.14	.30	1.00	.0	.91	-.3	Exam116L6CAM
117	29	60	-.14	.30	1.35	2.6	1.51**	2.0	Exam117L6CAM
137	29	60	-.14	.30	1.10	.8	1.04	.3	Exam137L6CEPS
139	29	60	-.14	.30	1.17	1.3	1.16	.7	Exam139L6CEPS
158	29	60	-.14	.30	1.15	1.2	1.06	.4	Exam158L6ENG
164	29	60	-.14	.30	1.22	1.7	1.39**	1.6	Exam164L6ENG
194	29	60	-.14	.30	1.20	1.6	1.15	.7	Exam194L6MED0560
61	28	60	-.23	.30	.95	-.4	.86	-.5	Exam061L5CAMS
141	28	60	-.23	.30	.93	-.6	.84	-.6	Exam141L6EEAL
156	28	60	-.23	.30	1.12	1.0	1.06	.3	Exam156L6ENG
183	28	60	-.23	.30	.95	-.4	1.05	.3	Exam183L6MED0560
185	28	60	-.23	.30	.88	-1.0	.78***	-.9	Exam185L6MED0560
95	27	60	-.32	.30	1.05	.4	.95	-.1	Exam095L5SCI
112	27	60	-.32	.30	.76	-2.1	.67***	-1.5	Exam112L6CAM
136	27	60	-.32	.30	.78	-2.0	.66***	-1.5	Exam136L6CEPS
138	27	60	-.32	.30	.87	-1.1	.78***	-.9	Exam138L6CEPS
147	27	60	-.32	.30	1.16	1.3	1.46**	1.7	Exam147L6EEAL
152	27	60	-.32	.30	.88	-1.0	.93	-.2	Exam152L6ENG
188	27	60	-.32	.30	.81	-1.6	.71	-1.2	Exam188L6MED0560
200	27	60	-.32	.30	1.00	.0	.91	-.3	Exam200L6SCI
202	27	60	-.32	.30	1.25	2.0	1.34**	1.3	Exam202L6SCI
52	26	60	-.41	.30	.94	-.5	.84	-.6	Exam052L4GEN
94	26	60	-.41	.30	.90	-.8	1.07	.3	Exam094L5SCI
114	26	60	-.41	.30	1.11	.9	1.02	.2	Exam114L6CAM
121	26	60	-.41	.30	1.24	1.9	1.82	2.7	Exam121L6CAM
122	26	60	-.41	.30	.84	-1.4	.73***	-1.1	Exam122L6CAM

134	**26**	**60**	**-.41**	**.30**	1.21	1.7	1.13	.6	Exam134L6CEPS
163	**26**	**60**	**-.41**	**.30**	1.05	.5	.97	.0	Exam163L6ENG
196	**26**	**60**	**-.41**	**.30**	1.17	1.4	1.14	.6	Exam196L6SCI
205	**26**	**60**	**-.41**	**.30**	.98	-.1	.91	-.3	Exam205L6SCI
78	**25**	**60**	**-.50**	**.30**	.94	-.5	.82	-.6	Exam078L5CEPS
93	**25**	**60**	**-.50**	**.30**	.93	-.6	.84	-.6	Exam093L5SCI
111	**25**	**60**	**-.50**	**.30**	.88	-1.0	.74***	-1.0	Exam111L5SCI
119	**25**	**60**	**-.50**	**.30**	.86	-1.2	.81	-.7	Exam119L6CAM
125	**25**	**60**	**-.50**	**.30**	1.02	.2	.99	.1	Exam125L6CAM
135	**25**	**60**	**-.50**	**.30**	1.12	1.0	.98	.0	Exam135L6CEPS
175	**25**	**60**	**-.50**	**.30**	.93	-.6	.84	-.5	Exam175L6MED0560
57	**24**	**60**	**-.59**	**.30**	.94	-.4	.79***	-.7	Exam057L5CAMS
83	**24**	**60**	**-.59**	**.30**	.87	-1.2	.72***	-1.0	Exam083L5CEPS
120	**24**	**60**	**-.59**	**.30**	.88	-1.1	.84	-.5	Exam120L6CAM
128	**24**	**60**	**-.59**	**.30**	1.04	.4	.93	-.2	Exam128L6CAM
145	**24**	**60**	**-.59**	**.30**	.91	-.8	.81	-.6	Exam145L6EEAL
155	**24**	**60**	**-.59**	**.30**	1.13	1.1	1.73**	2.3	Exam155L6ENG
174	**24**	**60**	**-.59**	**.30**	1.00	.1	.87	-.4	Exam174L6MED0560
179	**24**	**60**	**-.59**	**.30**	1.19	1.6	1.50**	1.7	Exam179L6MED0560
192	**24**	**60**	**-.59**	**.30**	1.10	.9	1.24	.9	Exam192L6MED0560
197	**24**	**60**	**-.59**	**.30**	1.26	2.1	1.54**	1.8	Exam197L6SCI
60	**23**	**60**	**-.68**	**.30**	1.00	.1	.90	-.2	Exam060L5CAMS
62	**23**	**60**	**-.68**	**.30**	.94	-.5	.79***	-.7	Exam062L5CAMS
74	**23**	**60**	**-.68**	**.30**	.85	-1.3	.71***	-1.0	Exam074L5LAW
115	**23**	**60**	**-.68**	**.30**	1.17	1.5	1.07	.3	Exam115L6CAM
126	**23**	**60**	**-.68**	**.30**	.88	-1.0	.79***	-.7	Exam126L6CAM
129	**23**	**60**	**-.68**	**.30**	.81	-1.7	.71***	-1.0	Exam129L6CAM
140	**23**	**60**	**-.68**	**.30**	.74	-2.4	.61***	-1.5	Exam140L6EEAL
143	**23**	**60**	**-.68**	**.30**	.96	-.3	1.27**	1.0	Exam143L6EEAL
149	**23**	**60**	**-.68**	**.30**	.98	-.1	.86	-.4	Exam149L6ENG
176	**23**	**60**	**-.68**	**.30**	.91	-.7	.78***	-.7	Exam176L6MED0560
180	**23**	**60**	**-.68**	**.30**	1.26	2.1	1.98**	2.8	Exam180L6MED0560
190	**23**	**60**	**-.68**	**.30**	.87	-1.1	.75***	-.9	Exam190L6MED0560
59	**22**	**60**	**-.77**	**.30**	.92	-.7	.85	-.4	Exam059L5CAMS
88	**22**	**60**	**-.77**	**.30**	.92	-.7	.76***	-.8	Exam088L5CEPS
92	**22**	**60**	**-.77**	**.30**	.88	-1.1	.76***	-.8	Exam092L5CEPS
109	**22**	**60**	**-.77**	**.30**	.97	-.3	.80	-.6	Exam109L5SCI
132	**22**	**60**	**-.77**	**.30**	.86	-1.2	.79***	-.6	Exam132L6CEPS
150	**22**	**60**	**-.77**	**.30**	1.13	1.1	2.22**	3.1	Exam150L6ENG
159	**22**	**60**	**-.77**	**.30**	1.06	.5	.93	-.1	Exam159L6ENG
168	**22**	**60**	**-.77**	**.30**	1.14	1.2	1.48**	1.5	Exam168L6ENG
178	**22**	**60**	**-.77**	**.30**	1.12	1.0	1.10	.4	Exam178L6MED0560
203	**22**	**60**	**-.77**	**.30**	1.22	1.8	1.30**	1.0	Exam203L6SCI
9	**21**	**60**	**-.86**	**.31**	1.15	1.2	.99	.1	Exam009L4GEN
15	**21**	**60**	**-.86**	**.31**	1.02	.2	.99	.1	Exam015L4GEN
67	**21**	**60**	**-.86**	**.31**	.94	-.4	.92	-.1	Exam067L5LAW
75	**21**	**60**	**-.86**	**.31**	.83	-1.5	.71***	-.9	Exam075L5LAW
97	**21**	**60**	**-.86**	**.31**	.91	-.7	.76***	-.7	Exam097L5SCI
105	**21**	**60**	**-.86**	**.31**	1.07	.6	1.02	.2	Exam105L5SCI
133	**21**	**60**	**-.86**	**.31**	1.03	.3	1.00	.1	Exam133L6CEPS
193	**21**	**60**	**-.86**	**.31**	.96	-.3	.88	-.3	Exam193L6MED0560
207	**21**	**60**	**-.86**	**.31**	1.31	2.4	1.72**	2.0	Exam207L6SCI
16	**20**	**60**	**-.96**	**.31**	1.06	.5	1.02	.2	Exam016L4GEN
19	**20**	**60**	**-.96**	**.31**	1.10	.8	.96	.0	Exam019L4GEN
65	**20**	**60**	**-.96**	**.31**	.89	-.9	1.28**	.9	Exam065L5CAMS
77	**20**	**60**	**-.96**	**.31**	.95	-.4	.82	-.5	Exam077L5CEPS
98	**20**	**60**	**-.96**	**.31**	.97	-.2	.84	-.4	Exam098L5SCI
110	**20**	**60**	**-.96**	**.31**	.91	-.7	.77***	-.6	Exam110L5SCI
118	**20**	**60**	**-.96**	**.31**	1.05	.5	.96	.0	Exam118L6CAM
7	**19**	**60**	**-1.05**	**.31**	.92	-.6	.76	-.6	Exam007L4GEN
20	**19**	**60**	**-1.05**	**.31**	1.05	.4	.97	.0	Exam020L4GEN
48	**19**	**60**	**-1.05**	**.31**	1.24	1.8	1.52**	1.4	Exam048L4GEN
80	**19**	**60**	**-1.05**	**.31**	1.02	.2	.92	-.1	Exam080L5CEPS

85	**19**	**60**	**-1.05**	**.31**	.75	-2.2	.59***	-1.3	Exam085L5CEPS
89	**19**	**60**	**-1.05**	**.31**	1.10	.8	2.19**	2.7	Exam089L5CEPS
90	**19**	**60**	**-1.05**	**.31**	.99	.0	.88	-.2	Exam090L5CEPS
100	**19**	**60**	**-1.05**	**.31**	.99	.0	.94	-.1	Exam100L5SCI
101	**19**	**60**	**-1.05**	**.31**	.98	-.1	.79	-.5	Exam101L5SCI
103	**19**	**60**	**-1.05**	**.31**	.96	-.3	.92	-.1	Exam103L5SCI
104	**19**	**60**	**-1.05**	**.31**	1.15	1.2	1.12	.4	Exam104L5SCI
157	**19**	**60**	**-1.05**	**.31**	.94	-.5	.88	-.2	Exam157L6ENG
53	**18**	**60**	**-1.15**	**.31**	.80	-1.6	.65***	-.9	Exam053L5CAMS
58	**18**	**60**	**-1.15**	**.31**	.95	-.4	.76***	-.6	Exam058L5CAMS
79	**18**	**60**	**-1.15**	**.31**	1.09	.7	1.00	.1	Exam079L5CEPS
106	**18**	**60**	**-1.15**	**.31**	1.00	.0	.81	-.4	Exam106L5SCI
130	**18**	**60**	**-1.15**	**.31**	.96	-.3	.88	-.2	Exam130L6CAM
169	**18**	**60**	**-1.15**	**.31**	.92	-.6	.80	-.5	Exam169L6ENG
177	**18**	**60**	**-1.15**	**.31**	.93	-.5	.95	.0	Exam177L6MED0560
11	**17**	**60**	**-1.25**	**.32**	.90	-.7	.72***	-.7	Exam011L4GEN
21	**17**	**60**	**-1.25**	**.32**	.79	-1.6	.61***	-1.0	Exam021L4GEN
23	**17**	**60**	**-1.25**	**.32**	.91	-.7	.70***	-.7	Exam023L4GEN
25	**17**	**60**	**-1.25**	**.32**	.94	-.4	.98	.1	Exam025L4GEN
71	**17**	**60**	**-1.25**	**.32**	.84	-1.2	.68***	-.8	Exam071L5LAW
91	**17**	**60**	**-1.25**	**.32**	1.06	.5	.86	-.2	Exam091L5CEPS
14	**16**	**60**	**-1.35**	**.32**	1.00	.1	1.65**	1.5	Exam014L4GEN
24	**16**	**60**	**-1.35**	**.32**	1.00	.0	.80	-.4	Exam024L4GEN
28	**16**	**60**	**-1.35**	**.32**	1.06	.5	.86	-.2	Exam028L4GEN
35	**16**	**60**	**-1.35**	**.32**	1.02	.2	1.05	.3	Exam035L4GEN
41	**16**	**60**	**-1.35**	**.32**	.89	-.8	.72***	-.6	Exam041L4GEN
44	**16**	**60**	**-1.35**	**.32**	.98	-.1	.82	-.3	Exam044L4GEN
68	**16**	**60**	**-1.35**	**.32**	.96	-.3	.94	.0	Exam068L5LAW
72	**16**	**60**	**-1.35**	**.32**	.90	-.7	.95	.0	Exam072L5LAW
96	**16**	**60**	**-1.35**	**.32**	1.05	.4	.92	-.1	Exam096L5SCI
189	**16**	**60**	**-1.35**	**.32**	.94	-.4	.78***	-.4	Exam189L6MED0560
195	**16**	**60**	**-1.35**	**.32**	1.08	.6	1.05	.3	Exam195L6SCI
26	**15**	**60**	**-1.46**	**.33**	1.18	1.2	1.23**	.6	Exam026L4GEN
29	**15**	**60**	**-1.46**	**.33**	.93	-.4	.77***	-.4	Exam029L4GEN
31	**15**	**60**	**-1.46**	**.33**	.87	-.8	.70***	-.6	Exam031L4GEN
45	**15**	**60**	**-1.46**	**.33**	1.13	.9	1.07	.3	Exam045L4GEN
49	**15**	**60**	**-1.46**	**.33**	.76	-1.7	.55***	-1.1	Exam049L4GEN
54	**15**	**60**	**-1.46**	**.33**	.87	-.8	.94	.0	Exam054L5CAMS
81	**15**	**60**	**-1.46**	**.33**	.94	-.4	1.00	.1	Exam081L5CEPS
86	**15**	**60**	**-1.46**	**.33**	1.01	.1	1.56**	1.3	Exam086L5CEPS
171	**15**	**60**	**-1.46**	**.33**	.92	-.5	.71***	-.6	Exam171L6ENG
206	**15**	**60**	**-1.46**	**.33**	1.09	.7	1.87**	1.7	Exam206L6SCI
1	**14**	**60**	**-1.57**	**.33**	.97	-.1	.76***	-.4	Exam001L4GEN
8	**14**	**60**	**-1.57**	**.33**	1.12	.8	.92	.0	Exam008L4GEN
32	**14**	**60**	**-1.57**	**.33**	.94	-.3	.75***	-.4	Exam032L4GEN
40	**14**	**60**	**-1.57**	**.33**	1.13	.8	.92	.0	Exam040L4GEN
50	**14**	**60**	**-1.57**	**.33**	1.12	.8	.95	.1	Exam050L4GEN
73	**14**	**60**	**-1.57**	**.33**	.97	-.1	.73***	-.5	Exam073L5LAW
99	**14**	**60**	**-1.57**	**.33**	.92	-.4	.72***	-.5	Exam099L5SCI
102	**14**	**60**	**-1.57**	**.33**	1.17	1.1	1.19	.5	Exam102L5SCI
107	**14**	**60**	**-1.57**	**.33**	.92	-.5	.66***	-.7	Exam107L5SCI
151	**14**	**60**	**-1.57**	**.33**	.96	-.2	.71***	-.5	Exam151L6ENG
182	**14**	**60**	**-1.57**	**.33**	.91	-.6	.66***	-.7	Exam182L6MED0560
5	**13**	**60**	**-1.68**	**.34**	1.09	.6	.89	-.1	Exam005L4GEN
6	**13**	**60**	**-1.68**	**.34**	.84	-.9	.73***	-.4	Exam006L4GEN
34	**13**	**60**	**-1.68**	**.34**	1.02	.2	.88	-.1	Exam034L4GEN
46	**13**	**60**	**-1.68**	**.34**	.92	-.5	.67***	-.6	Exam046L4GEN
51	**13**	**60**	**-1.68**	**.34**	.99	.0	.81	-.2	Exam051L4GEN
55	**13**	**60**	**-1.68**	**.34**	1.13	.8	1.24**	.6	Exam055L5CAMS
56	**13**	**60**	**-1.68**	**.34**	.85	-.9	.62***	-.7	Exam056L5CAMS
66	**13**	**60**	**-1.68**	**.34**	.99	.0	.74***	-.4	Exam066L5LAW
84	**13**	**60**	**-1.68**	**.34**	.92	-.4	.74***	-.4	Exam084L5CEPS
172	**13**	**60**	**-1.68**	**.34**	.90	-.6	.74***	-.4	Exam172L6MED0560

12	**12**	**60**	**-1.80**	**.35**	1.06	.4	1.12	.4	Exam012L4GEN
39	**12**	**60**	**-1.80**	**.35**	.85	-.8	.66***	-.6	Exam039L4GEN
42	**12**	**60**	**-1.80**	**.35**	.98	-.1	.76***	-.3	Exam042L4GEN
70	**12**	**60**	**-1.80**	**.35**	1.05	.3	.84	-.1	Exam070L5LAW
76	**12**	**60**	**-1.80**	**.35**	.90	-.5	.68***	-.5	Exam076L5LAW
82	**12**	**60**	**-1.80**	**.35**	.98	.0	.68***	-.5	Exam082L5CEPS
13	**11**	**60**	**-1.93**	**.36**	.99	.0	.82	-.1	Exam013L4GEN
63	**11**	**60**	**-1.93**	**.36**	1.01	.1	.82	-.2	Exam063L5CAMS
64	**11**	**60**	**-1.93**	**.36**	.91	-.4	.65***	-.5	Exam064L5CAMS
87	**11**	**60**	**-1.93**	**.36**	.87	-.6	.61***	-.6	Exam087L5CEPS
3	**10**	**60**	**-2.06**	**.37**	.99	.0	.76***	-.2	Exam003L4GEN
18	**10**	**60**	**-2.06**	**.37**	.96	-.1	.81	-.1	Exam018L4GEN
22	**10**	**60**	**-2.06**	**.37**	1.00	.1	.72***	-.3	Exam022L4GEN
30	**10**	**60**	**-2.06**	**.37**	1.06	.3	2.18**	1.7	Exam030L4GEN
43	**10**	**60**	**-2.06**	**.37**	1.03	.2	.74***	-.3	Exam043L4GEN
108	**10**	**60**	**-2.06**	**.37**	.86	-.6	.61***	-.5	Exam108L5SCI
4	**9**	**60**	**-2.21**	**.39**	.94	-.2	.77***	-.2	Exam004L4GEN
10	**9**	**60**	**-2.21**	**.39**	1.08	.4	.78***	-.1	Exam010L4GEN
27	**9**	**60**	**-2.21**	**.39**	1.01	.1	.99	.2	Exam027L4GEN
37	**9**	**60**	**-2.21**	**.39**	1.13	.6	1.37**	.7	Exam037L4GEN
38	**9**	**60**	**-2.21**	**.39**	1.05	.3	.84	.0	Exam038L4GEN
17	**8**	**60**	**-2.36**	**.40**	1.13	.6	1.03	.3	Exam017L4GEN
36	**8**	**60**	**-2.36**	**.40**	.92	-.2	.68***	-.3	Exam036L4GEN
47	**8**	**60**	**-2.36**	**.40**	1.21	.9	1.56**	.9	Exam047L4GEN
69	**7**	**60**	**-2.53**	**.42**	1.12	.5	1.20	.5	Exam069L5LAW
2	**6**	**60**	**-2.72**	**.45**	1.04	.2	.79***	.0	Exam002L4GEN
33	**6**	**60**	**-2.72**	**.45**	.93	-.1	.54***	-.5	Exam033L4GEN
MEAN	20.9	60.0	-.92	.32	1.01	.1	.97	.0	
P.SD	7.4	.0	.74	.03	.12	.9	.30	.8	

Notes. [a]Language Use; [b]Listening; [c]Reading. [d]Error acceptable value = less than 0.20; large unacceptable error values in bold. [e]acceptable fit range for high-stakes test = 0.8 to 1.20 (1.0 is perfect fit); **underfit over 1.20; ***overfit less than 0.80.

```
Measure Item - Map - Person
         <rare>|<more>
    4       X  +
               |
               |
               |
            X  |
               |
            X  |
               |
    3          +
               |
              T|
           XX  |
            X  |
            X  |
               |
               |
    2          +
               |
            X  |
               |
            X  |
            X S|
           XX  |
            X  |
    1      XX  +
          XXX  |  Exam124L6CAM
               |
               |  Exam142L6EEAL
            X  |T Exam146L6EEAL     Exam181L6MED0560  Exam198L6SCI
            X  |  Exam153L6ENG      Exam154L6ENG      Exam160L6ENG
                  Exam167L6ENG      Exam187L6MED0560
          XXX  |  Exam113L6CAM      Exam127L6CAM      Exam144L6EEAL
                  Exam170L6ENG      Exam184L6MED0560  Exam191L6MED0560
                  Exam199L6SCI      Exam201L6SCI
            X  |  Exam186L6MED0560  Exam204L6SCI
    0       X M+  Exam123L6CAM      Exam131L6CEPS     Exam148L6EEAL
                  Exam161L6ENG      Exam162L6ENG      Exam165L6ENG
                  Exam166L6ENG      Exam173L6MED0560
         XXXX  |S Exam116L6CAM      Exam117L6CAM      Exam137L6CEPS
                  Exam139L6CEPS     Exam158L6ENG      Exam164L6ENG
                  Exam194L6MED0560
         XXXX  |  Exam061L5CAMS     Exam141L6EEAL     Exam156L6ENG
                  Exam183L6MED0560  Exam185L6MED0560
            X  |  Exam052L4GEN      Exam094L5SCI      Exam095L5SCI
                  Exam112L6CAM      Exam114L6CAM      Exam121L6CAM
                  Exam122L6CAM      Exam134L6CEPS     Exam136L6CEPS
                  Exam138L6CEPS     Exam147L6EEAL     Exam152L6ENG
                  Exam163L6ENG      Exam188L6MED0560  Exam196L6SCI
                  Exam200L6SCI      Exam202L6SCI      Exam205L6SCI
           XX  |  Exam078L5CEPS     Exam093L5SCI      Exam111L5SCI
                  Exam119L6CAM      Exam125L6CAM      Exam135L6CEPS
                  Exam175L6MED0560
           XX  |  Exam057L5CAMS     Exam060L5CAMS     Exam062L5CAMS
                  Exam074L5LAW      Exam083L5CEPS     Exam115L6CAM
                  Exam120L6CAM      Exam126L6CAM      Exam128L6CAM
                  Exam129L6CAM      Exam140L6EEAL     Exam143L6EEAL
                  Exam145L6EEAL     Exam149L6ENG      Exam155L6ENG
                  Exam174L6MED0560  Exam176L6MED0560  Exam179L6MED0560
                  Exam180L6MED0560  Exam190L6MED0560  Exam192L6MED0560
                  Exam197L6SCI
         XXXX  |  Exam059L5CAMS     Exam088L5CEPS     Exam092L5CEPS
                  Exam109L5SCI      Exam132L6CEPS     Exam150L6ENG
                  Exam159L6ENG      Exam168L6ENG      Exam178L6MED0560
                  Exam203L6SCI
          XXX  |M Exam009L4GEN      Exam015L4GEN      Exam067L5LAW
                  Exam075L5LAW      Exam097L5SCI      Exam105L5SCI
                  Exam133L6CEPS     Exam193L6MED0560  Exam207L6SCI
```

```
-1      XX  +  Exam007L4GEN      Exam016L4GEN      Exam019L4GEN
               Exam020L4GEN      Exam048L4GEN      Exam065L5CAMS
               Exam077L5CEPS     Exam080L5CEPS     Exam085L5CEPS
               Exam089L5CEPS     Exam090L5CEPS     Exam098L5SCI
               Exam100L5SCI      Exam101L5SCI      Exam103L5SCI
               Exam104L5SCI      Exam110L5SCI      Exam118L6CAM
               Exam157L6ENG
        XX  |  Exam053L5CAMS     Exam058L5CAMS     Exam079L5CEPS
               Exam106L5SCI      Exam130L6CAM      Exam169L6ENG
               Exam177L6MED0560
         X  |  Exam011L4GEN      Exam021L4GEN      Exam023L4GEN
               Exam025L4GEN      Exam071L5LAW      Exam091L5CEPS
       XXX S|  Exam014L4GEN      Exam024L4GEN      Exam028L4GEN
               Exam035L4GEN      Exam041L4GEN      Exam044L4GEN
               Exam068L5LAW      Exam072L5LAW      Exam096L5SCI
               Exam189L6MED0560  Exam195L6SCI
        XX  |  Exam026L4GEN      Exam029L4GEN      Exam031L4GEN
               Exam045L4GEN      Exam049L4GEN      Exam054L5CAMS
               Exam081L5CEPS     Exam086L5CEPS     Exam171L6ENG
               Exam206L6SCI
        XX  |S Exam001L4GEN      Exam005L4GEN      Exam006L4GEN
               Exam008L4GEN      Exam032L4GEN      Exam034L4GEN
               Exam040L4GEN      Exam046L4GEN      Exam050L4GEN
               Exam051L4GEN      Exam055L5CAMS     Exam056L5CAMS
               Exam066L5LAW      Exam073L5LAW      Exam084L5CEPS
               Exam099L5SCI      Exam102L5SCI      Exam107L5SCI
               Exam151L6ENG      Exam172L6MED0560  Exam182L6MED0560
            |  Exam012L4GEN      Exam039L4GEN      Exam042L4GEN
               Exam070L5LAW      Exam076L5LAW      Exam082L5CEPS
        XX  |  Exam013L4GEN      Exam063L5CAMS     Exam064L5CAMS
               Exam087L5CEPS
-2          +
            |  Exam003L4GEN      Exam018L4GEN      Exam022L4GEN
               Exam030L4GEN      Exam043L4GEN      Exam108L5SCI
            |  Exam004L4GEN      Exam010L4GEN      Exam027L4GEN
               Exam037L4GEN      Exam038L4GEN
         X  |T Exam017L4GEN      Exam036L4GEN      Exam047L4GEN
            |  Exam069L5LAW
            |
           T|  Exam002L4GEN      Exam033L4GEN
            |
-3          +
      <freq>|<less>
```

Figure O1. Person-Item Map

Appendix P: Detailed test takers' questionnaire frequency analyses results

The following tables give some information about the administration of the Moodle-hosted test based on the questionnaire frequency analyses results. Table P1 shows that there was a total of 174 questionnaire respondents coming from 14 classes or sections in different programs and levels.

