

A RASCH APPROACH TO VALIDATION OF
CEFR-ALIGNED READING TESTS FOR TESTING
UNDERGRADUATE READING COMPREHENSION

BY

MOHAMED ISMAIL FOUZUL KAREEMA

A thesis submitted in fulfilment of the requirement for the
degree of Doctor of Philosophy in Education

Kulliyyah of Education
International Islamic University Malaysia

MARCH 2023

ABSTRACT

In the ESL context, reading is an important skill necessary for academic success. Similarly, reading tests commonly are conducted in order to find out the students' ability in comprehending texts so that appropriate teaching and learning instructions are provided to enhance the skill. Applying the latest developments in testing reading and test validation, this study focused on three important objectives. The first was to produce valid and reliable instruments to measure the academic reading comprehension ability of university students in Sri Lanka by adapting the CEFR-aligned tests. The second was to examine the reading ability of students of the four faculties at SEUSL, using these validated instruments. The third objective was to investigate the students' achievement level in the cognitive processes of reading based on Khalifa and Weir's (2009) model of reading. To achieve these three objectives, 13 texts were adapted along with their (127) items from the CEFR-aligned LRN materials, and four testlets were produced. Eight cognitive processes of reading, namely Word Recognition (WR), Lexical Access (LA), Syntactic Parsing (SP), Establishing Prepositional Meaning (EPM), Inferencing (I), Building a Mental Model (BMM), Creating Text Level Structure (CTLS), and Creating Inter-Textual Representation (CITR), which are arranged hierarchically, were measured. A single test had 40 selected-response objective items including eleven common items, which had been used as anchoring items to horizontally equate four tests. The concurrent analysis of the Rasch measurement model was used to examine the psychometric properties of the tests. The findings revealed the validity and reliability of the tests and the strength of using the Rasch model for test equating. The findings also discovered that, while there was inconsistency in the hierarchical order of the cognitive processes of reading, there was consistency among the LOT (except for EPM) and the HOT processes, and the items within the same process did not have the same difficulty level, which indicates that certain cognitive processes can be used across different difficulty levels. The results also showed that 843 students, 93.5% out of 902, scored the CEFR B1 and B2 levels, which were identified as the minimum requirement for academic success in the ESL context. In addition, students' reading performance was measured according to their degree programmes with English as a-medium of instruction, and the results showed that students from the FE outperformed their counterparts in FAS, FMC, and FAC in the reading test. The study had several theoretical and practical implications in language testing and validation, and testing reading.