Table P1. *Questionnaire Respondents' Disciplinary Areas and Courses/Levels*

		Disciplinary Areas*								
Level	Course Code	GEN	COM	SCI	MED/NUR	ENG	Law	AGR	EEAL	Totals By Level
4	340	46								46
5	450		9	15			11	13		48
6	560				17					80
6	604		8	9		22		15	9	
Totals By Discipline		46	17	24	17	22	11	28	9	174

Notes. GEN = General English; COM = Commerce; SCI = Sciences; MED/NURS = Medicine/Nursing; ENG = Engineering; AGR = Agriculture; EEAL = Education, English specialists, Arts, and Law.

To give more context about the questionnaire respondents who sat the test, Table P2 gives frequency details on testing session days, sections, levels and courses. It should be noted here that the test was administered over 10 testing sessions (labelled D1 to D14) to questionnaire respondents enrolled in 14 sections (labeled S1 to S14) from different levels or courses, as shown in Table P2.

Table P2. *Testing Sessions Per Day and Section From Different Levels/Courses*

Testing Day	Section	Level	Course Code	Frequenc y	Percent	Valid Percent	Cumulative Percent
D1	S1	4	340 GEN	11	6.3	6.3	6.3
D2	S2 & S3	4	340 GEN	20	11.5	11.5	24.1
D3	S4	5	450 COM	9	5.2	5.2	29.3
D4	S5	6	560 MED/NURS	17	9.8	9.8	39.1
D5	S6 & S7	4 & 6	340 GEN & 604 ENG	37	21.3	21.3	60.3
D6	S8 & S9	6 & 5	604 COM & 450 SCI	23	13.2	13.2	73.6
D7	S10 & S11	6	604 AGR & 604 SCI	24	13.8	13.8	87.4
D8	S12	6	604 EEAL	9	5.2	5.2	92.5
D9	S13	5	450 AGR	13	7.5	7.5	100.0
D10	S14	5	450 LAW	11	6.3	6.3	12.6
Total				174	100.0	100.0	

Notes. GEN = General English; COM = Commerce; SCI = Sciences; MED/NURS = Medicine/Nursing; ENG = Engineering; AGR = Agriculture; EEAL = Education, English specialists, Arts, and Law; D1 to D10 = Testing days from day 1 to day 10; S1 to S14 = Sections from S1 to S14.

The questionnaire respondents took the test in different venues, as illustrated in Table P3. These testing venues were five computer laboratories (labeled V1, V2, V4, V5, and V7).

Table P3. *Testing Venues For Each Testing Session Including Sections, Levels, and Courses*

Venue	Testing Day	Section	Level	Course Code	Frequency	Percent	Valid Percent	Cumulative Percent
V1	D3 D5 D7 D9	S4 S6 S10 S13	5 4 6 5	450 COM 340 GEN 604 AGR 450 AGR	52	29.9	29.9	29.9
V2	D5 D6	S7 S9	6 5	604 ENG 450 SCI	37	21.3	21.3	51.1
V4	D1 D2	S1 S3	4 4	340 GEN 340 GEN	19	10.9	10.9	62.1
V5	D2 D6	S2 S8	4 6	340 GEN 604 COM	20	11.5	11.5	73.6
V7	D4 D7 D8 D10	S5 S11 S12 S14	6 6 6 5	560 MED/NURS 604 SCI 604 EEAL 450 LAW	46	26.4	26.4	100.0
Total					174	100.0	100.0	

Notes. V1 to V7 = Labels for test session venues, namely computer laboratories.

Table P4 gives the percentages of questionnaire respondents for every option selected. Two columns have also been added to the results showing the ‘agree’ and ‘strongly agree’ responses in a broad agreement category, and the ‘disagree’ and ‘strongly disagree’ responses in a broad disagreement category.

Table P4. *Frequency Analysis on Five Point Likert-Type Scale Questionnaire Items*

Item #	Question text	Strongly agree (5)	Agree (4)	Agreement (5&4)	Neutral (3)	Disagree (2)	Strongly disagree (1)	Disagreement (2&1)	No answer (9)
Q5	Liked test-taking experience	16.1%	45.4%	61.5%	28.2%	7.5%	2.3%	9.8%	.6%
Q6	Easy test navigation	58.0%	30.5%	88.5%	5.7%	2.9%	1.7%	4.6%	1.1%
Q7	Sufficient test timing	20.7%	34.5%	55.2%	18.4%	20.7%	5.2%	25.9%	.6%
Q8	Liked split screen mode for reading tests	48.3%	35.6%	83.9%	9.2%	4.6%	1.7%	6.3%	.6%
Q9	Appropriate background theme	36.2%	43.7%	79.9%	13.2%	2.9%	1.1%	4.0%	2.9%
Q10	Liked presence of count-down timer	63.2%	25.9%	89.1%	5.7%	3.4%	.6%	4.0%	1.1%
Q11	Good listening sound quality	30.5%	34.5%	65.0%	13.8%	16.7%	4.6%	21.3%	-
Q12	Liked receiving instant Moodle feedback/test results	40.2%	32.8%	73.0%	14.9%	5.7%	6.3%	12.1%	-
Q13	Clear and easy test procedures and instructions	46.0%	36.2%	82.2%	13.2%	2.3%	.6%	2.9%	1.7%
Q14	Liked typing responses	14.4%	37.9%	52.3%	31.0%	9.8%	4.6%	14.4%	2.3%
Q15	Liked using new technology	25.9%	36.2%	62.1%	18.4%	10.9%	5.7%	16.6%	2.9%
Q16	Test reflecting true language ability	14.9%	35.1%	50.0%	27.6%	12.1%	8.0%	20.1%	2.3%
Q17	Would like to take Moodle tests as official exams	10.9%	19.0%	29.9%	24.7%	19.0%	25.3%	44.3%	1.1%
Q20	Technical problems present during exam	6.3%	12.1%	18.4%	13.2%	39.1%	28.7%	67.8%	.6%

Q21	Efficient network	44.8%	40.2%	85%	8.0%	5.2%	1.7%	6.9%	-
Q22	Audio files loading quickly	47.7%	39.1%	86.8%	8.6%	3.4%	1.1%	4.5%	-
Q23	Computer working properly	59.2%	33.9%	93.1%	5.7%	.6%	-	.6%	.6%
Q24	Headphones working properly	43.1%	31.0%	74.1%	14.4%	9.2%	2.3%	11.5%	-
Q25	Successful log-in process	60.3%	33.3%	93.6%	4.0%	1.7%	.6%	2.3%	-
Q26	Clear pictures and graphs	44.8%	33.3%	78.1%	17.2%	4.6%	-	4.6%	-
Q27	Inappropriate font size	5.2%	16.1%	21.3%	18.4%	30.5%	29.9%	60.4%	-
Q28	Test was too long and had too many sections	28.7%	33.9%	62.6%	26.4%	8.0%	2.3%	10.3%	.6%
Q29	Staring at computer screen for long causing loss of concentration	32.8%	31.0%	63.8%	24.1%	8.0%	4.0%	12.0%	-
Q30	Staring at computer screen for long causing eye fatigue	33.9%	38.5%	72.4%	19.0%	4.6%	2.3%	6.9%	1.7%
Q31	Needed to take notes during test	17.2%	37.9%	55.1%	24.1%	11.5%	6.3%	17.8%	2.9%
Q32	Have enough experience with technology	25.9%	45.4%	71.3%	17.8%	7.5%	2.3%	9.8%	1.1%
Q33	Need extra technical training	17.2%	28.7%	45.9%	18.4%	20.7%	14.4%	35.1%	.6%

Appendix Q: Boxplots of questionnaire analysis results

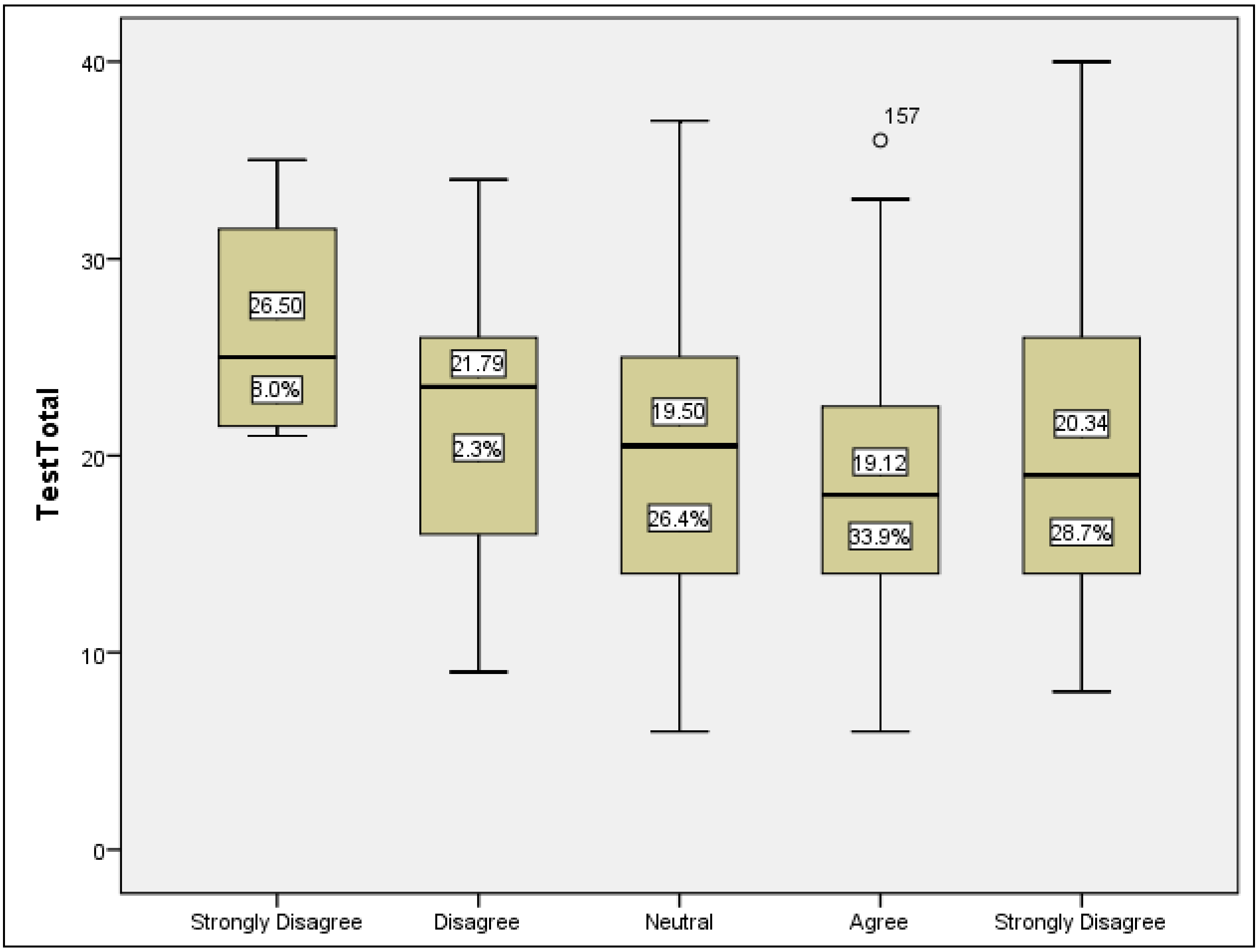


Figure Q1. Test length: Agreement (X Axis) and mean test scores (Y Axis)

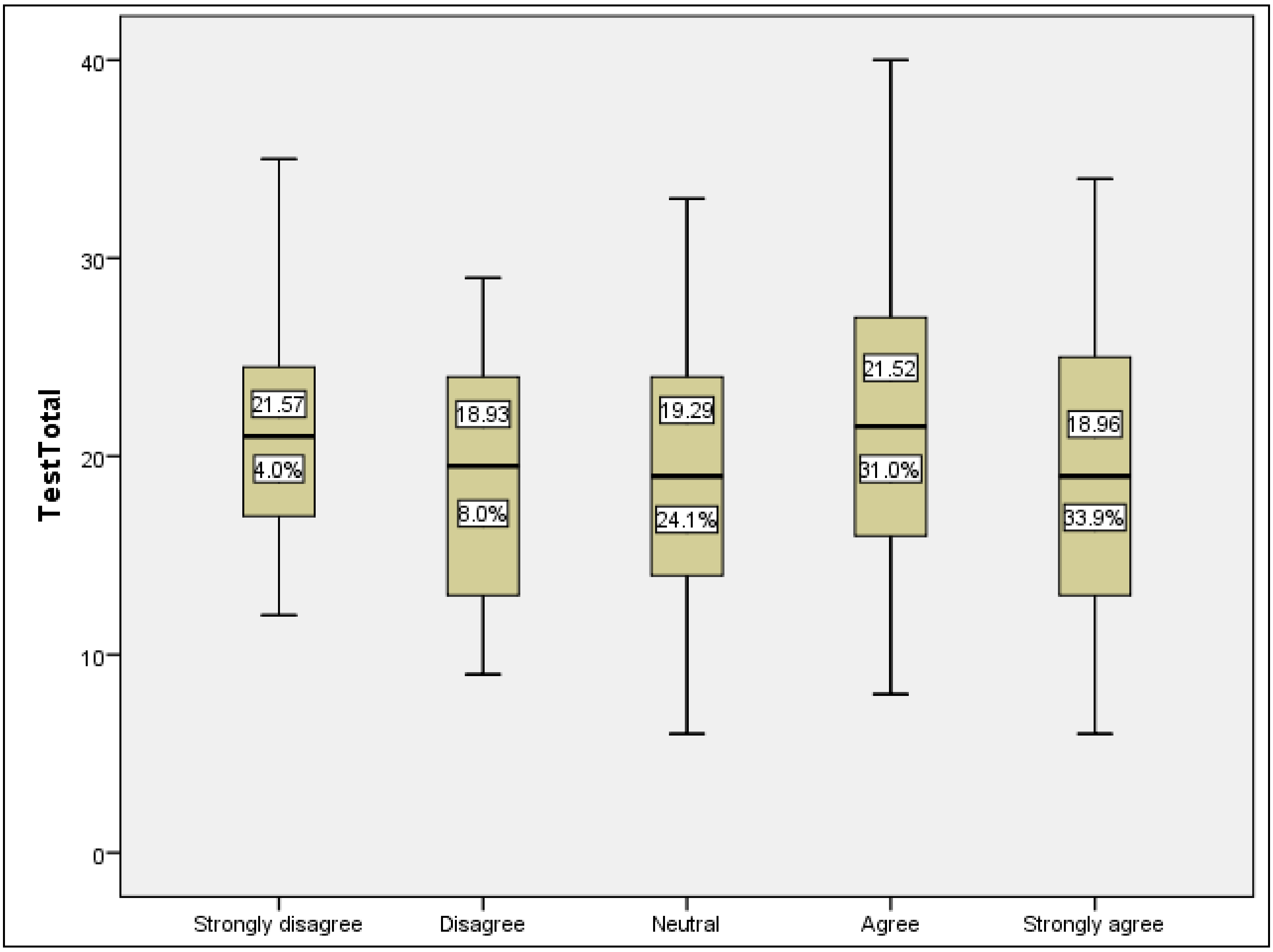


Figure Q2. Concentration loss: Agreement (X Axis) and mean test scores (Y Axis).

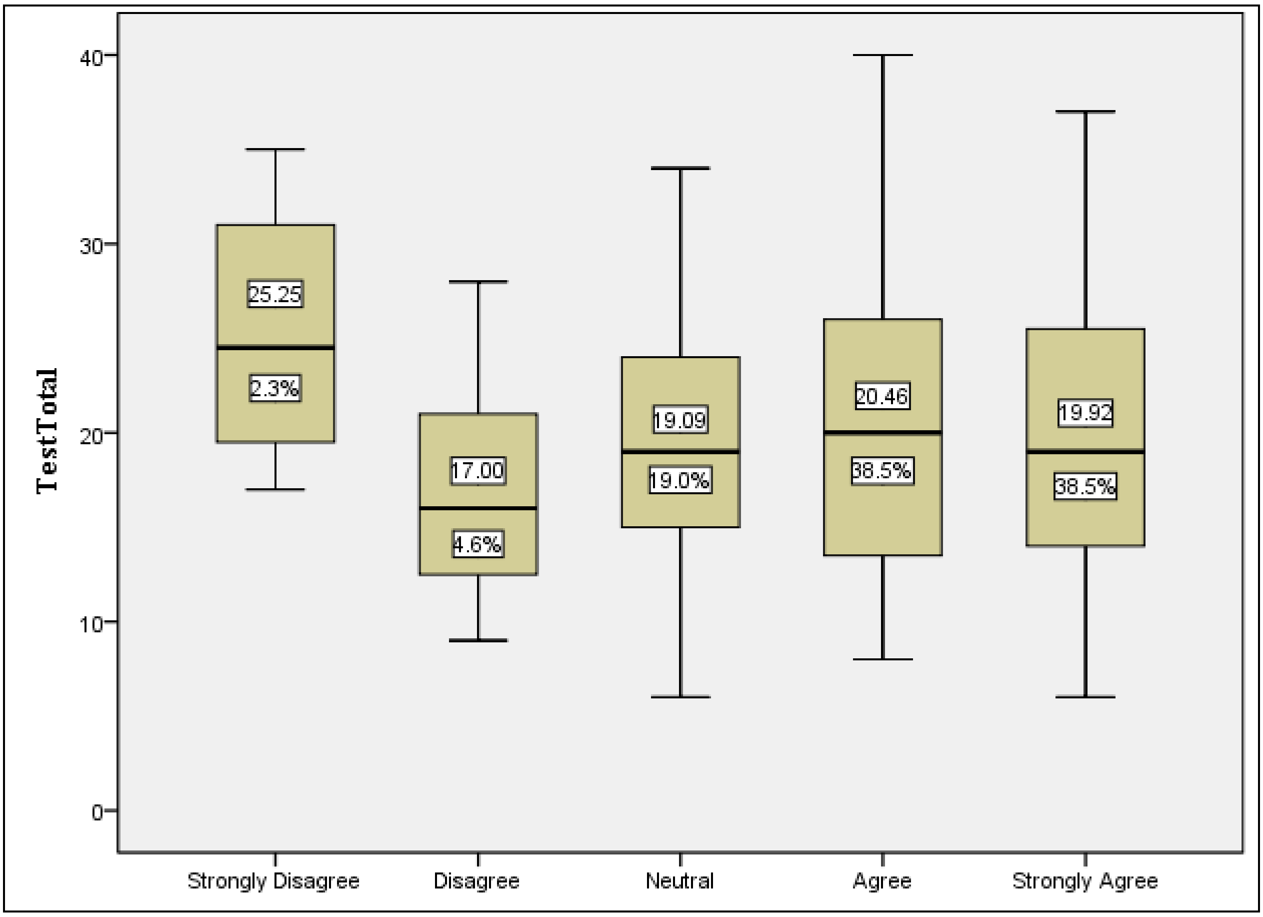


Figure Q3. Eye fatigue: Agreement (X Axis) and mean test scores (Y Axis).

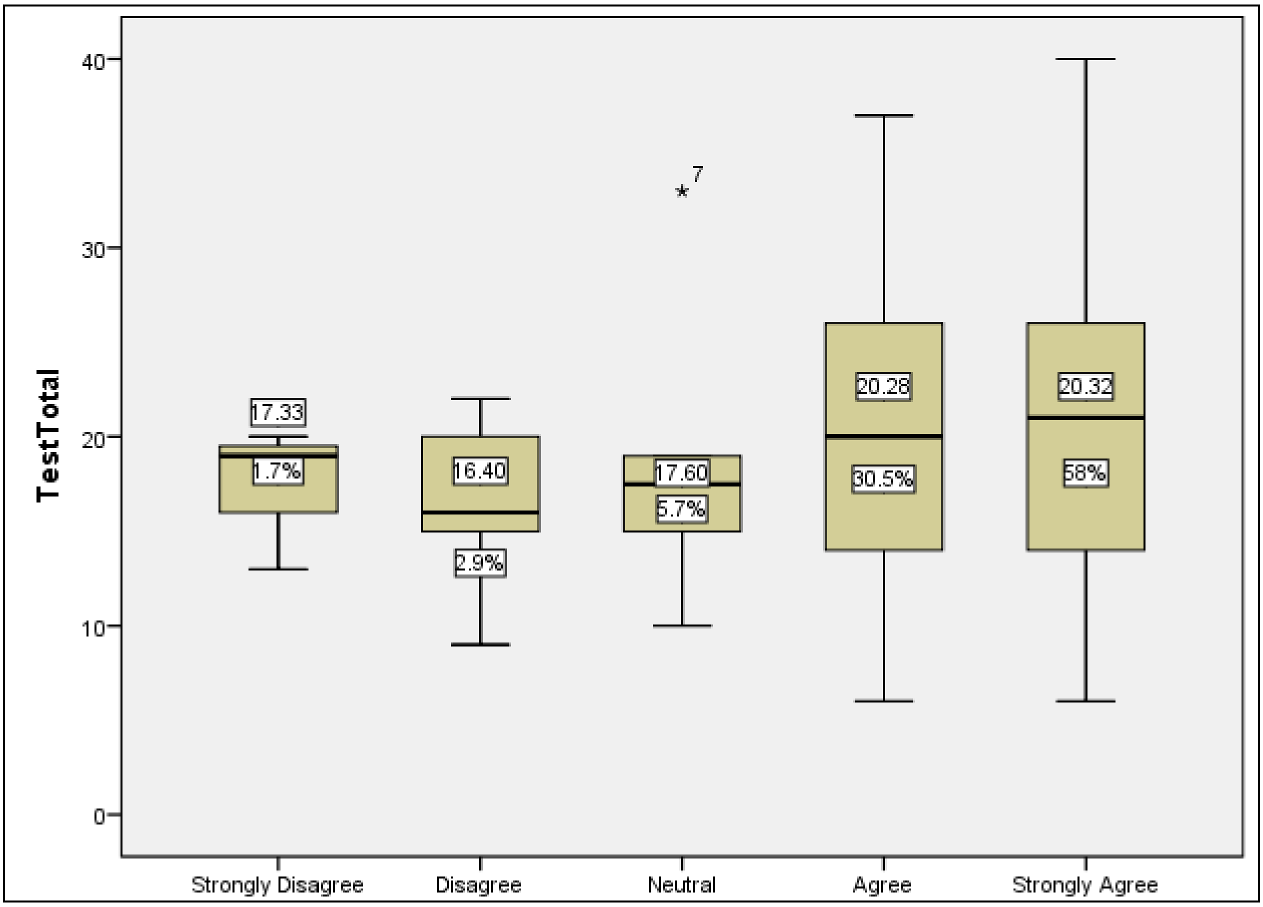


Figure Q4. Ease of navigation: Agreement (X Axis) and mean test scores (Y Axis).

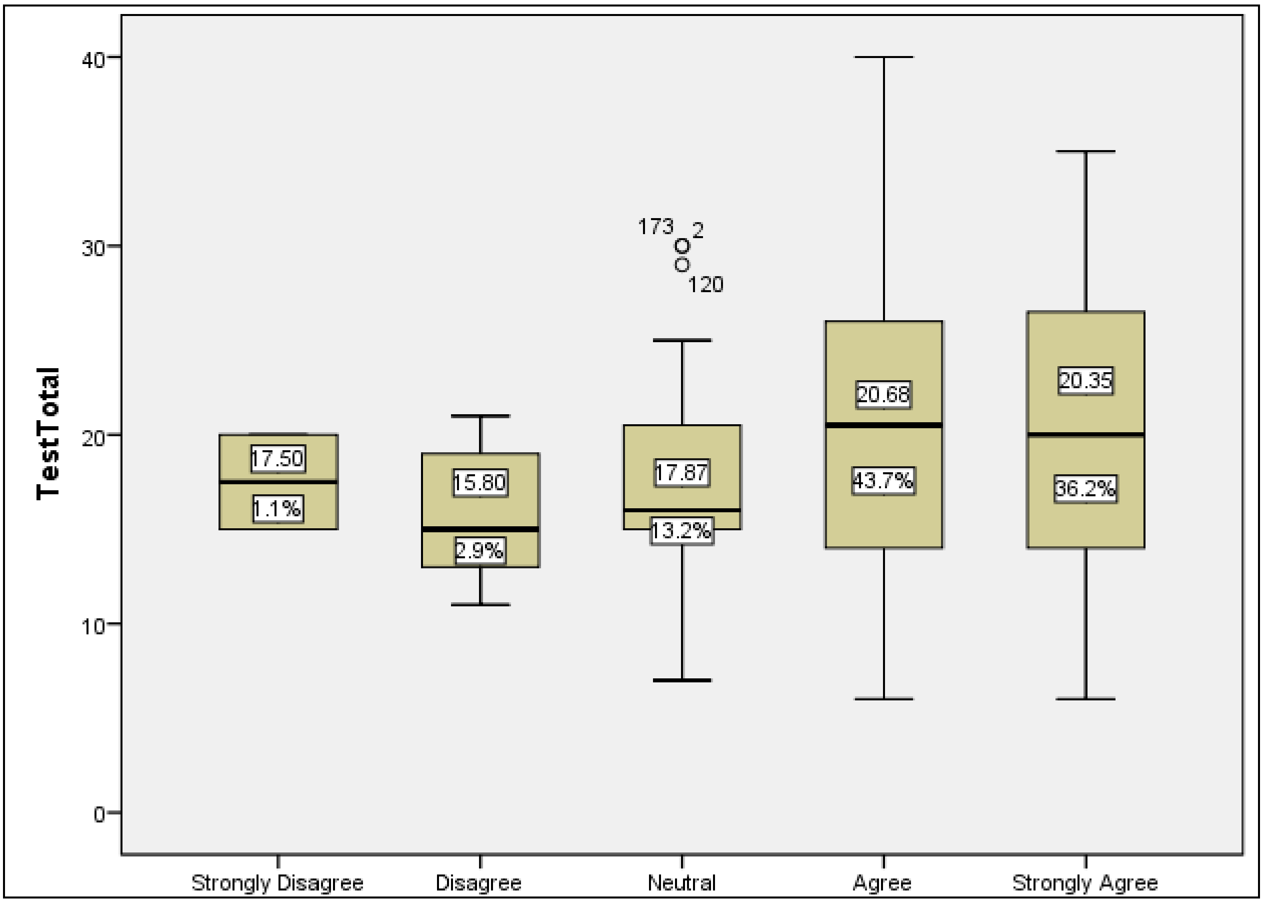


Figure Q5. Appropriate background colour: Agreement (X Axis) and mean test scores (Y Axis).

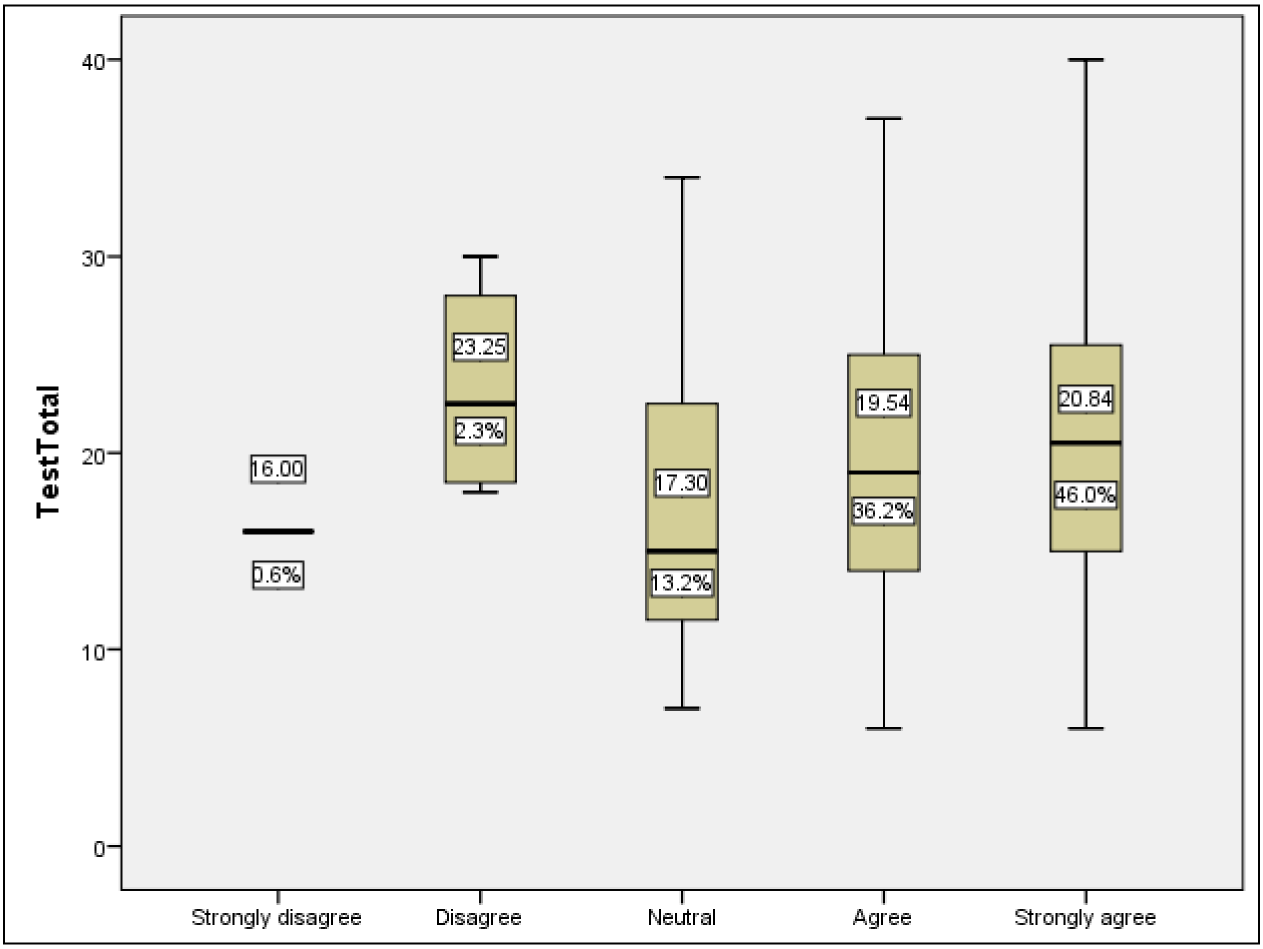


Figure Q6. Clarity of procedures and instructions: Agreement (X Axis) and mean test scores (Y Axis).

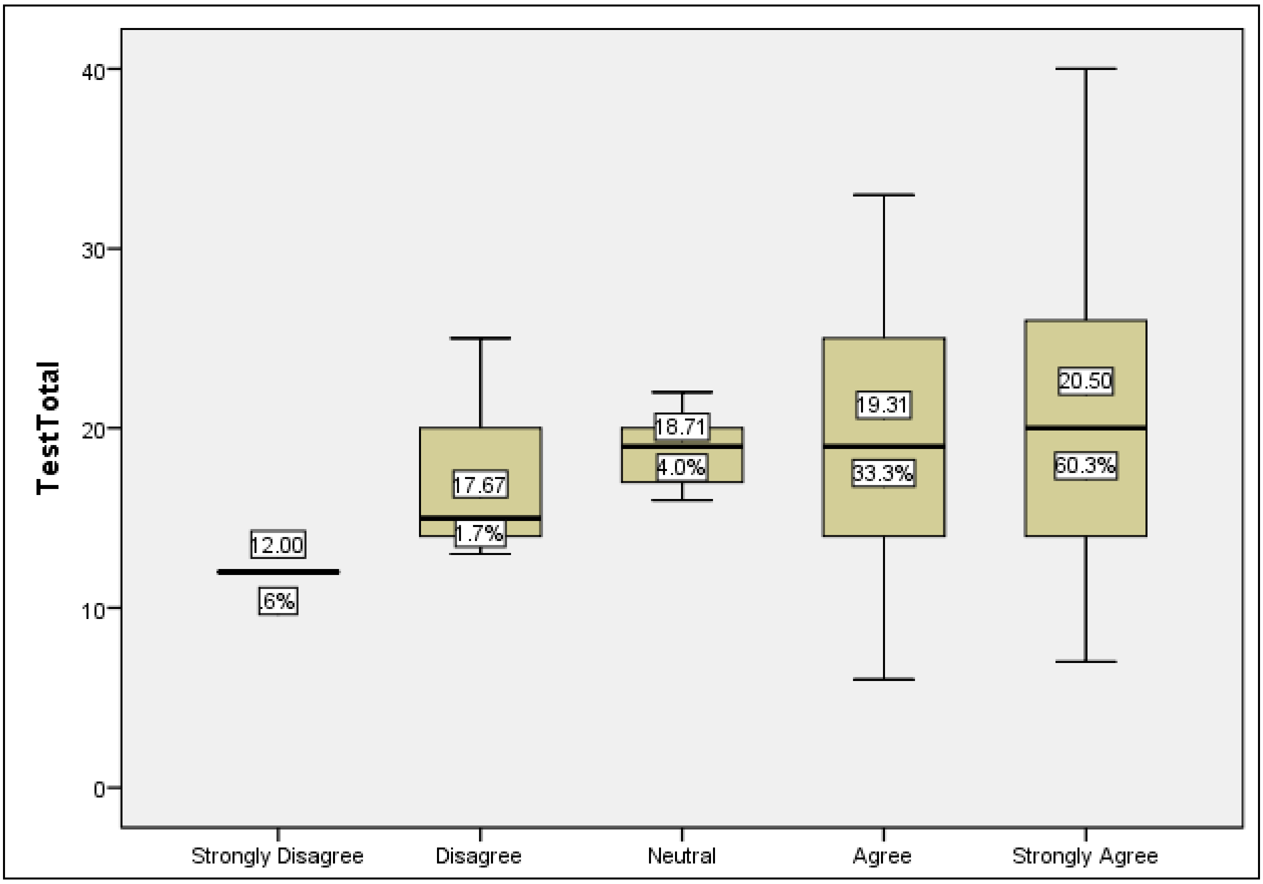


Figure Q7. Ease of test login: Agreement (X Axis) and mean test scores (Y Axis).

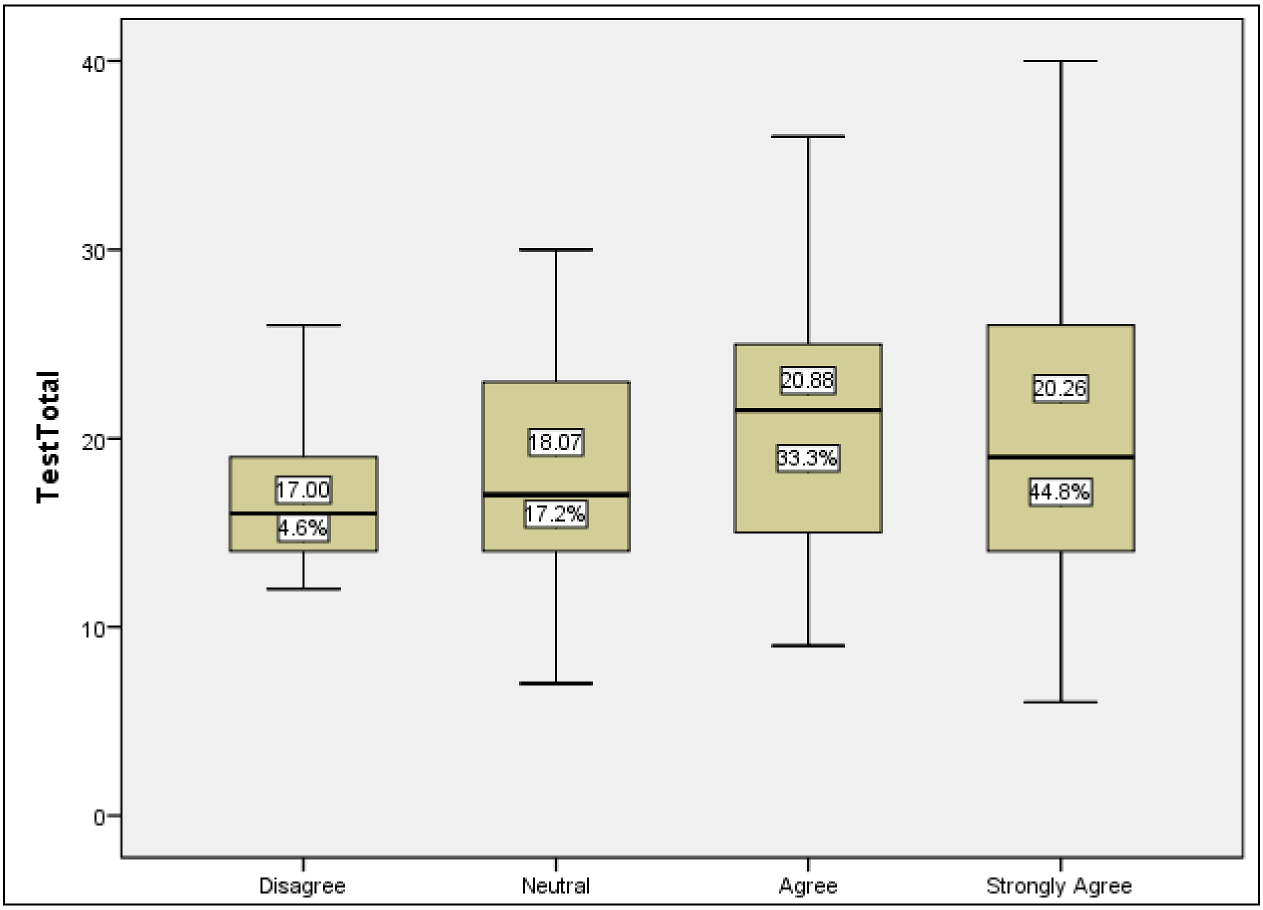


Figure Q8. Clarity of pictures and graphs: Agreement (X Axis) and mean test scores (Y Axis).

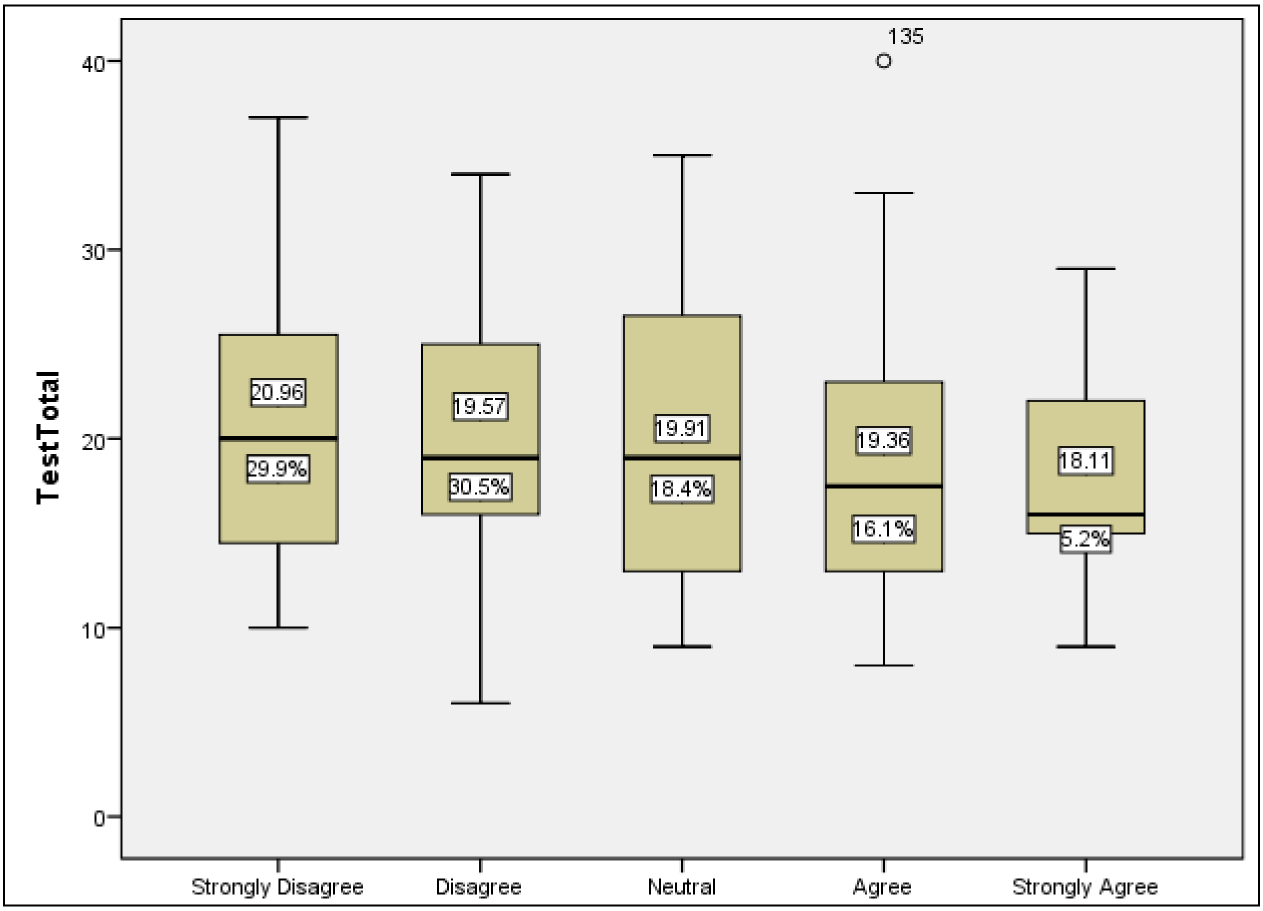


Figure Q9. Inappropriate font size: Agreement (X Axis) and mean test scores (Y Axis).

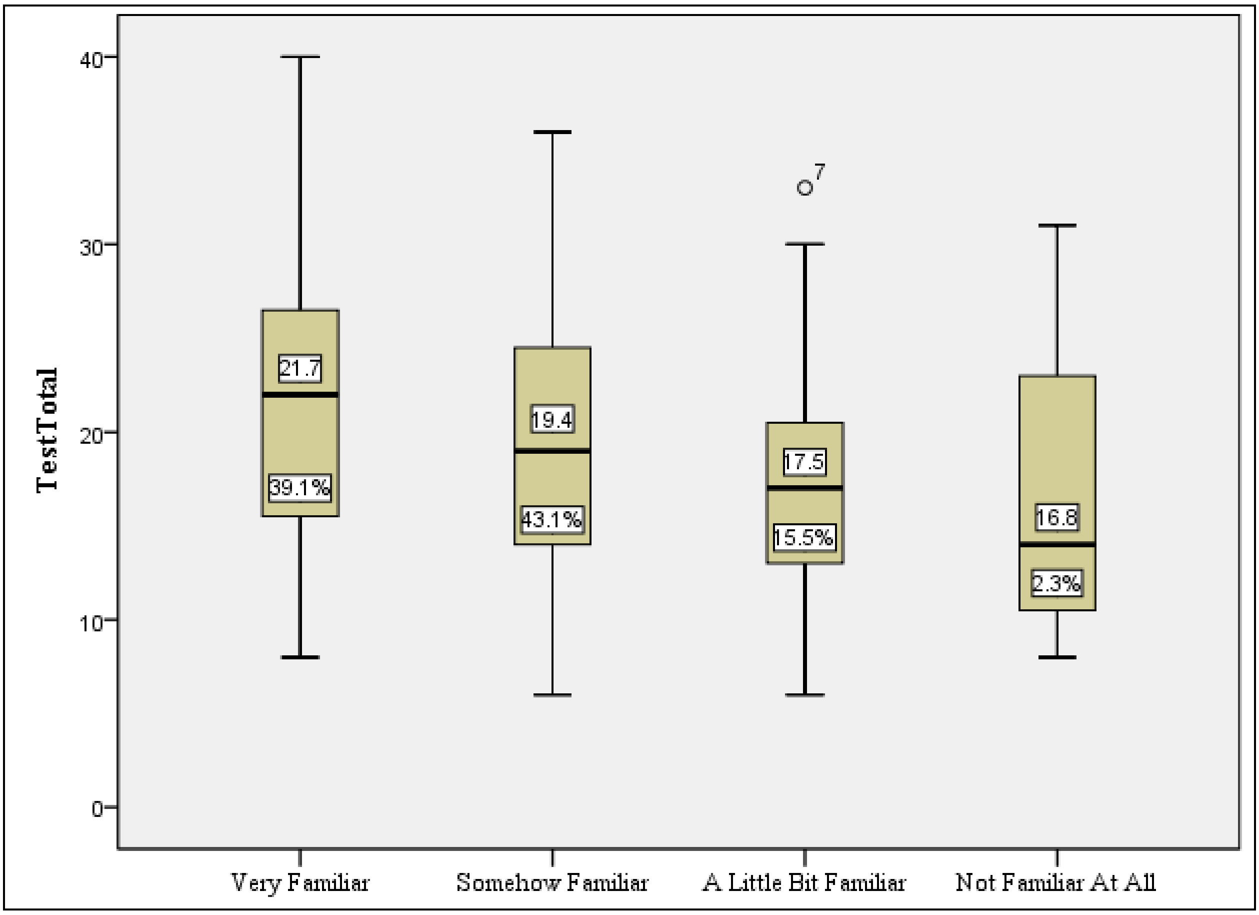


Figure Q10. Familiarity with Moodle tests (X Axis) and mean test scores (Y Axis).

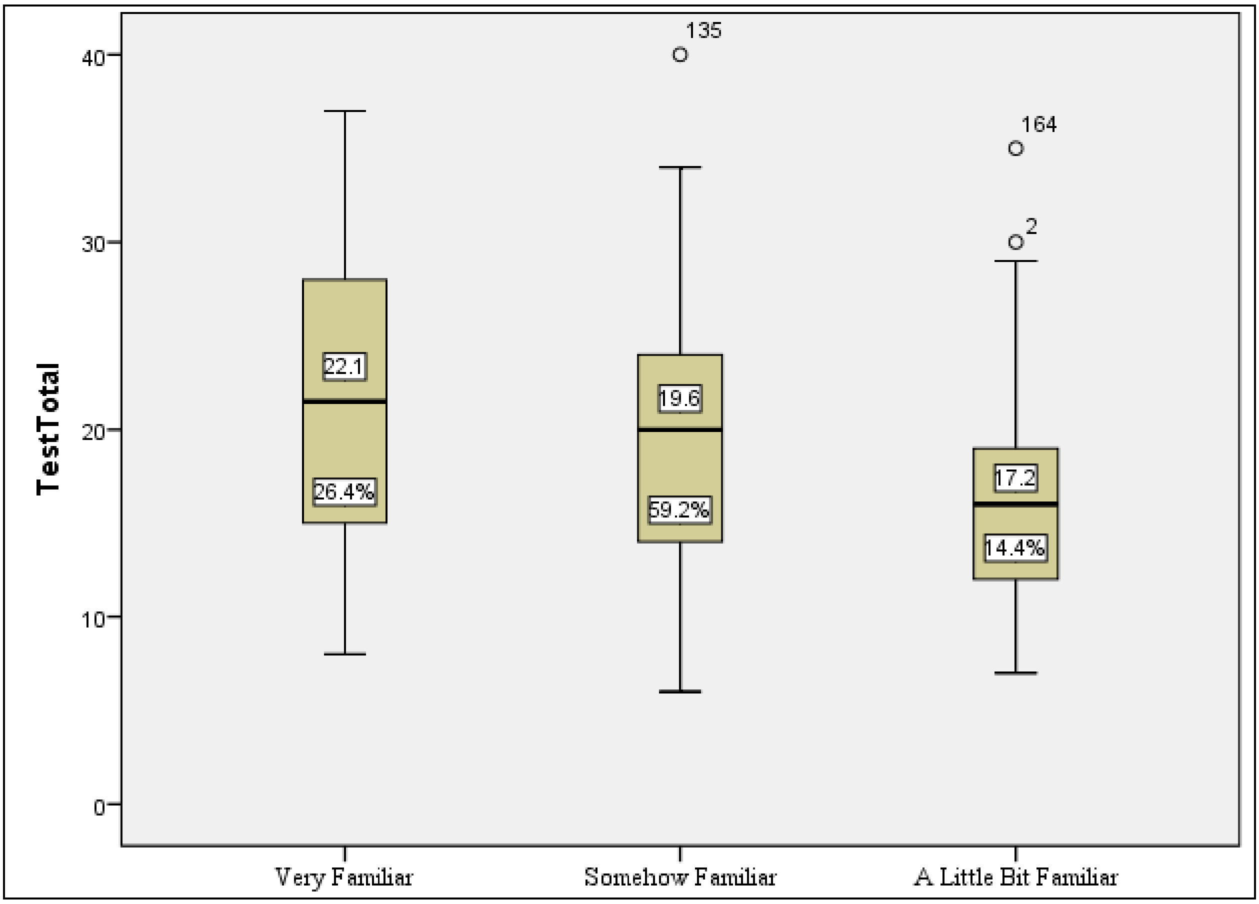


Figure Q11. Familiarity with computers (X Axis) and mean test scores (Y Axis).