ملخص البحث

تعد القراءة مهارة ضرورية للنجاح الأكاديمي. ولذا يتم إجراء اختبارات القراءة من أجل معرفة قدرة الطلاب على فهم النصوص بحيث يتم توفير التعليمات المناسبة لتعليم تنمية المهارة. لذا ركزت هذه الدراسة على ثلاثة أهداف مهمة مواكبة في ذلك أحدث التطورات في مجال اختبار اللغة وفعاليتها. الهدف الأول هو إنتاج أدوات صالحة وموثوقة لقياس قدرة فهم القراءة لدى طلاب جامعيين سريلنكيين. والهدف الثاني هو فحص القدرة على طلاب جامعة الجنوب الشرقي بسريلنكا من خلال تكييف الاختبارات المتوافقة مع معيار القراءة لطلاب الكليات الأربع باستخدام التكييف، بعد أن تم قياسها بواسطة هذه الأدوات التي تم التحقق من صحتها. وأما الهدف الثالث فقد استكشف مستوى تحصيل الطلاب في العمليات المعرفية للقراءة بناءً على نموذج خليفة ووير (٢٠٠٩) للقراءة. ولتحقيق هذه الأهداف الثلاثة فإن الدراسة قد تبنت ١٣ نصًا مع الأسئلة ١٢٧ مما سمح بإعداد أربعة اختبارات. وقد تم ذلك عن طريق تبني ثمان عمليات للقراءة المعرفية وهي: التعرف على الكلمات، واستخدام المعجم، والتحليل النحوي، وإنشاء معان الجر، والاستدلال، وبناء نموذج عقلي، وإنشاء بنية لمستوى النص، وإنشاء تمثيل لما بين النصوص. وقد تم ترتيب هذه العمليات بشكل هرمي. وقد احتوى كل اختبار على ٤٠ عنصرًا لإجابات عن أسئلة متعددة الخيارات بما في ذلك أحد عشر عنصرًا مشتركًا والتي تم استخدامها كعناصر إرساء لمعادلة أربعة اختبارات أفقيّة. وقد تم استخدام التحليل المتزامن (RASCH) لنموذج قياس فحص الخصائص السيكمومترية للاختبارات. كشفت النتائج عن صحة وموثوقية الاختبارات وقوة استخدامه نموذج التحليل المتزامن (RASCH) لمعادلة الاختبار. كما أسفرت النتائج على أنه وفي ظل وجود تضارب في الترتيب الهرمي للعمليات المعرفية للقراءة، وجد أيضًا اتساق في مستوى التفكير البسيط باستثناء عملية إنشاء معاني الجر ومستوى التفكير الأعلى، ولم يكن للعناصر الموجودة في العملية نفس مستويات الصعوبة، مما يشير إلى أن بعض العمليات المعرفية يمكن استخدامها عبر مستويات مختلفة الصعوبة. كما أظهرت النتائج أيضًا أن ٨٤٣ طالبًا (٩٣,٥٪) من أصل ٩٠٢ طالبًا حصلوا على مستويات في مواد شبكة مصادر التعلم (CEFR) في المستويين (B1) و (B2) واللذين يعتبران الحد الأدنى من متطلبات النجاح الأكاديمي في اللغة الإنجليزية باعتبارها لغة ثانية. بالإضافة إلى ذلك، فقد تم قياس أداء الطلاب في القراءة وفقًا للبرامج الأكاديمية التي يزاولونها والتي تبنت اللغة الإنجليزية كأداة للتعلم. وقد أظهرت النتائج بأن الطلاب الذين ينتمون إلى كلية الهندسة تفوقوا في اختبار القراءة على نظرائهم في كلية العلوم التطبيقية وكذلك كلية الإدارة والتجارة وكلية الآداب والثقافة. كما توصلت الدراسة الحالية إلى مجموعة من النتائج ذات الطابع النظري والعملية المتعلقة بالاختبارات اللغوية والتصديق واختبارات القراءة.

APPROVAL PAGE

The thesis of Mohamed Ismail Fouzul Kareema has has been approved by the following:

Ainol Madziah Zubairi
Supervisor

Noor Lide Abu Kassim
Co-Supervisor

Kamal J I Badrasawi
Co-Supervisor

Mohamad Sahari Nordin
Internal Examiner

Raja Safinas Raja Harun
External Examiner

Noor Mohammad Osmani
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Mohamed Ismail Fouzul Kareema

Signature

Date

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

**DECLARATION OF COPYRIGHT AND AFFIRMATION OF
FAIR USE OF UNPUBLISHED RESEARCH**

**A RASCH APPROACH TO VALIDATION OF
CEFR-ALIGNED READING TESTS FOR TESTING
UNDERGRADUATE READING COMPREHENSION**

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

Copyright © 2023 Mohamed Ismail Fouzul Kareema and International Islamic University Malaysia.
All rights reserved.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may only be used by others in their writing with due acknowledgement.
2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.
3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Mohamed Ismail Fouzul Kareema

.....
Signature

.....
Date

This thesis is dedicated to my loving parents for the foundation of what I turned out to be in life.

To my dedicated husband MCM. Sathif

And to my wonderful children: Abdullah, Fathih Sadhaf, and Fikrath Sadhafa

May this serve as an inspiration to all of you!

ACKNOWLEDGEMENTS

First of all, all glory and thanks belong to Allah, **أَلْحَمْدُ لِلَّهِ**. I would like to express my everlasting sincere gratitude to my supervisor, Prof. Dr. Ainol Madziah Zubairi for all her highly appreciated guidance, insightful comments, tolerance, and constant inspiration. My heartfelt appreciation goes to my supervisory committee members: Prof. Dr. Noor Lide Abu Kassim and Assoc. Prof. Dr. Kamal Jamil Badrasawi. They contributed immensely with their constructive comments, prominent support, and continuous encouragement that helped me immeasurably in shaping this research.