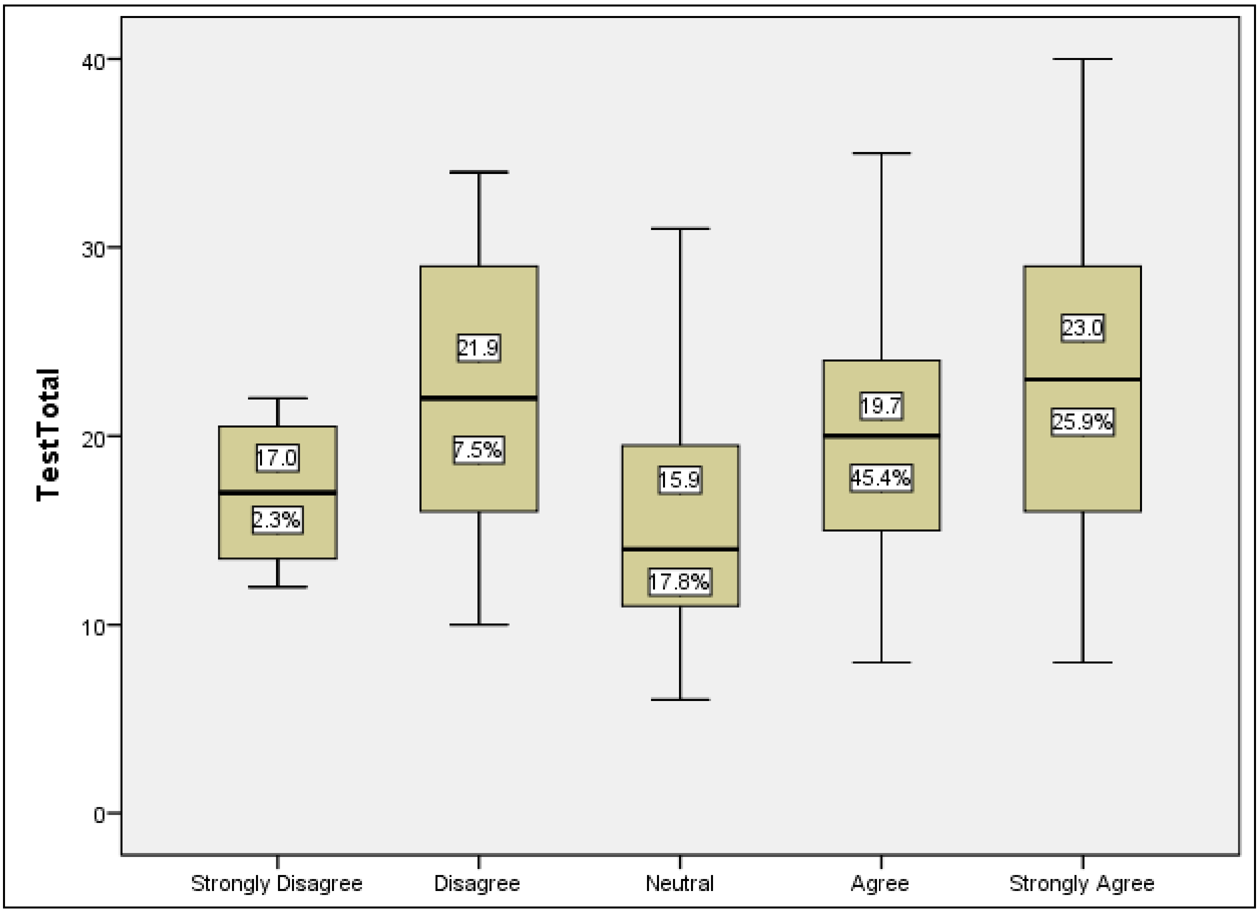


Figure Q12. Enough technology experience: Agreement (X Axis) and mean test scores (Y Axis).

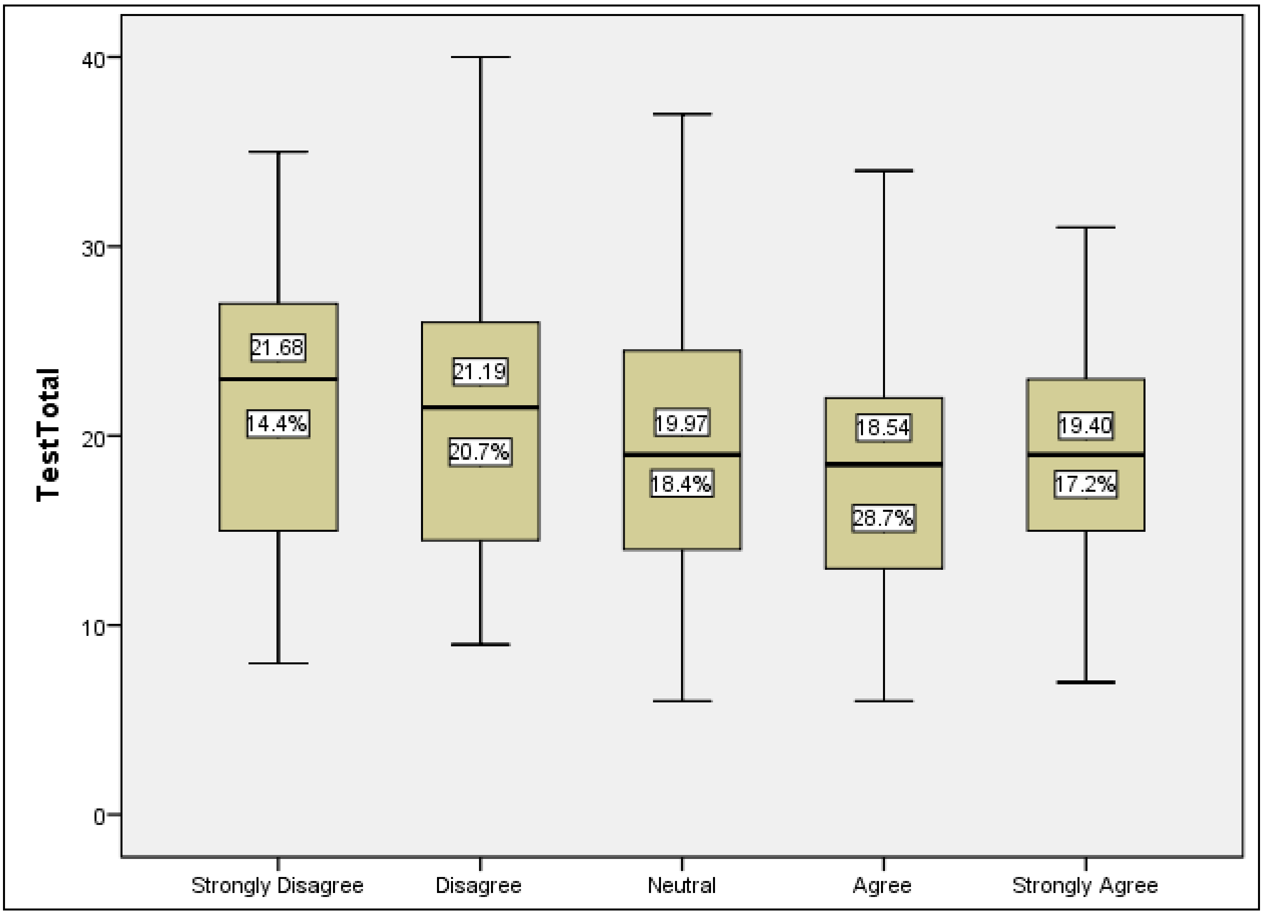


Figure Q13. Need extra technical training: Agreement (X Axis) and mean test scores (Y Axis).

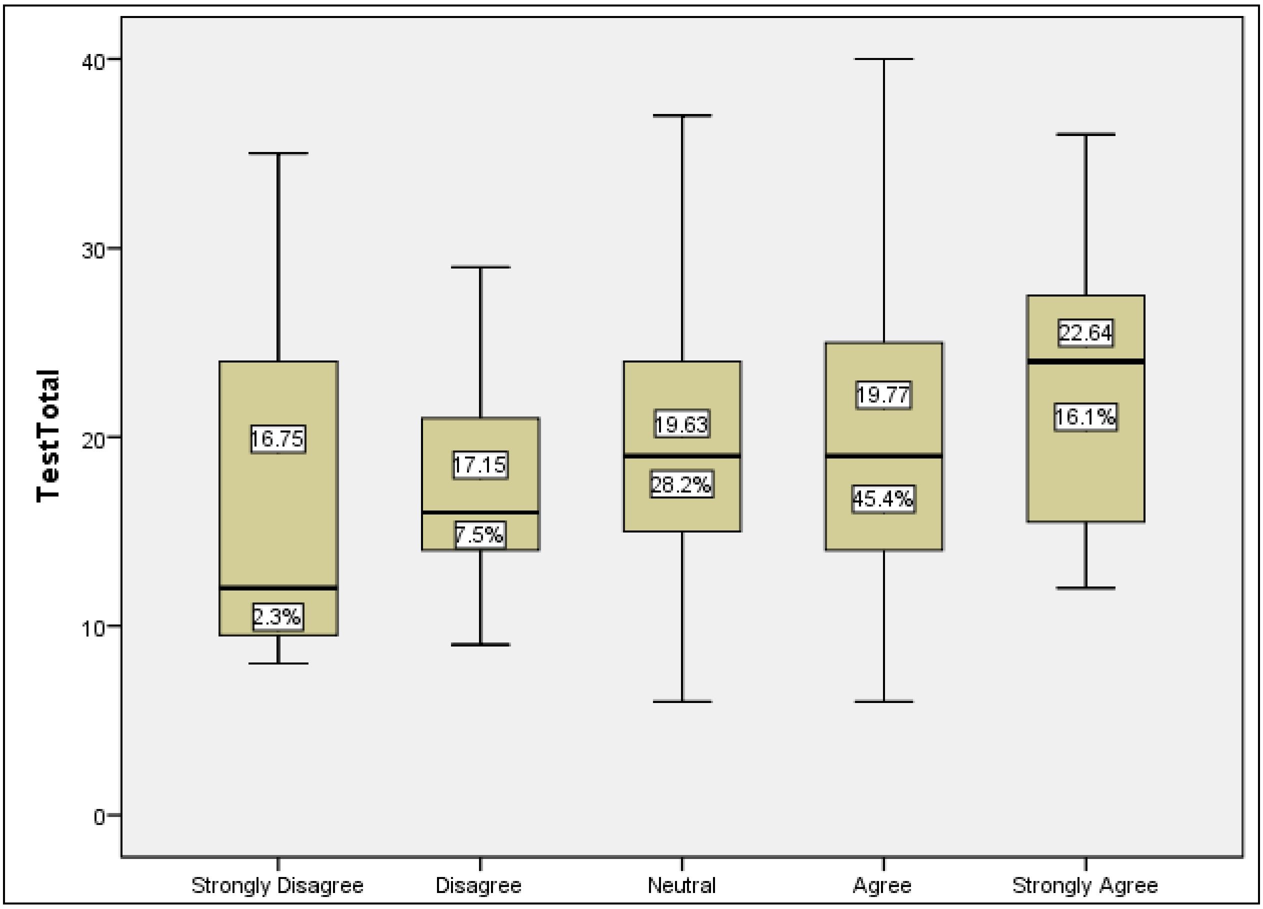


Figure Q14. Liked test-taking experience: Agreement (X Axis) and mean test scores (Y Axis).

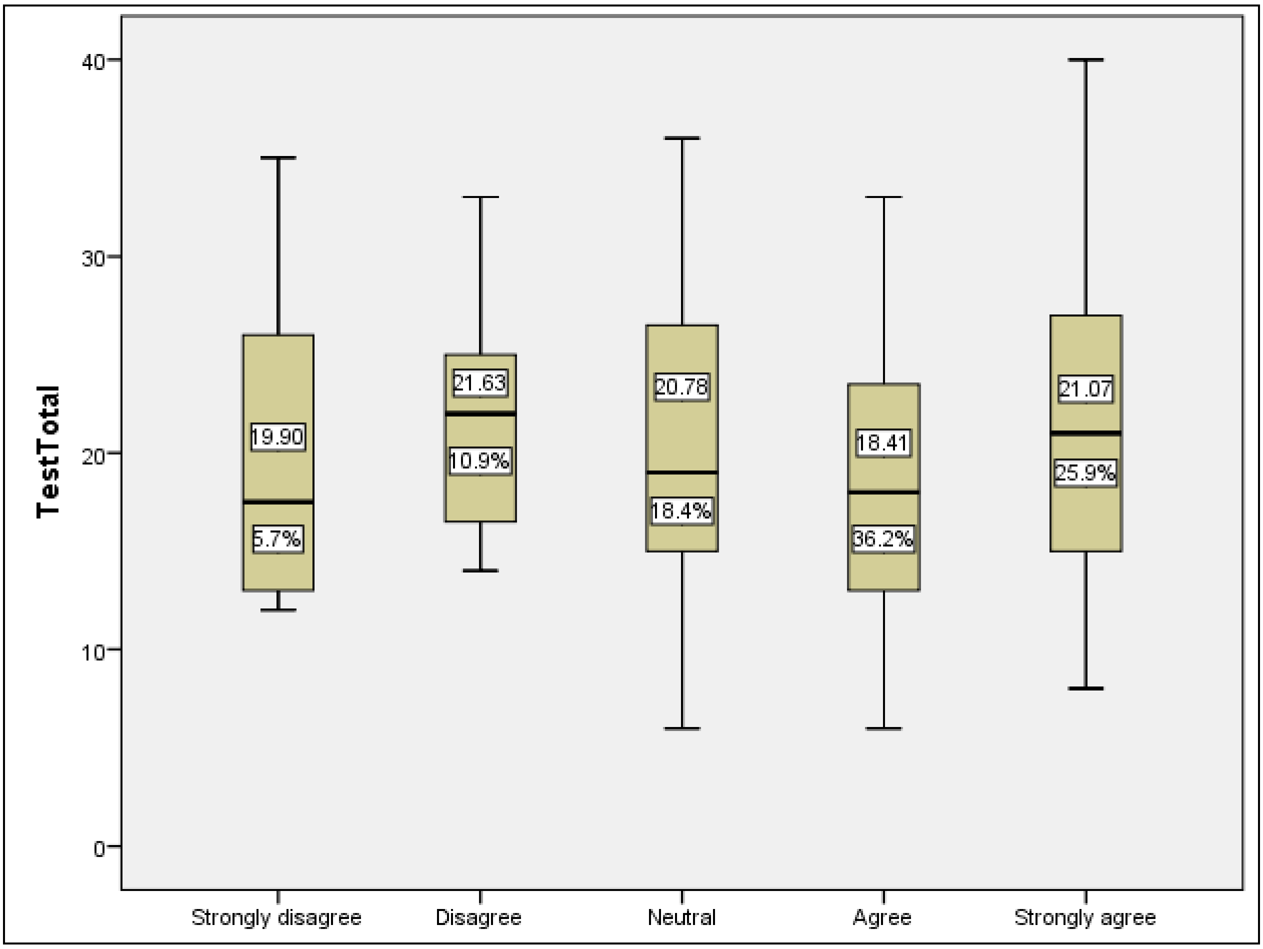


Figure Q15. Liked using new technology: Agreement (X Axis) and mean test scores (Y Axis).

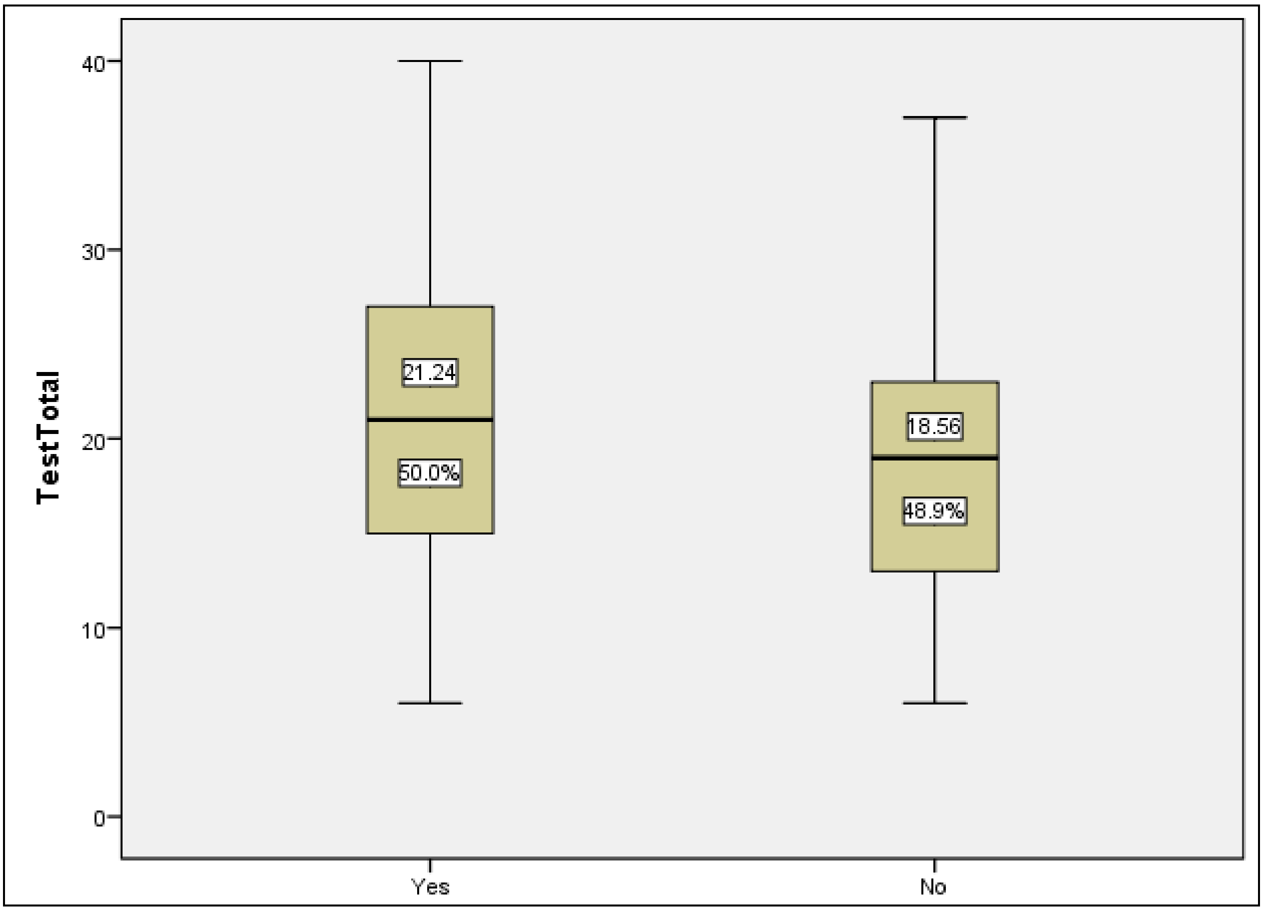


Figure Q16. Liked taking the test on Moodle: Agreement (X Axis) and mean test scores (Y Axis).

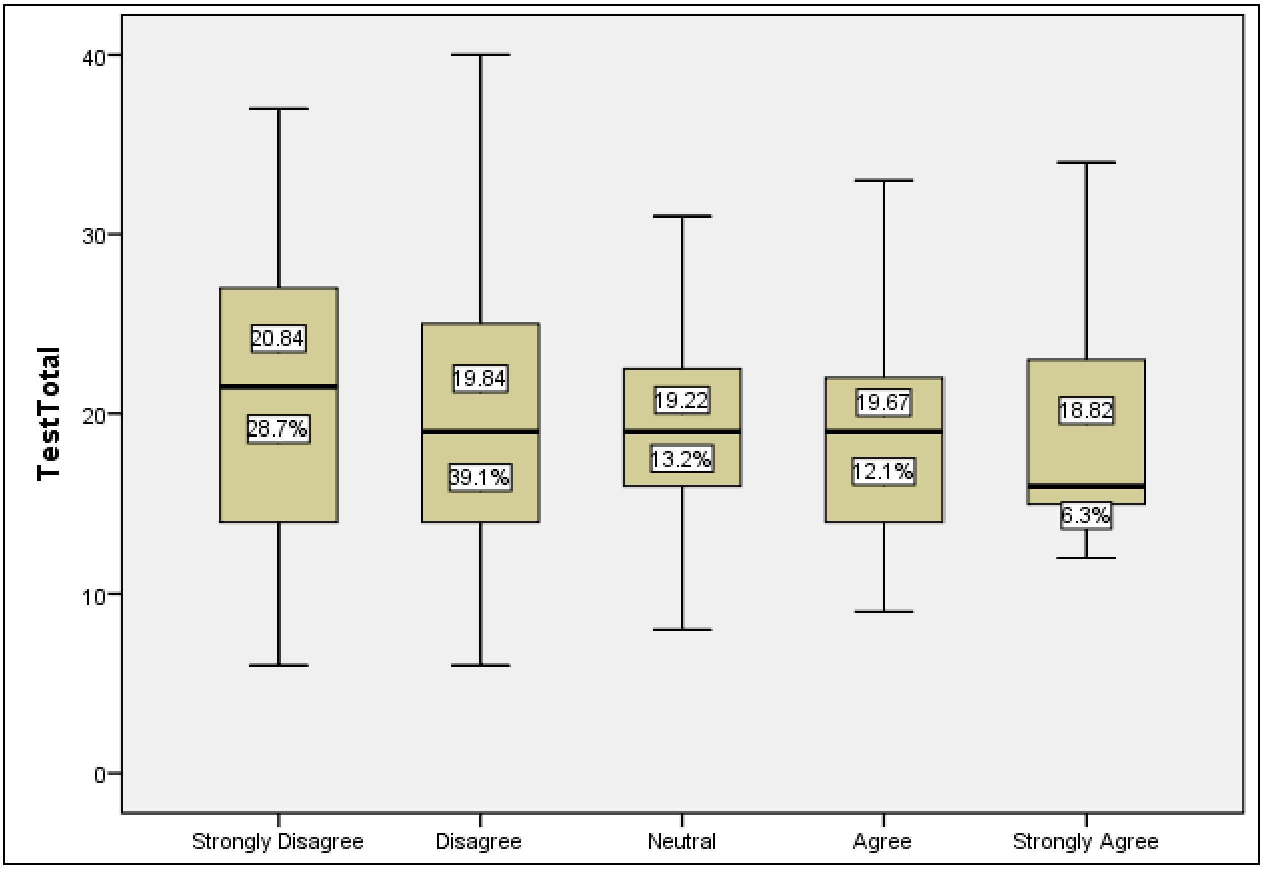


Figure Q17. Technical problems: Agreement (X Axis) and mean test scores (Y Axis).

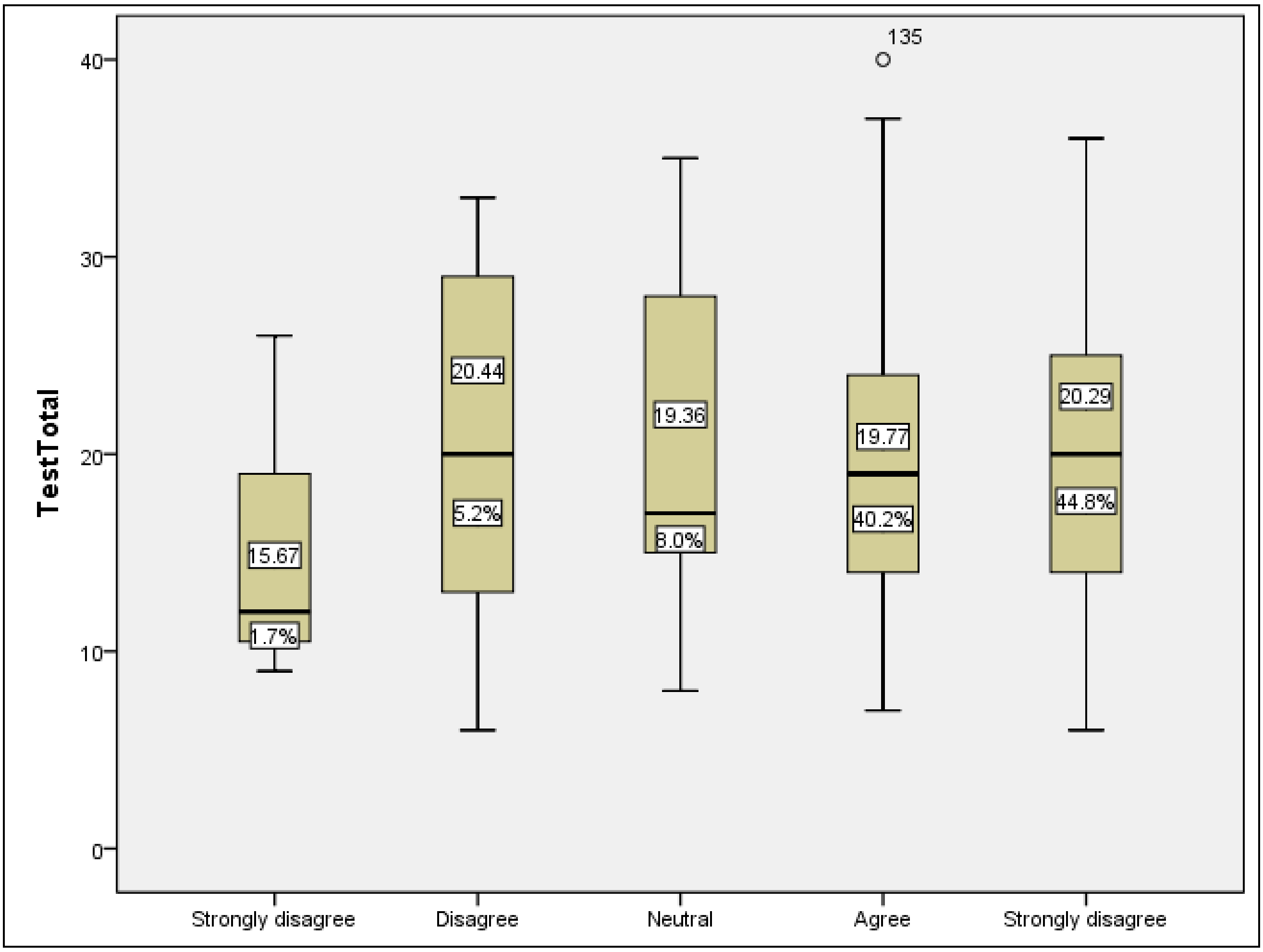


Figure Q18. Efficient network: Agreement (X Axis) and mean test scores (Y Axis).

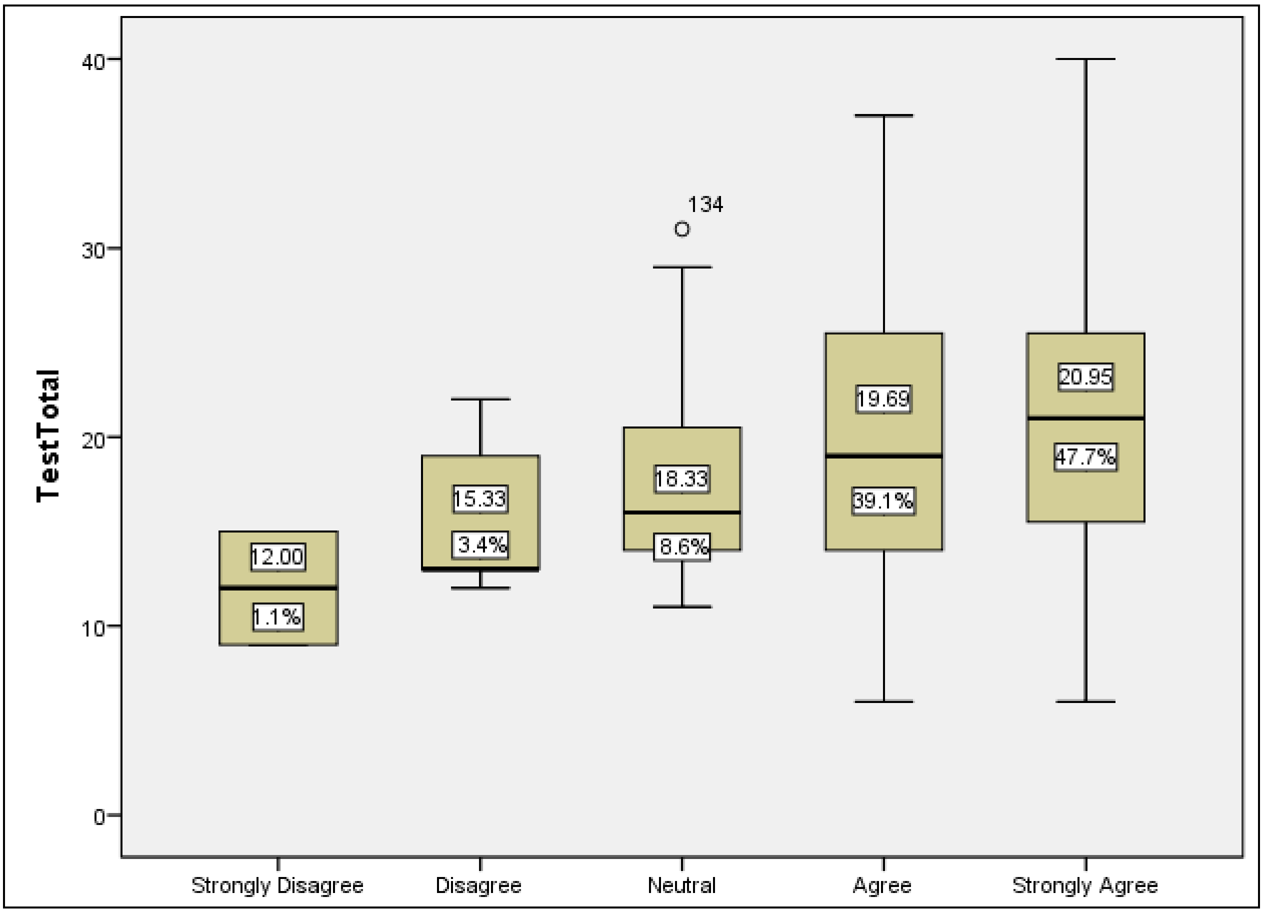


Figure Q19. Audio file loaded quickly: Agreement (X Axis) and mean test total scores (Y Axis).

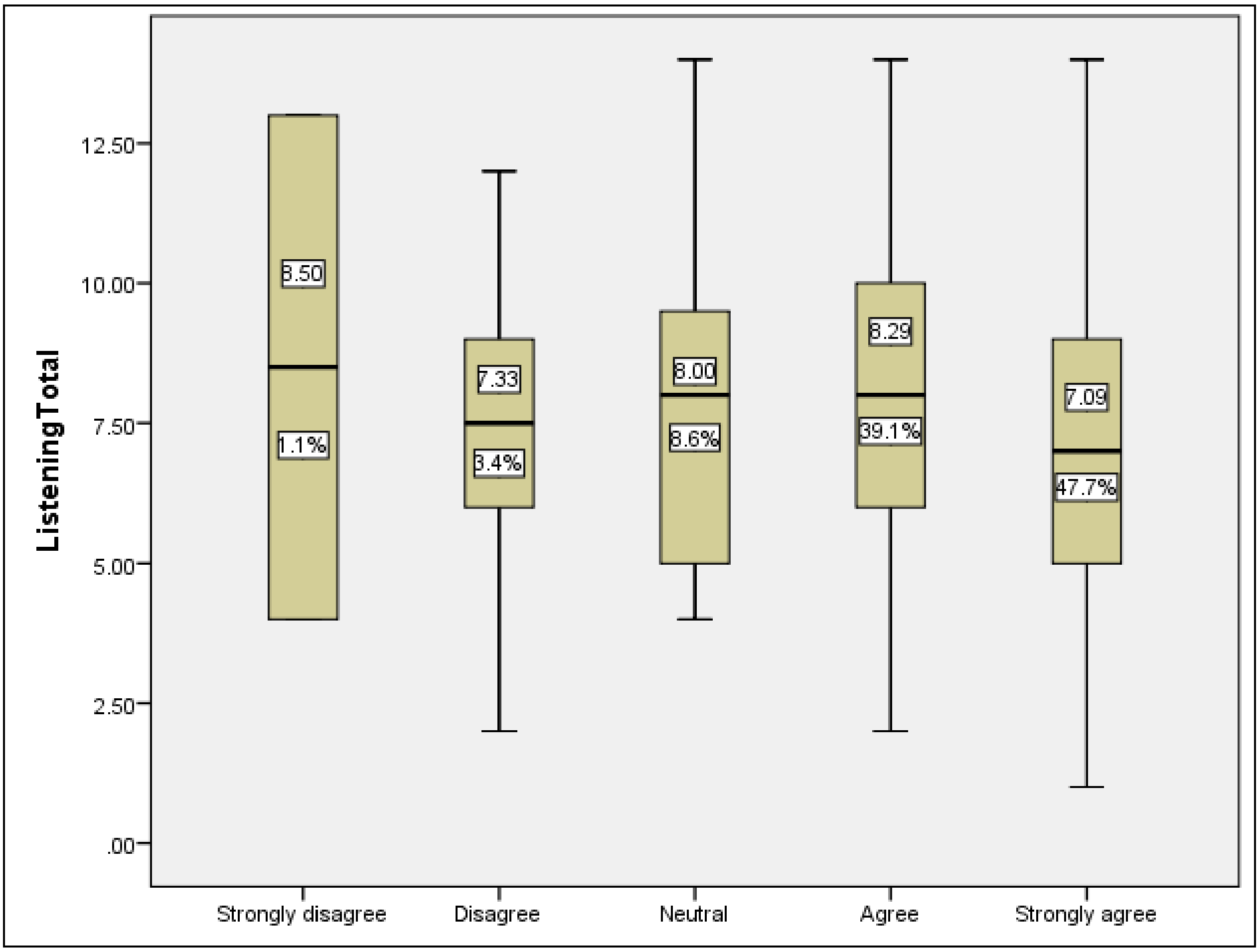


Figure Q20. Audio file loaded quickly: Agreement (X Axis) and mean listening test scores (Y Axis).

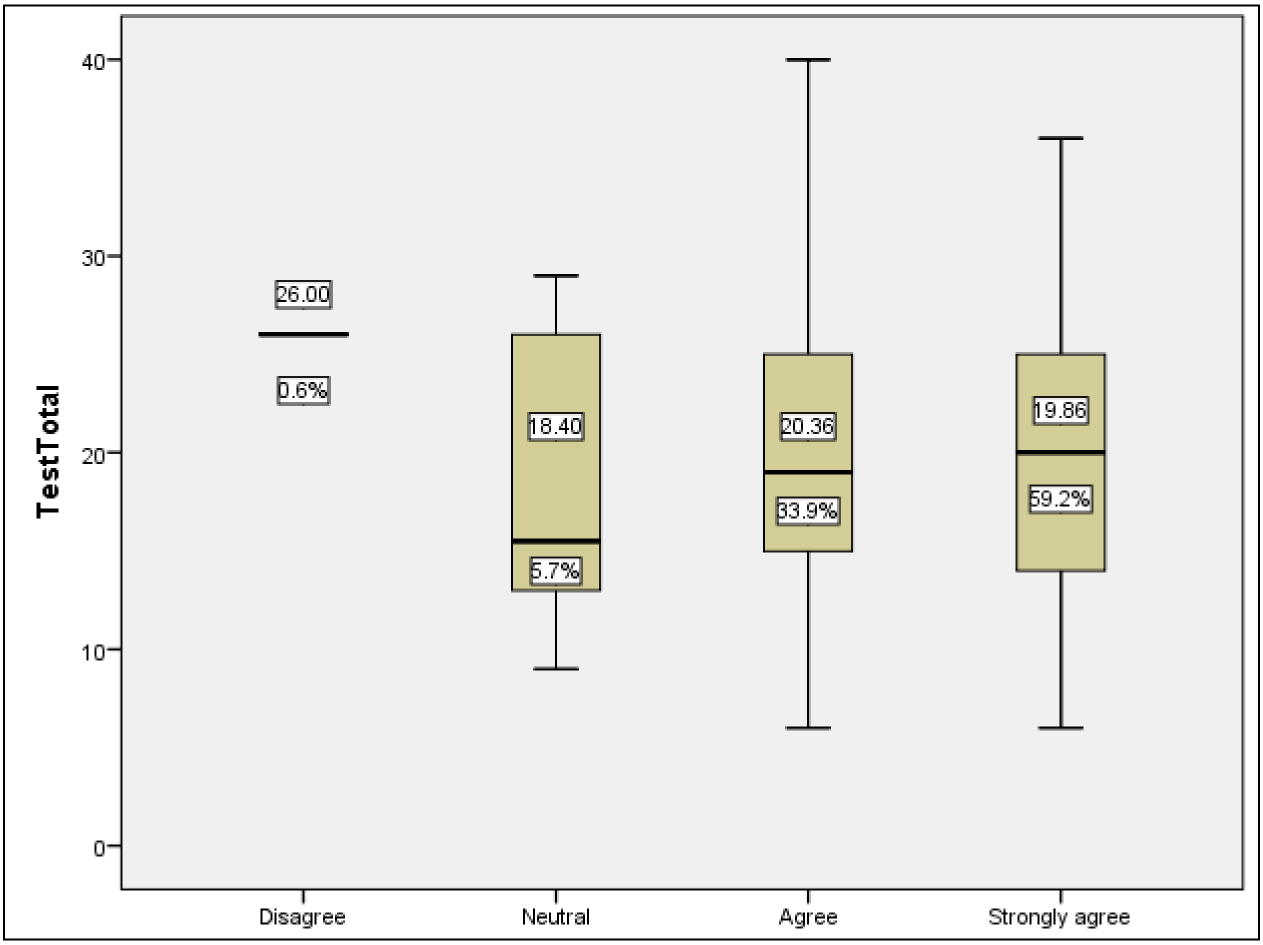


Figure Q21. Computer working properly: Agreement (X Axis) and mean test scores (Y Axis).

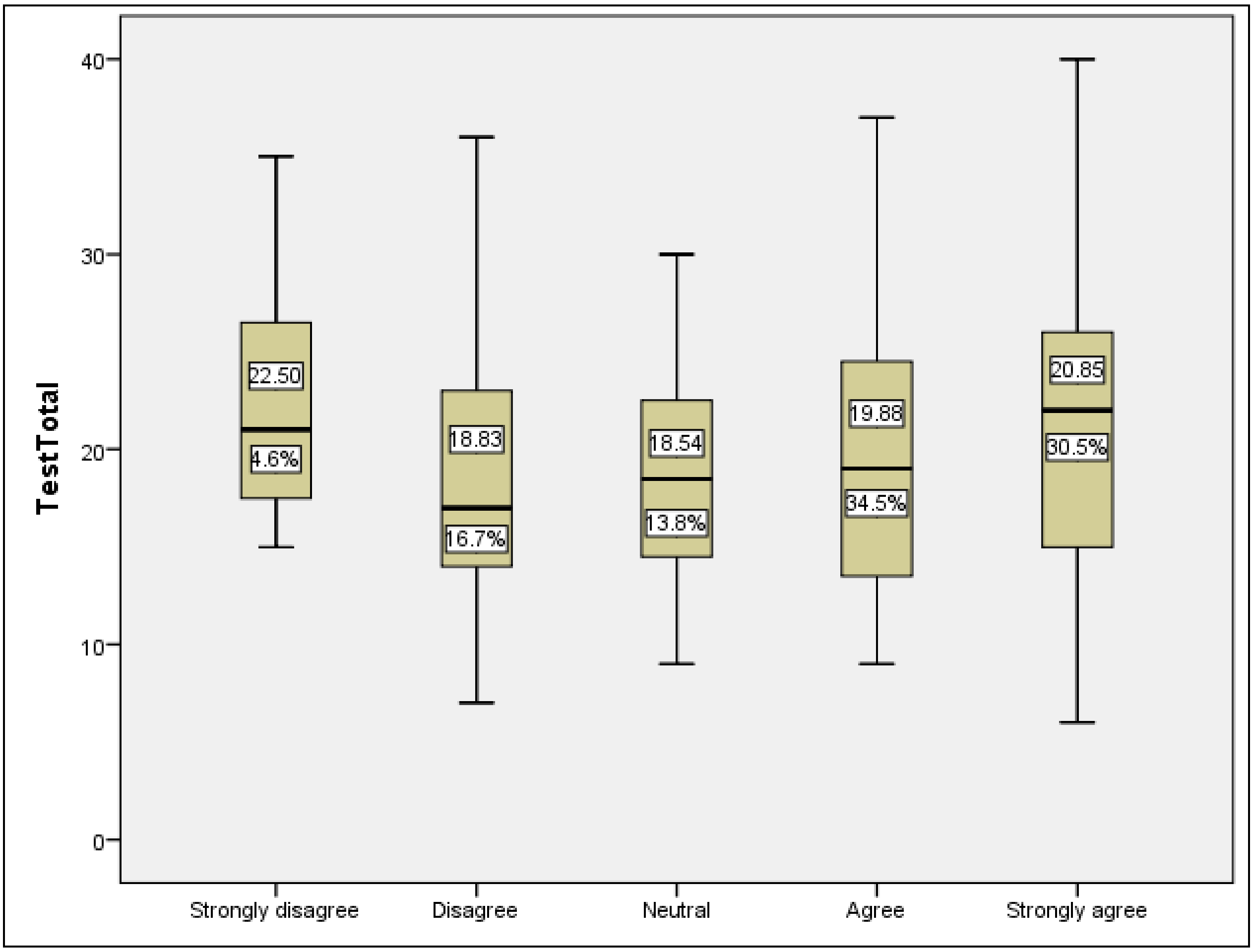


Figure Q22. Sound quality: Agreement (X Axis) and mean test total scores (Y Axis).

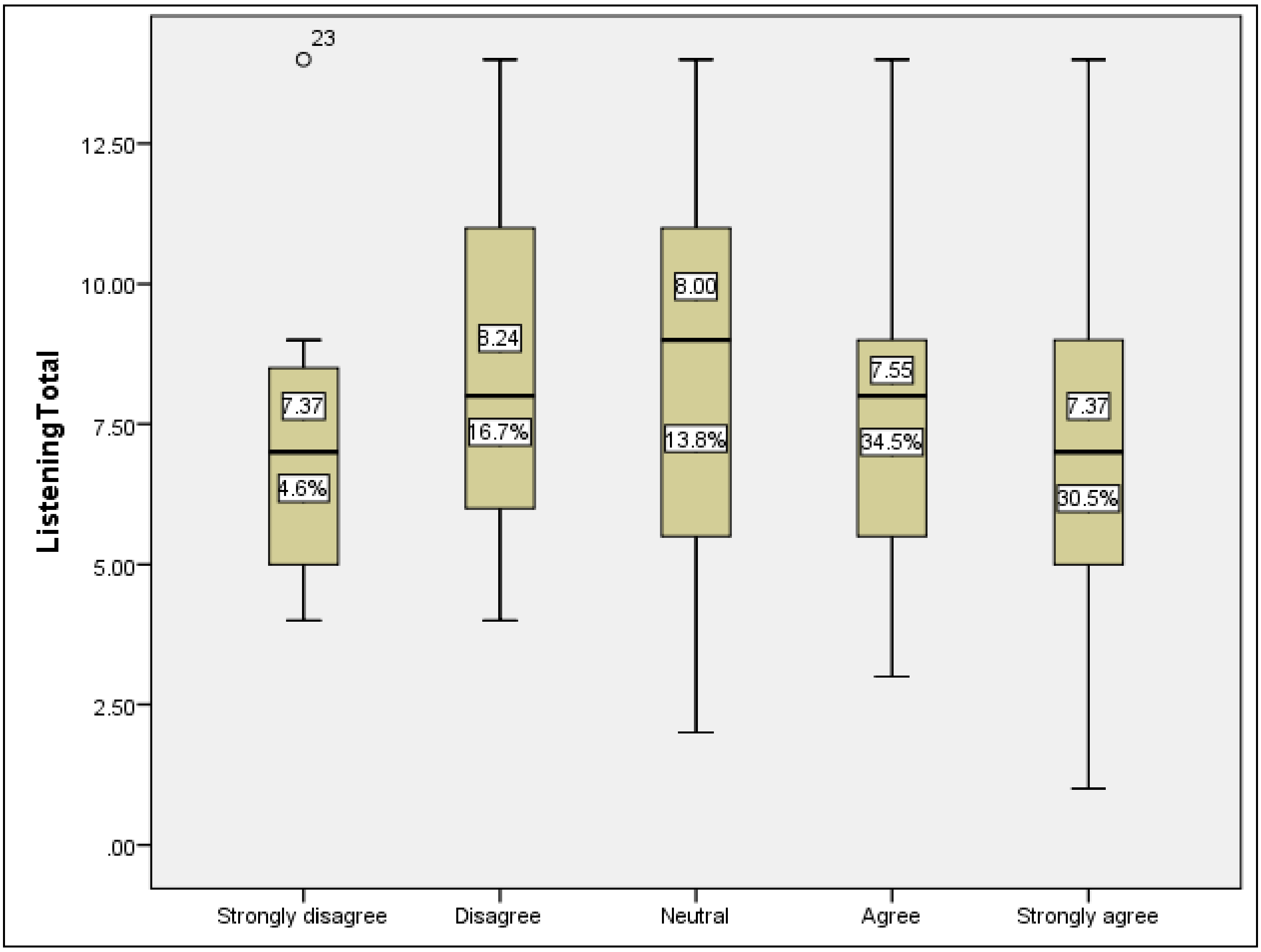


Figure Q23. Sound quality: Agreement (X Axis) and mean listening test scores (Y Axis).

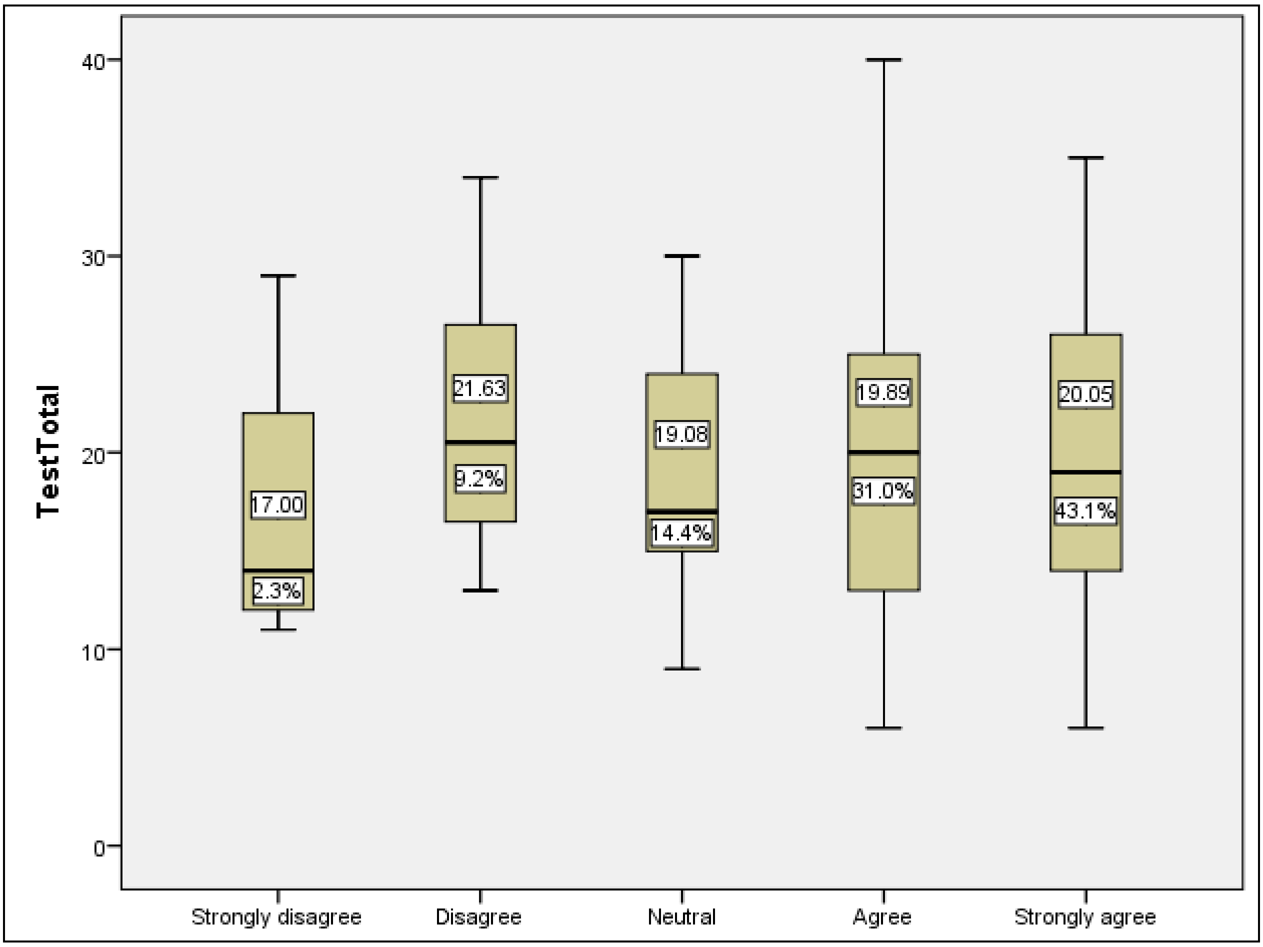


Figure Q24. Headphones quality: Agreement (X Axis) and mean test total scores (Y Axis).

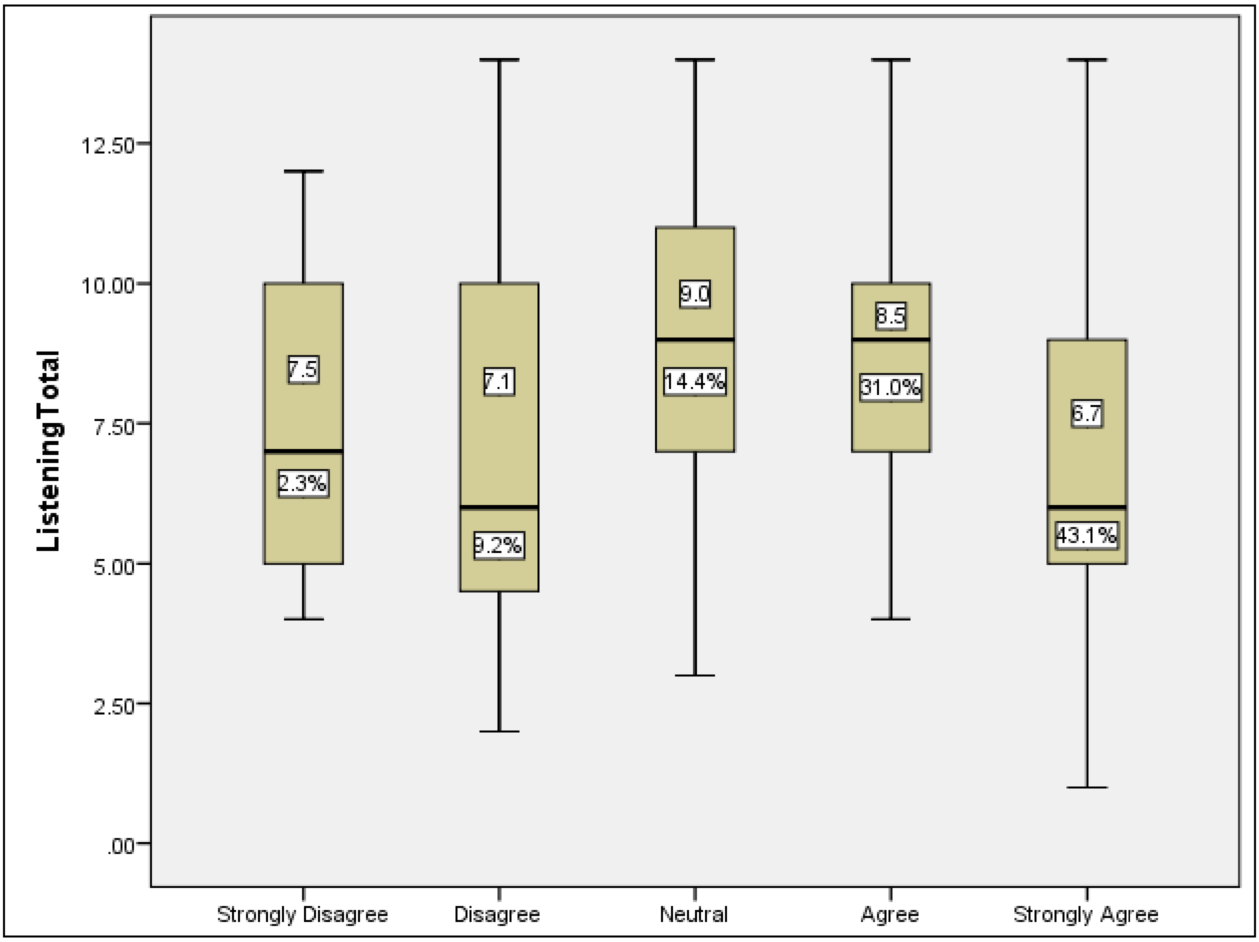


Figure Q25. Headphones quality: Agreement (X Axis) and mean listening test scores (Y Axis).

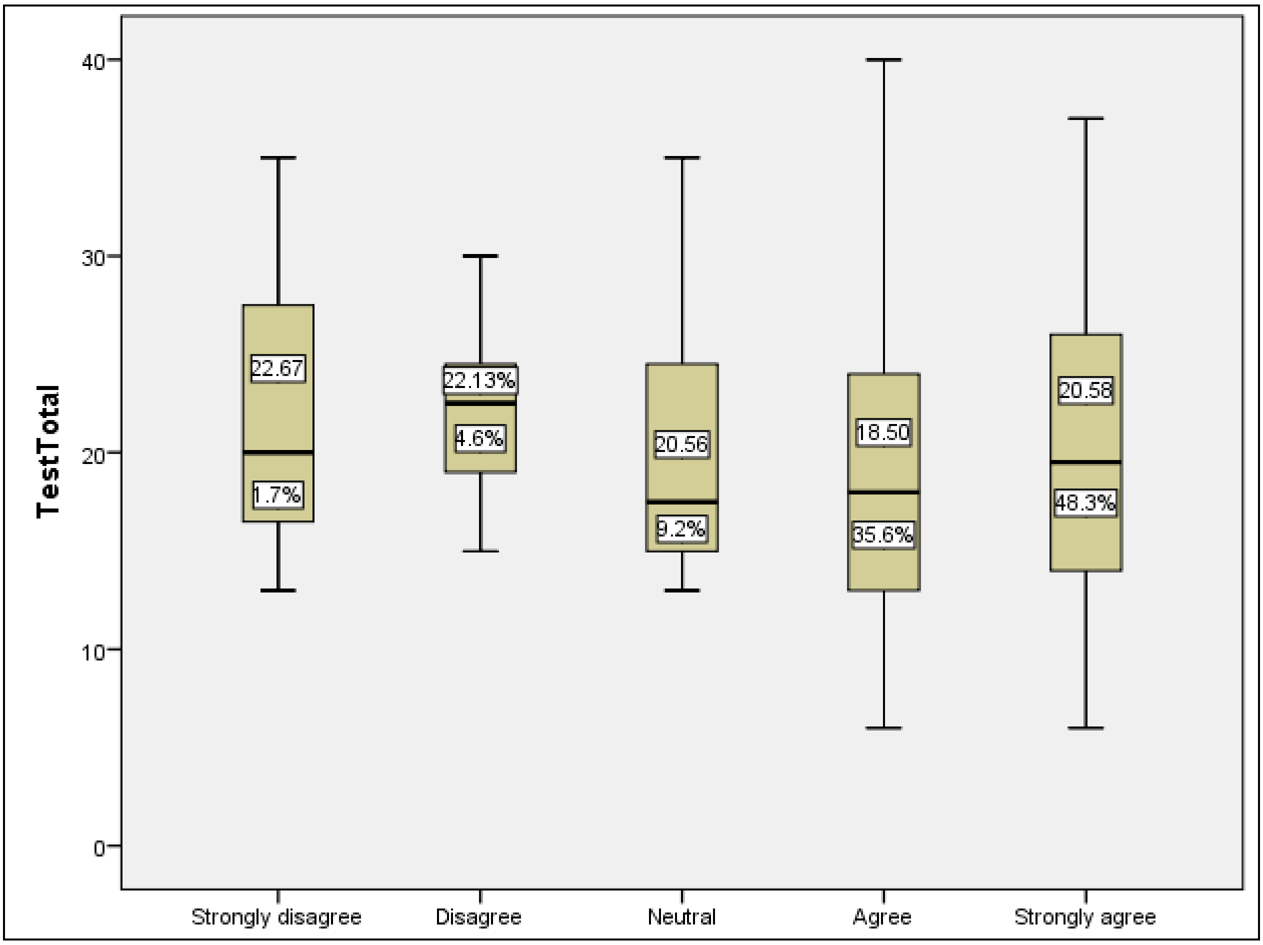


Figure Q26. Liked the split screen mode for reading tests: Agreement (X Axis) and mean test total scores (Y Axis).

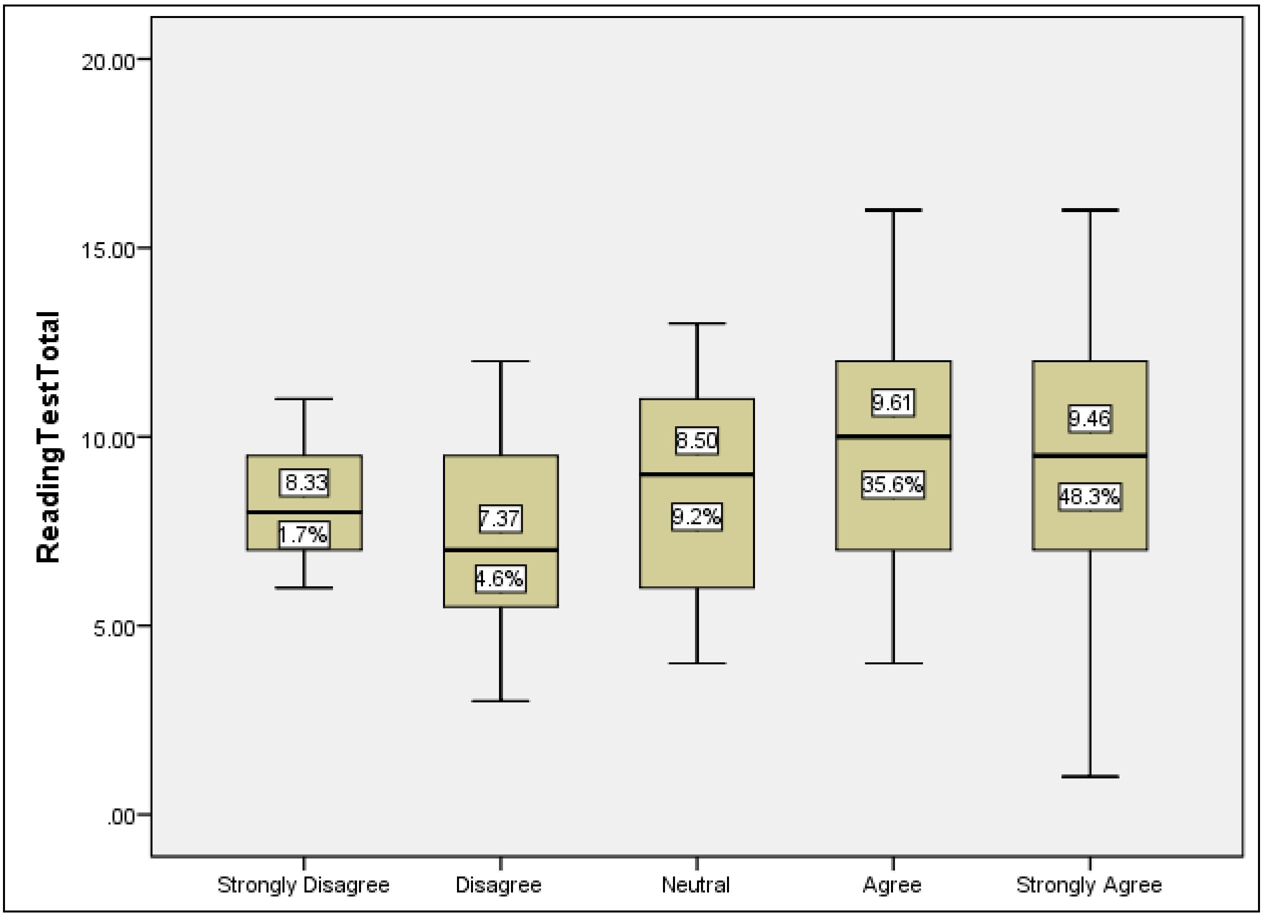


Figure Q27. Liked the split screen mode for reading tests: Agreement (X Axis) and mean reading test scores (Y Axis).

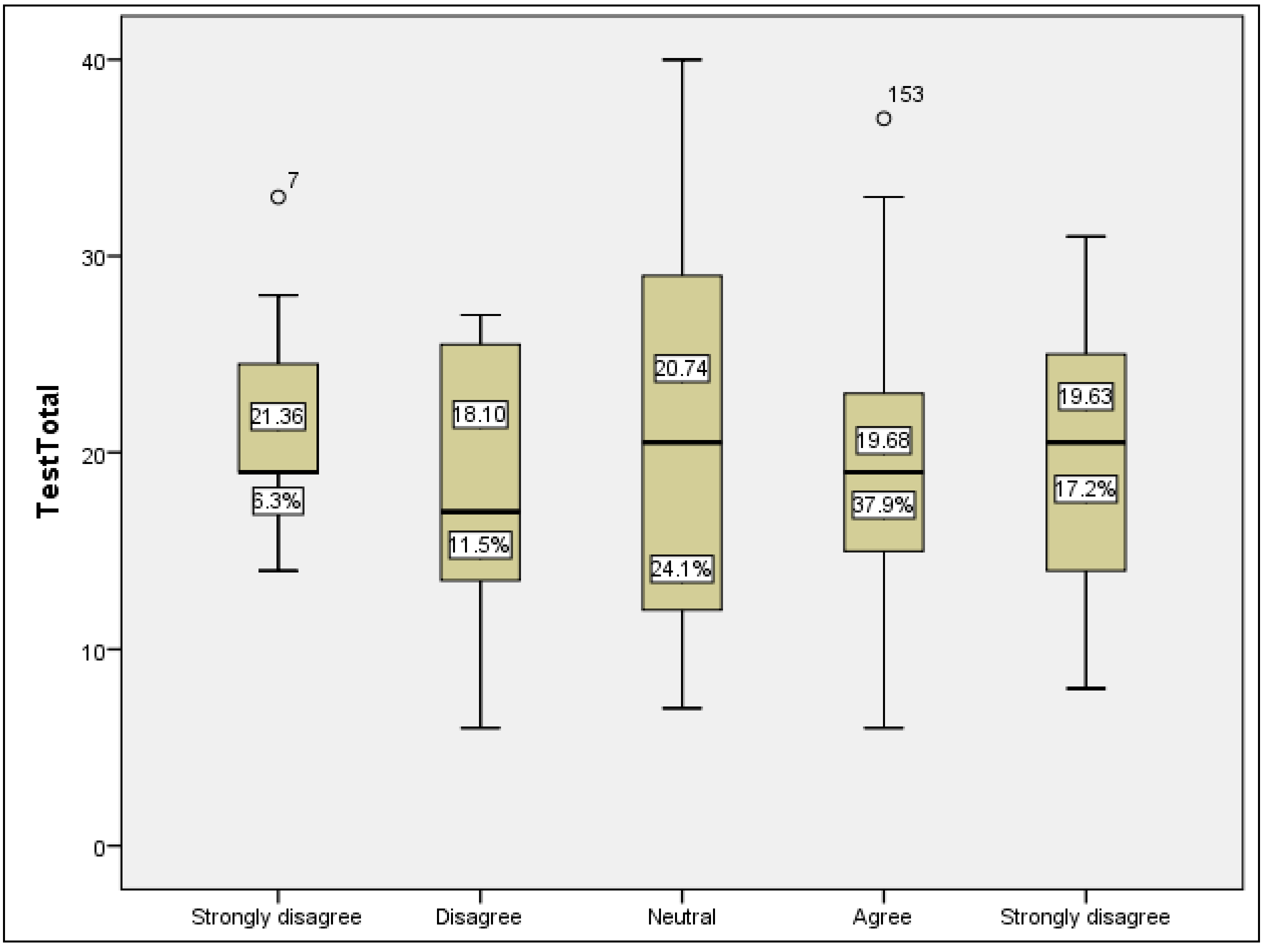


Figure Q28. Needed to take notes: Agreement (X Axis) and mean test scores (Y Axis).

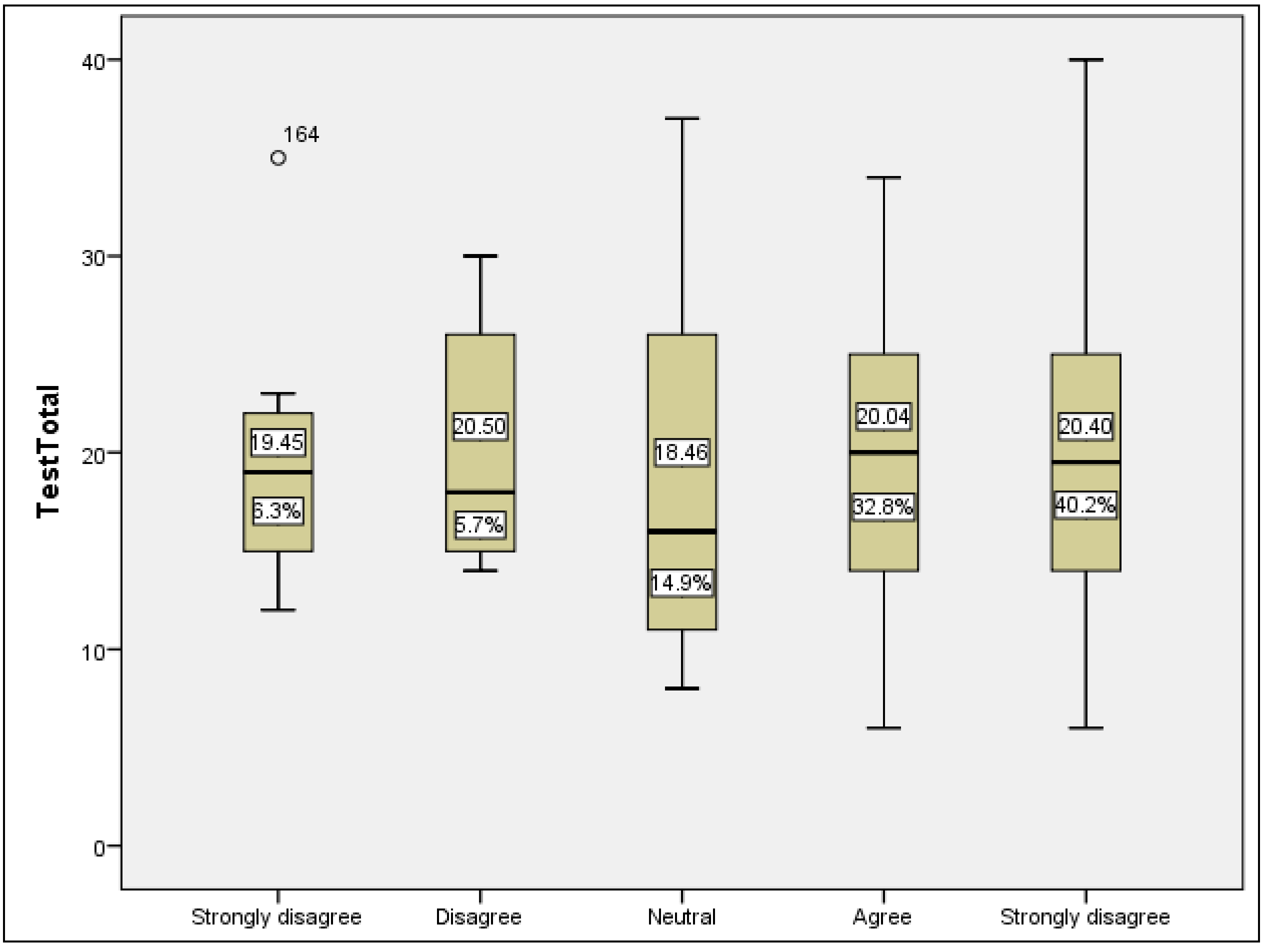


Figure Q29. Liked Moodle feedback: Agreement (X Axis) and mean test scores (Y Axis).

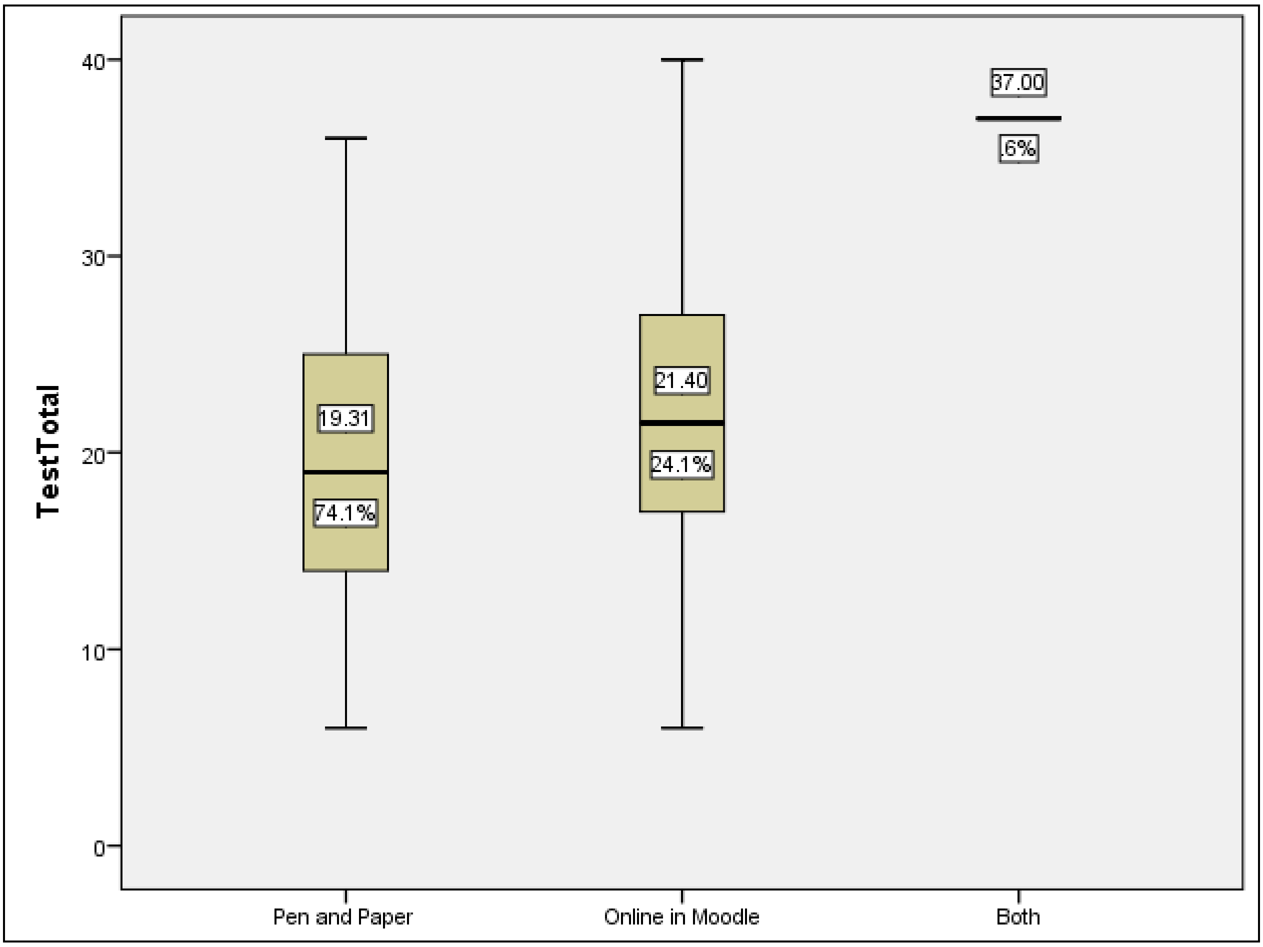


Figure Q30. Testing format preference: Agreement (X Axis) and mean test scores (Y Axis).

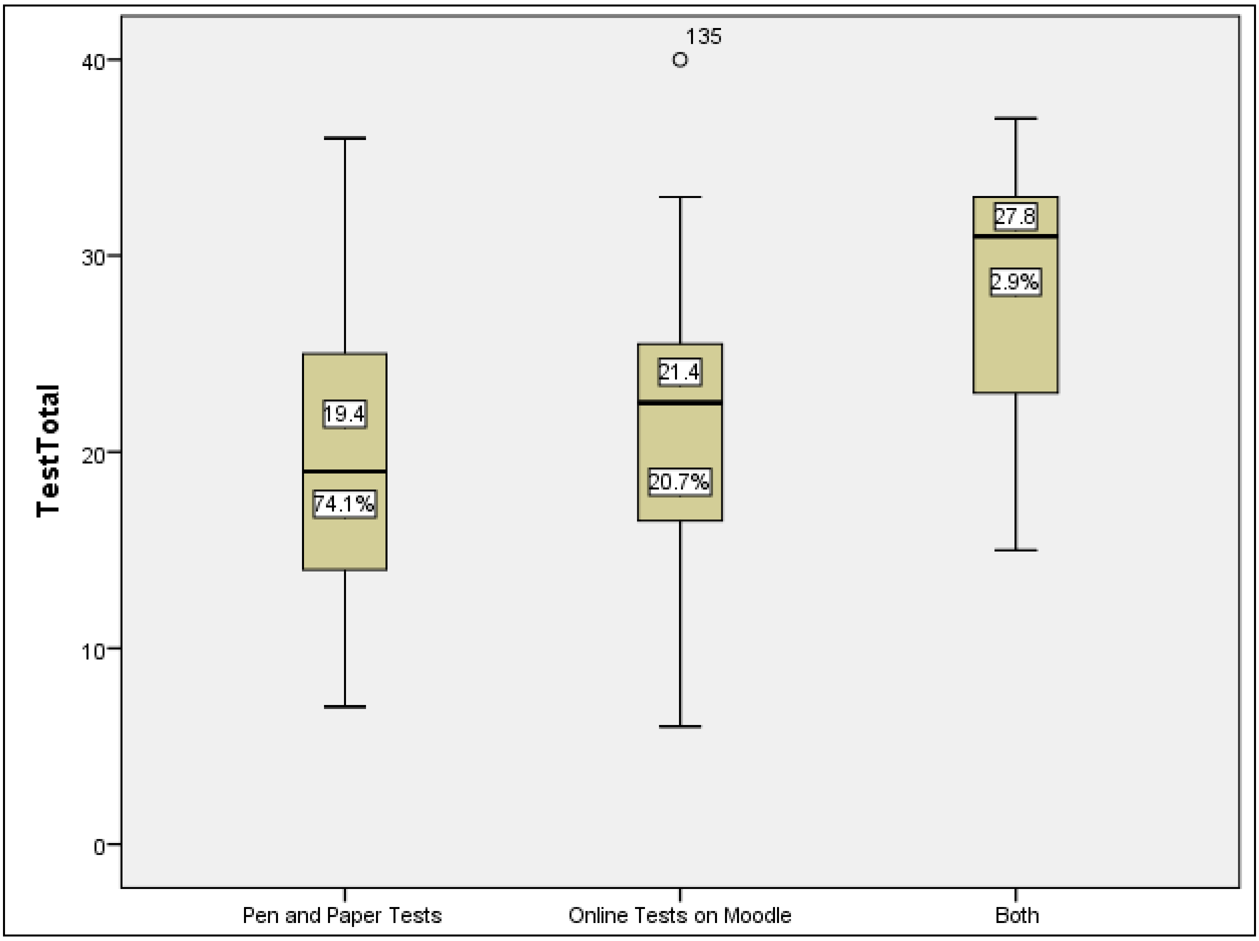


Figure Q31. I would perform best when using: Test format (X Axis) and mean test scores (Y Axis).

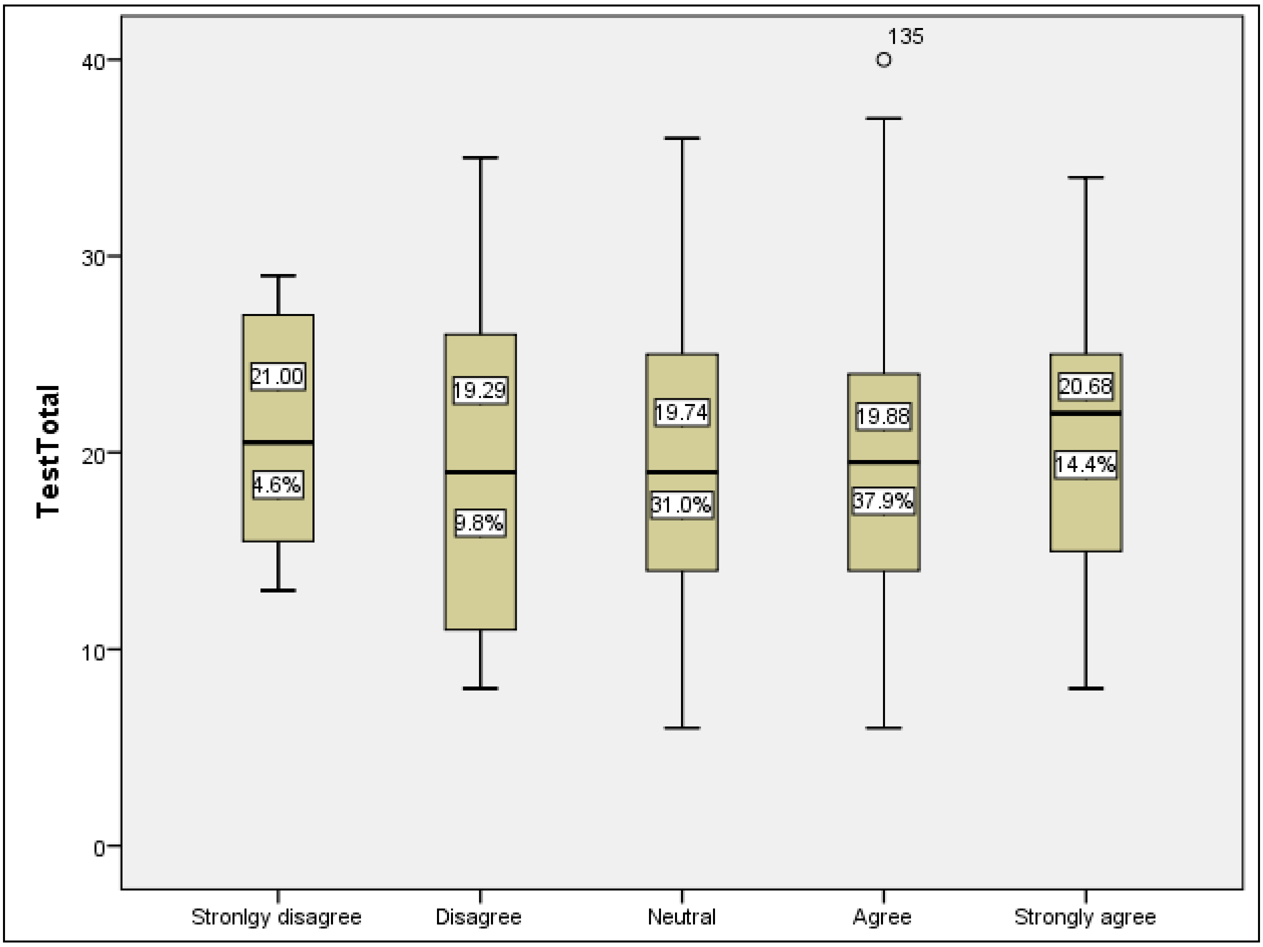


Figure Q32. Liked typing responses: Agreement (X Axis) and mean test total scores (Y Axis).

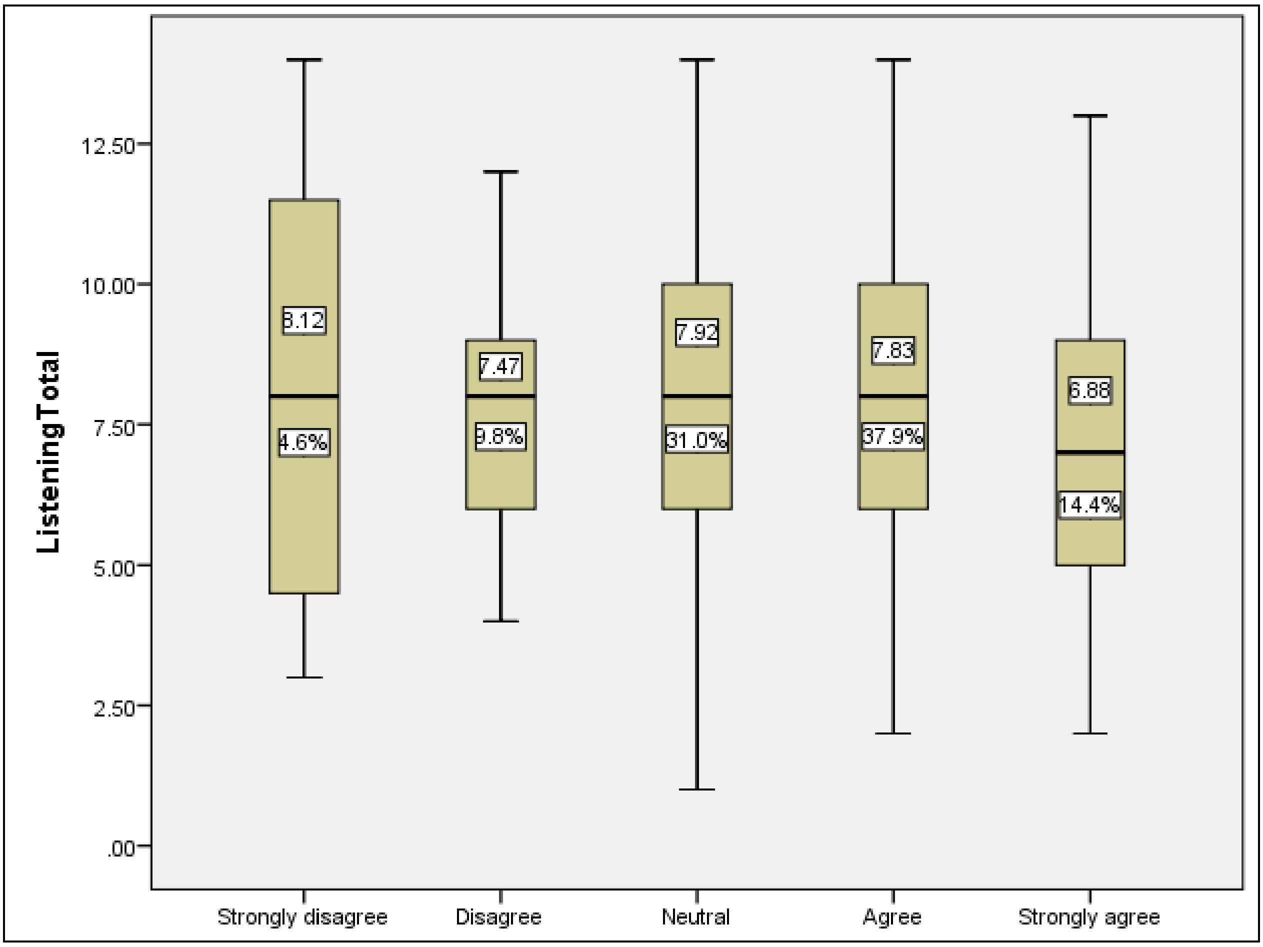


Figure Q33. Liked typing responses: Agreement (X Axis) and mean listening test scores (Y Axis).

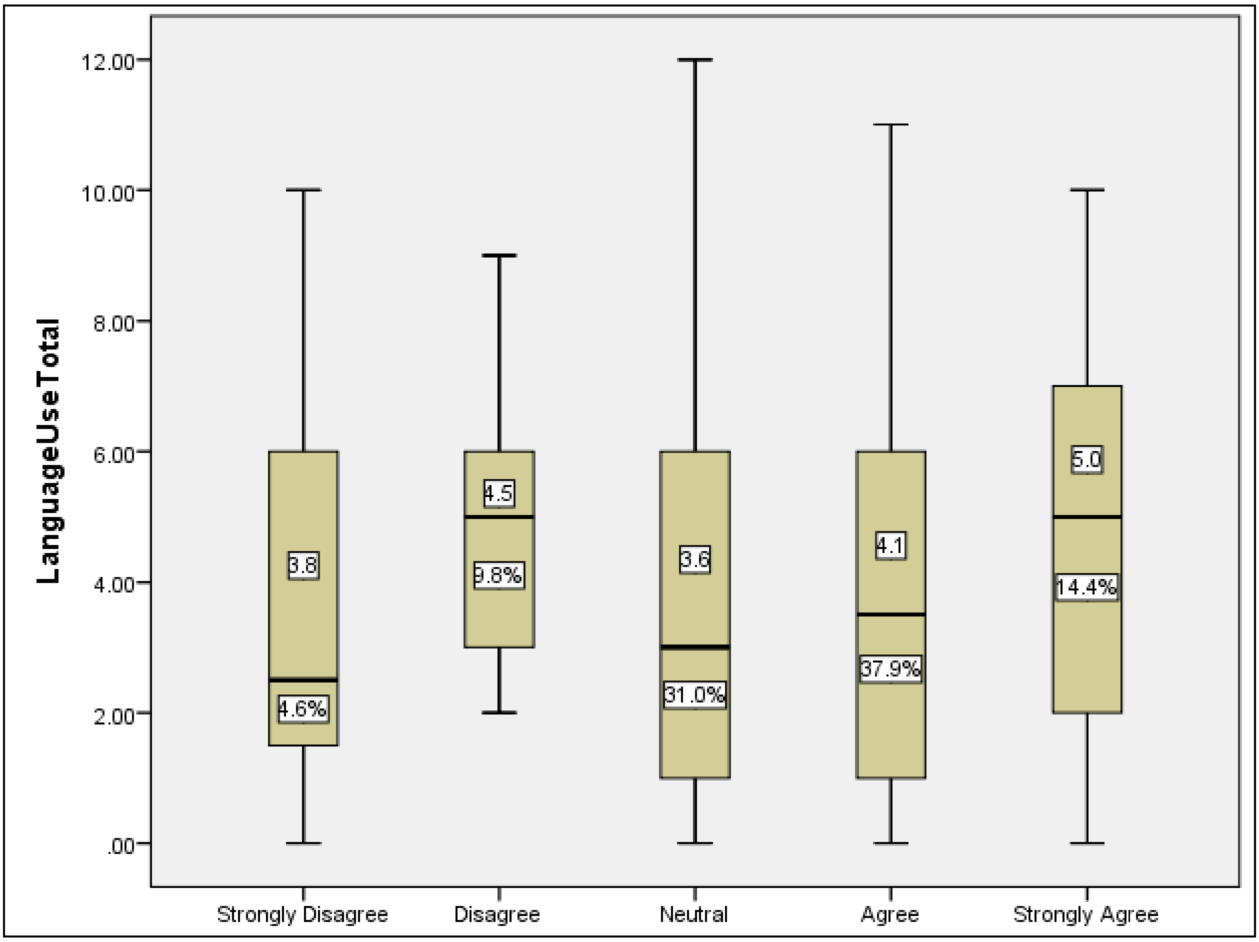


Figure Q34. Typing responses: Agreement (X Axis) and mean test scores (Y Axis).

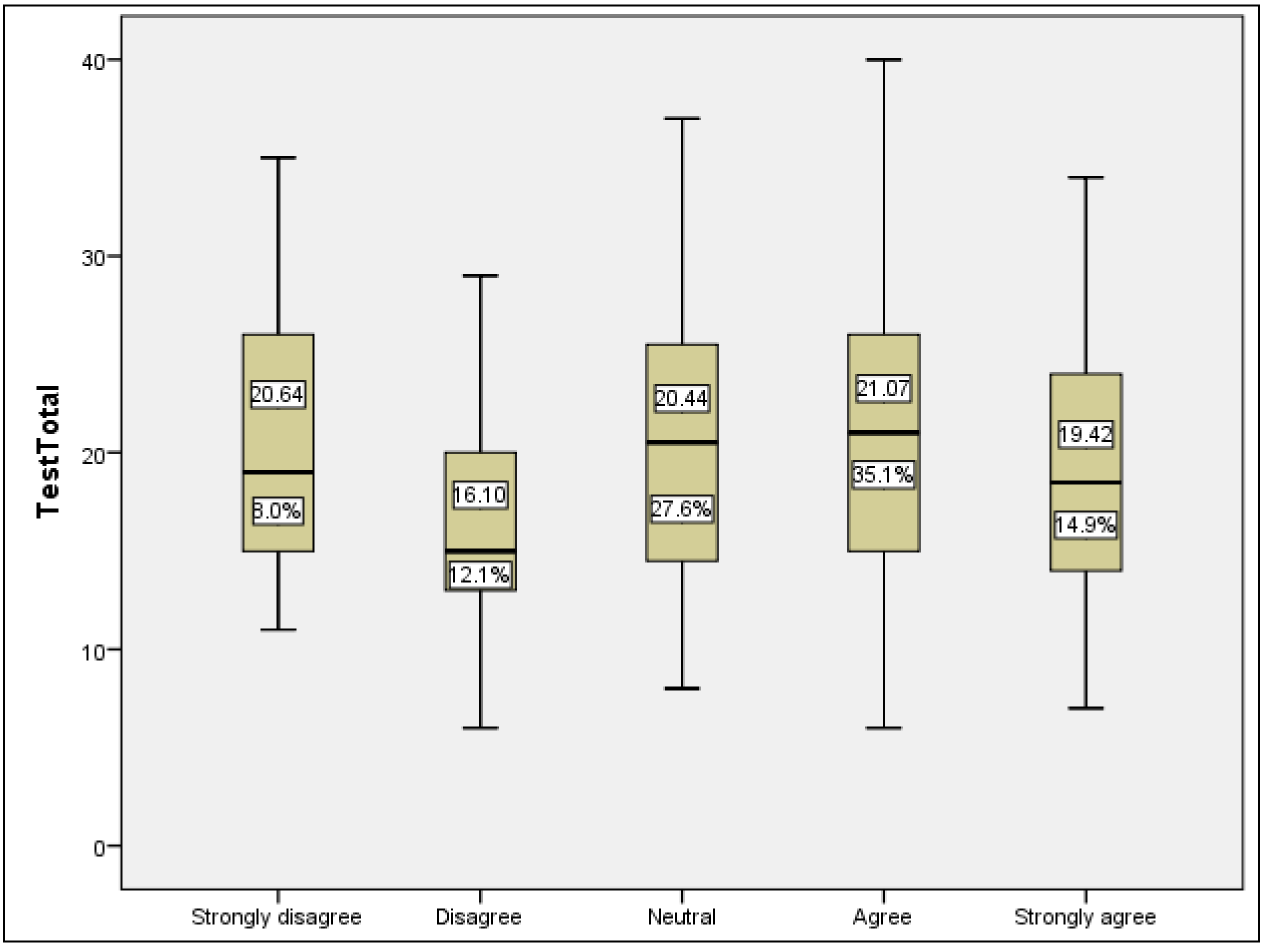


Figure Q35. Test reflecting true language ability: Agreement (X Axis) and mean test scores (Y Axis).

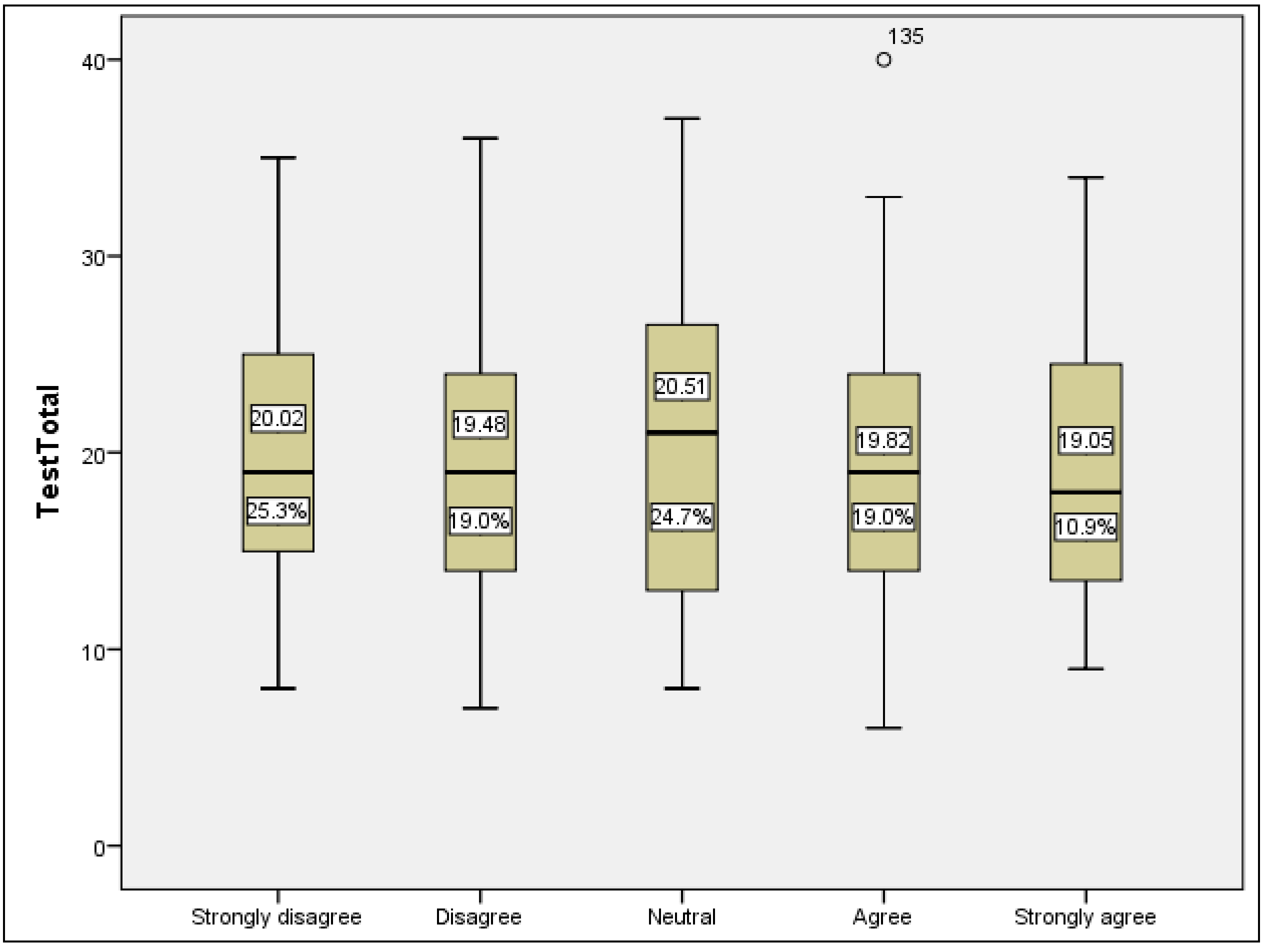


Figure Q36. Would like to take Moodle tests as official exams (Likert): Agreement (X Axis) and mean test scores (Y Axis).

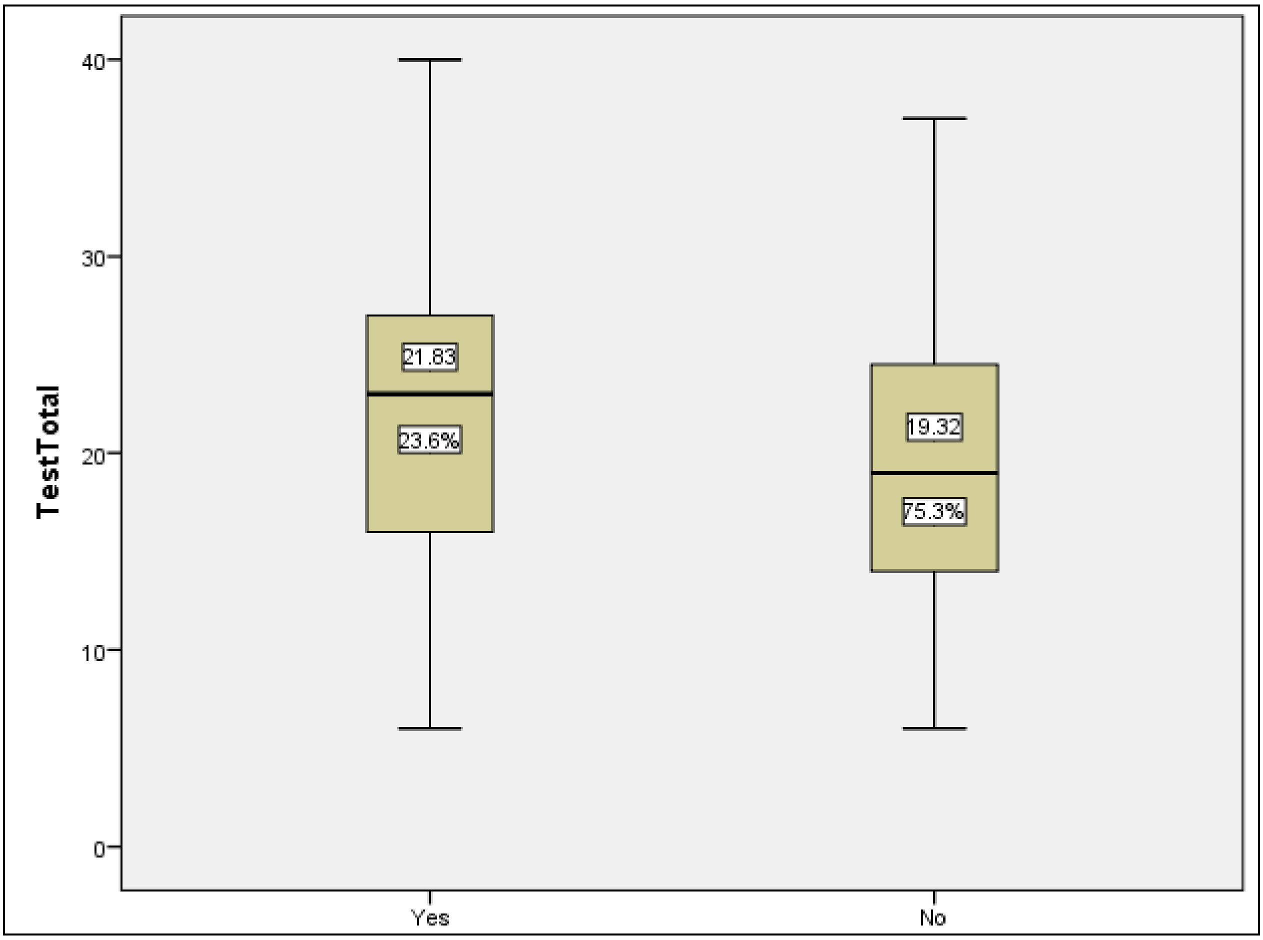


Figure Q37. Would you like to take official exams (like mid-terms, finals, placement tests, exit tests, and so forth) on Moodle to take decisions about the level of your language proficiency? (Yes/No): Agreement (X Axis) and mean test scores (Y Axis).

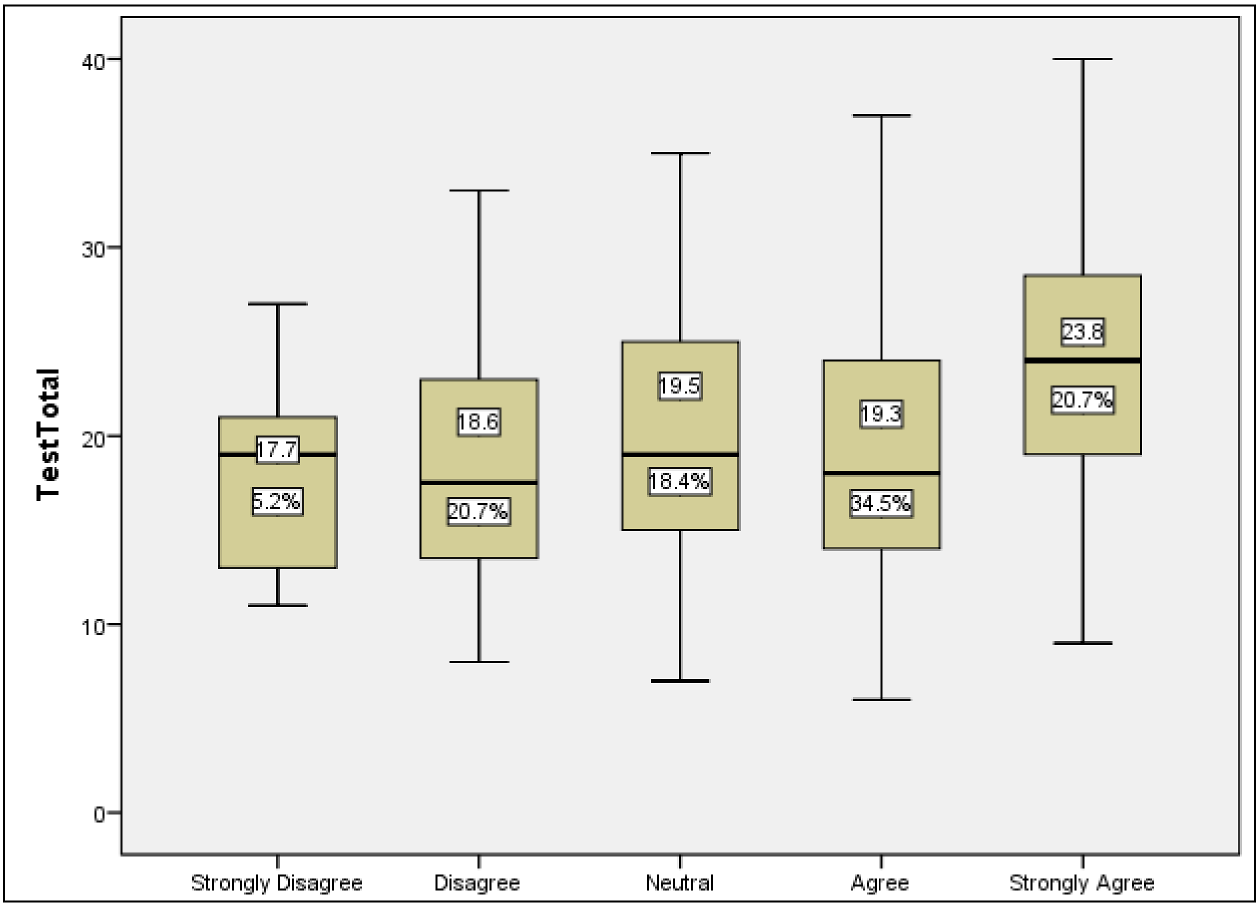


Figure Q38. Sufficiency of test timing: Agreement (X Axis) and mean test scores (Y Axis).

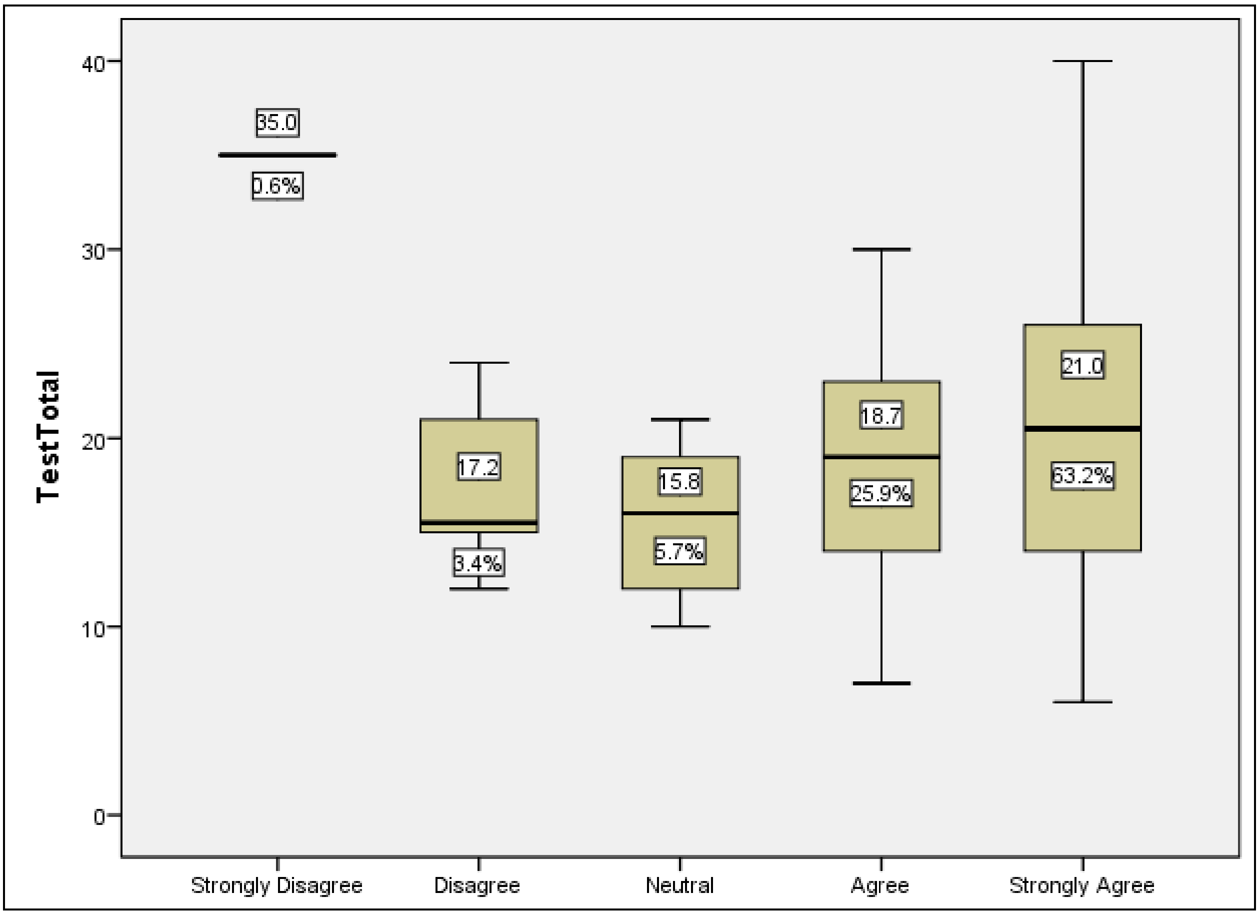


Figure Q39. Count-down timer: Agreement (X Axis) and mean test scores (Y Axis).

Appendix R: Detailed tables of questionnaire analysis results

Table R1. *Kruskal-Wallis Test: Test Length: Test Scores and Agreement*

Q28: The test was too long as it consisted of too many sections	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	4	14	46	59	50
Test score	26.5	21.8	19.5	19.1	20.3
SD	6.5	7.7	7.7	6.5	7.7

Notes. Not significant, $H(5, n = 174) = 4.99$, $p = .417$.

Table R2. *Kruskal-Wallis Test: Concentration Loss: Test Scores and Agreement*

Q29: Staring at the computer screen for a long period of time made me lose my concentration.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	7	14	42	54	57
Test score	21.6	18.9	19.3	21.5	19.0
SD	7.6	6.8	6.7	7.7	7.3

Notes. Not significant, $H(4, n = 174) = 3.83$, $p = .430$).

Table R3. *Kruskal-Wallis Test: Eye Fatigue: Test Scores and Agreement*

Q30: Staring at the computer screen for a long period of time caused me eye fatigue.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	4	8	33	67	67
Test score	25.3	17.0	19.1	20.5	19.9
SD	7.7	6.3	5.9	8.0	7.4

Notes. Not significant, $H(5, n = 174) = 3.56$, $p = .614$.

Table R4. *Kruskal-Wallis Test: Ease of Test Navigation: Test Score and Agreement*

Q6: Overall, the test was easy to navigate by moving from one page displaying a subtest to another.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	3	5	10	53	101
Test score	17.3	16.4	17.6	20.3	20.3
SD	3.8	5.0	6.4	7.8	7.3

Notes. Not significant, $H(5, n = 174) = 3.761$, $p = .584$.

Table R5. *Kruskal-Wallis Test: Appropriateness of Background Colour: Test Score and Agreement*

Q9: I think the background theme (colours) of the test was appropriate	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	2	5	23	76	63
Test score	17.5	15.8	17.9	20.7	20.4
SD	3.5	4.2	6.3	7.7	7.4

Notes. Not significant, *H*(5, *n* = 174) = 4.720, p = .451.

Table R6. *Kruskal-Wallis Test: Clarity of Procedures and Instructions: Test Score and Agreement*

Q13: Test procedures and instructions given were clear and easy to follow	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	1	4	23	63	80
Test score	16.0	23.3	17.3	19.5	20.9
SD	NA	5.7	7.4	7.2	7.4

Notes. Not significant: *H*(5, *n* = 174) = 5.833, p = .323.

Table R7. *Kruskal-Wallis Test: Ease of Test Login: Test Score and Agreement*

Q25: I was able to successfully log onto Moodle and the online test	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	1	3	7	58	105
Test score	12.0	17.7	18.7	19.3	20.5
SD	NA	6.4	2.3	7.1	7.6

Notes. Not significant: *H*(4, *n* = 174) = 2.65, p = .618.

Table R8. *Kruskal-Wallis Test: Clarity of Pictures and Graphs: Test Score and Agreement*

Q26: Pictures and graphs were clear	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	0	8	30	58	78
Test score	NA	17.0	18.1	20.9	20.3
SD	NA	4.5	6.5	6.8	8.0

Notes. Not significant, *H*(3, *n* = 174) = 4.35, p = .226.

Table R9. *Kruskal-Wallis Test: Inappropriate Font Size: Test Score and Agreement*

Q27: The font size was NOT appropriate	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	52	53	32	28	9
Test score	21.0	19.6	19.9	19.4	18.1
SD	7.0	6.6	8.4	8.0	6.9

Notes. Not significant, *H*(4, *n* = 174) = 1.96, p = .743.

Table R10. *Kruskal-Wallis Test: Familiarity with Moodle Tests: Test Score and Agreement*

Q3: Your level of familiarity with tests or quizzes on Moodle: (Very familiar; Somehow familiar; A little bit familiar; Not familiar at all)	Very familiar	Somehow familiar	A little bit familiar	Not familiar at all
N	68	75	27	4
Test score	21.7	19.4	17.5	16.8
SD	7.5	6.9	6.8	10.0

Notes. Significant, *H*(3, *n* = 174) = 7.899, *p* = .048, *r* = 0.05;
Not significant in post hoc comparisons.

Table R11. *Kruskal-Wallis Test: Familiarity with Computers: Test Score and Agreement*

Q4: Your level of computer-literacy or familiarity with computers: (Very familiar; Somehow familiar; A little bit familiar; Not familiar at all)	*Very familiar	Somehow familiar	*A little bit familiar	Not familiar at all
N	46	103	25	0
Test score	22.1	19.6	17.2	NA
SD	7.6	7.0	6.8	NA

Notes. Significant, *H*(2, *n* = 174) = 7.58, *p* = .023, *r* = 0.04;
*Post hoc pairwise comparisons, *p* = .020, *r* = 0.49.

Table R12. *Kruskal-Wallis Test: Enough Technology Experience: Test Score and Agreement*

Q32: I have enough experience with technology to take tests on Moodle	Strongly disagree	Disagree	*Neutral	Agree	*Strongly agree
N	4	13	31	79	45
Test score	17.0	21.9	15.9	19.7	23.0
SD	4.4	7.2	6.8	6.2	8.2

Notes. Significant, *H*(5, *n* = 174) = 18.80, *p* = .002; *r* = 0.11.
*Post hoc pairwise comparisons, *p* = .001; *r* = 0.62.

Table R13. *Kruskal-Wallis Test: Need Extra Technical Training: Test Score and Agreement*

Q33: I will need extra technical training before I am ready to take online exams.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	25	36	32	50	30
Test score	21.7	21.2	20.0	18.5	19.4
SD	7.5	7.5	7.8	7.1	6.7

Notes. Not significant, *H*(5, *n* = 174) = 4.44, *p* = .487.

Table R14. *Kruskal-Wallis Test: Liked Test-Taking Experience: Test Score and Agreement*

Q5: Overall, I liked this test-taking experience	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	4	13	49	79	28
Test score	16.8	17.2	19.6	19.8	22.6
SD	12.3	5.8	7.1	7.3	7.3

Notes. Not significant, *H*(5, *n* = 174) = 7.06, *p* = .216.

Table R15. *Kruskal-Wallis Test: Liked Using New Technology: Test Score and Agreement*

Q15: I liked using new technology to take this test.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	10	19	32	63	45
Test score	19.9	21.6	20.8	18.4	21.1
SD	8.0	5.6	7.8	6.9	7.9

Notes. Not significant, *H*(5, *n* = 174) = 5.82, *p* = .324.

Table R16. *Kruskal-Wallis Test: Liked Taking Moodle Test: Test Score and Agreement*

Q34: Did you like taking the test on Moodle? (Yes/No)	Yes	No
N	87	85
Test score	21.2	18.6
SD	7.5	6.9

Notes. Not significant, *H*(2, *n* = 174) = 5.93, *p* = .052.

Table R17. *Kruskal-Wallis Test: Technical Problems: Test Score and Agreement*

Q20: There were technical problems during the exam.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	50	68	23	21	11
Test score	20.8	19.8	19.2	19.7	18.8
SD	8.4	7.1	5.8	7.5	6.6

Notes. Not significant, *H*(5, *n* = 174) = 1.49, *p* = .914.

Table R18. *Kruskal-Wallis Test: Network Efficiency: Test Score and Agreement*

Q21: The network was efficient and did not slow down while taking the test.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	3	9	14	70	78
Test score	15.7	20.4	19.4	19.8	20.3
SD	9.1	9.3	8.5	7.0	7.2

Notes. Not significant, *H*(4, *n* = 174) = 1.67, *p* = .796.

Table R19. *Kruskal-Wallis Test: Speed of Audio File Loading: Test Score and Agreement*

Q22: The audio file in the listening loaded quickly.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
	Overall test[a]				
N	2	6	15	68	83
Test score	12.0	15.3	18.3	19.7	21.0
SD	4.2	4.1	6.4	7.2	7.5
	Listening test[b]				
N	2	6	15	68	83
Test score	8.5	7.3	8.0	8.3	7.1
SD	6.4	3.4	3.1	3.0	2.6

Notes. [a]Not significant, *H*(4, *n* = 174) = 8.14, *p* = .087.
[b]Not significant, *H*(4, *n* = 174) = 6.50, *p* = .165.

Table R20. *Kruskal-Wallis Test: Computer Working Properly During the Exam: Test Score and Agreement*

Q23: The computer worked properly during the exam.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	0	1	10	59	103
Test score	NA	26.00	18.4	20.4	19.9
SD	NA	NA	7.1	7.5	7.2

Notes. Not significant, *H*(4, *n* = 174) = 3.09, *p* = .544.

Table R21. *Kruskal-Wallis Test: Sound Quality: Test Score and Agreement*

Q11: Sound quality of the listening tests was good.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
	Overall test[a]				
N	8	29	24	60	53
Test score	22.5	18.8	18.5	19.9	20.9
SD	6.8	6.9	6.0	7.4	8.0
	Listening test[b]				
N	8	29	24	60	53
Test score	7.4	8.2	8.0	7.6	7.4
SD	3.2	2.7	3.4	2.5	3.1

Notes. [a]Not significant, *H*(4, *n* = 174) = 3.48, *p* = .482.
[b]Not significant, *H*(4, *n* = 174) = 2.72, *p* = .606.

Table R22. *Kruskal-Wallis Test: Headphones Quality: Test Score and Agreement*

Q24: The headphones worked properly during the exam.	Strongly disagree	Disagree	Neutral[c]	Agree[d]	Strongly agree[cd]
	Overall test[a]				
N	4	16	25	54	75
Test score	17.0	21.6	19.1	19.9	20.1
SD	8.2	6.2	5.7	7.7	7.7
	Listening test[b]				
N	4	16	25	54	75
Test score	7.5	7.1	9.0	8.5	6.7
SD	3.4	3.4	2.8	2.6	2.7

Notes. [a]Not significant, $H(4, n = 174) = 2.09$, $p = .720$.
[b]Significant, $H(4, n = 174) = 19.01$, $p = .001$, $r = 0.11$.
[c]Significant in post hoc comparisons, $p = .006$, $r = 0.40$.
[d]Significant in post hoc comparisons, $p = .005$, $r = 0.31$.

Table R23. *Kruskal-Wallis Test: Split Screen Mode for Reading Tests: Test Score and Agreement*

Q8: I liked the split screen mode for the reading tests where the reading texts were on the left side of the screen and the questions were on the right side.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
	Overall test[a]				
N	3	8	16	62	84
Test score	22.7	22.1	20.6	18.5	20.6
SD	11.2	4.7	7.1	7.2	7.5
	Reading test[b]				
N	3	8	16	62	84
Test score	8.3	7.4	8.5	9.6	9.5
SD	2.5	2.9	3.0	3.1	3.3

Notes. [a]Not significant, $H(5, n = 174) = 4.169$, $p = .525$.
[b]Not significant, $H(5, n = 174) = 6.783$, $p = .237$.

Table R24. *Kruskal-Wallis Test: Needing To Take Notes during the Test: Test Score and Agreement*

Q31: I needed to take notes during the test.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	11	20	42	66	30
Test score	21.4	18.1	20.7	19.7	19.6
SD	5.7	6.3	9.2	6.7	6.5

Notes. Not significant, $H(5, n = 174) = 2.22$, $p = .818$.

Table R25. *Kruskal-Wallis Test: Moodle Instant Feedback: Test Score and Agreement*

Q12: I liked that Moodle showed me instant feedback/test results at the end of the test.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	11	10	26	57	70
Test score	19.5	20.5	18.5	20.0	20.4
SD	6.4	6.4	8.2	7.7	6.9

Notes. Not significant, *H*(4, *n* = 174) = 1.85, *p* = .763.

Table R26. *Kruskal-Wallis Test: Testing Format Preference: Test Score and Agreement*

Q18: Which format of testing do you prefer? a) pen and paper b) online in Moodle	Pen and paper	Online in Moodle	Neutral
N	129	42	1
Test score	19.3	21.4	37.0
SD	7.2	7.1	NA

Notes. Not significant, *H*(3, *n* = 174) = 5.98, *p* = .113.

Table R27. *Kruskal-Wallis Test: Which Testing Format Students Would Perform Best on: Test Score and Agreement*

Q19: I think I would perform best when using: a) pen and paper tests. b) online tests on Moodle.	Pen and paper	Online in Moodle	Neutral
N	129	36	5
Test score	19.4	21.4	27.8
SD	7.1	7.2	8.8

Notes. Significant, *H*(3, *n* = 174) = 8.30, *p* = .040, *r* = 0.05;
Not significant in post hoc comparisons.

Table R28. *Kruskal-Wallis Test: Typing Responses: Test Score and Agreement*

Q14: I liked typing my responses for some questions.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
	Overall test[a]				
N	8	17	54	66	25
Test score	21.0	19.3	19.7	19.9	20.7
SD	6.2	9.1	7.3	7.4	6.9
	Listening test[b]				
N	8	17	54	66	25
Test score	8.1	7.5	7.9	7.8	6.9
SD	4.3	2.1	2.8	3.0	2.9
	Language use test[c]				
N	8	17	54	66	25
Test score	3.8	4.5	3.6	4.1	5.0
SD	3.4	1.9	3.4	2.9	2.8

Notes. [a]Not significant, *H*(5, *n* = 174) = .955, *p* = .966.
[b]Not significant, *H*(5, *n* = 174) = 4.08, *p* = .538.
[c]Significant, *H*(5, *n* = 174) = 13.53, *p* = .019, *r* = 0.08;
Not significant in post hoc comparisons.

Table R29. *Kruskal-Wallis Test: Test Reflecting True Language Ability: Test Score and Agreement*

Q16: I think that the test reflected my true language ability.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	14	21	48	61	26
Test score	20.6	16.1	20.4	21.1	19.4
SD	7.1	6.0	7.2	7.7	7.5

Notes. Not significant, *H*(5, *n* = 174) = 7.87, *p* = .164.

Table R30. *Kruskal-Wallis Test: Would Like to Take Moodle Official Exams: Test Score and Agreement*

Q17: I would like to take such online tests on Moodle as official exams (e.g. mid-terms, finals, Placement Test, Exit Test).	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	44	33	43	33	19
Test score	20.0	19.5	20.5	19.8	19.1
SD	6.6	7.3	7.8	7.9	7.1

Notes. Not significant, *H*(5, *n* = 174) = 1.17, *p* = .948.

Table R31. *Kruskal-Wallis Test: Would Like to Take Moodle Official Exams: Test Score and Agreement*

Q35: Would you like to take official exams (like mid-terms, finals, placement tests, exit tests, and so forth) on Moodle to take decisions about the level of your language proficiency? (Yes/No)	Yes	No
N	41	131
Test score	21.8	19.3
SD	7.3	7.3

Notes. Not significant, $H(2, n = 174) = 4.358$, $p = .113$.

Table R32. *Kruskal-Wallis Test: Sufficiency of Test Timing: Test Score and Agreement*

Q7: Test timing was sufficient for all test sections.	Strongly disagree	Disagree[a]	Neutral	Agree[b]	Strongly agree[ab]
N	9	36	32	60	36
Test score	17.7	18.6	19.5	19.3	23.8
SD	5.3	6.4	7.1	7.4	7.4

Notes. Significant, $H(5, n = 174) = 15.61$, $p = .008$, $r = 0.10$.

[a]Post hoc pairwise comparisons, $p = .036$, $r = 0.51$.

[b]Post hoc pairwise comparisons, $p = .048$, $r = 0.44$.

Table R33. *Kruskal-Wallis Test: Count-down Timer: Test Score and Agreement*

Q10: I liked the presence of the count-down timer to help me submit my answers to the test questions within the given test time.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
N	1	6	10	45	110
Test score	35.0	17.2	15.8	18.7	21.0
SD	NA	4.5	4.1	6.5	7.6

Notes. Significant, $H(5, n = 174) = 11.83$, $p = .037$, $r = 0.07$;
Not significant in post hoc comparisons.