I am obliged to thank the expert panel, particularly: AP. Dr. Nor Liza Ali, AP. Dr. Ting Su Hie, AP. Dr. Adlina Ariffin, Dr. Nicola Latimer, Dr. Zailani Jusoh, Dr. Rifa Mahroof and others for their irreplaceable services and training sessions. My special thanks go to Prof. Mike Linacre for his clarifications on the Rasch. I am indebted to the copyright permission of the LRN to utilize their materials, without which I could not have carried out my study smoothly.

I am thankful to the Ministry of Higher Education Sri Lanka for its AHEAD project, which granted me a scholarship to pursue my study. As well, I wish to recognize the support of the South Eastern University of Sri Lanka for granting me study leave. Both academic and administrative staff members of the SEUSL immensely supported my study. My deepest gratitude and respect go to the HoD of the DELT: Dr AMM. Navaz, for facilitating all research work and my data collection processes. I highly appreciate my colleagues: Mr Sameem, Mr Abdul Rahuman, Mr Al-Ihsan, Ms Firzan, Ms Hoorul, Ms Shakira, Ms Rifka, Ms Hanan, and all others at the DELT for their commitment to all of the minutiae, big and small, is which at the heart of the study. I am appreciative of Prof. MAM. Fazil, Dr Rifa Mahroof, Dr Aslam Saja, Ms Rashidha, and Mr Riswan for their moral and academic support. I must not forget the assistance of the SEUSL respondents in the two stages of data collection.

I would like to express my sincere thanks to the IIUM, its lecturers, and the staff of KOED, who were extremely helpful and supportive. Especially, I am thankful to all my fellow PhD friends for the continuous support, prayers, sharing, caring, and for all the fun we have had.

I would like to extend my deepest gratitude to my family: my husband, MCM. Sathif and my loving children: Abdullah, Fathih and Fikrath. You have been put to the ultimate test of patience, your dedication has lasted the longest, but your faith has also been tested, and your support has been crucial. I am grateful to my parents, Al-Hajjah ALM Ismail and AL. Saboora Beevi for their blessings, love and care throughout my life. My special thanks go to my late father-in-law, Al-Haj ALM Cassim, whose encouraging words remain a blessing in my heart. I will never forget my mother-in-law's prayers for my prosperity. As well, I am grateful to my sisters, brothers, uncle, brothers and sisters-in-law, all other family members, and friends for their love, prayers, and support.

Finally, I express my gratitude to other unnamed individuals, who assisted in the success of this work. May Allah SWT bless you all. Thank you all.

TABLE OF CONTENTS

Abstract	ii
Abstract in Arabic	iii
Approval Page.....	iv
Declaration	v
Copyright Page.....	vi
Dedication	vi
Acknowledgements	viii
List of Tables	xiv
List of Figures	xvi
CHAPTER ONE: INTRODUCTION	1
1.1 Introduction.....	1
1.2 Background of the Study	1
1.2.1 Reading in the Second Language	3
1.2.2 Influence of Reading for Academic Success	4
1.2.3 Reading Skill for English Medium Instruction (EMI)	5
1.2.4 Assessing Reading	6
1.2.5 CEFR.....	7
1.2.6 CEFR Level for reading to achieve academic success.....	8
1.2.7 Current trend of Reading assessment in Sri Lanka	9
1.2.8 Reading from Islamic Perspectives	11
1.3 Problem Statement.....	11
1.3.1 Research Objectives	16
1.3.2 Research Questions	16
1.4 Rationale for the Study	17
1.5 Significance of the Study.....	17
1.6 Limitations of the Study	18
1.7 Operational Definition of Terms	19
1.8 Organization of the Study.....	21
CHAPTER TWO: LITERATURE REVIEW.....	22
2.1 Introduction.....	22
2.2 English Language Teaching (ELT) in Sri Lanka.....	22
2.2.1 The History of English in Sri Lanka	22
2.2.1.1 English Language Teaching in Colonial Sri Lanka	23
2.2.1.2 Present ELT Trend in Sri Lanka	24
2.2.1.2.1 ELT at School Level	24
2.2.1.2.2 ELT at University Level	25
2.2.2 Issues in ELT in Sri Lanka.....	27
2.2.3 English Language Testing and Evaluation in Sri Lanka.....	28
2.2.3.1 Previous Research in Language Testing in Sri Lanka	31
2.2.3.1.1 Research Studies on Washback.....	31
2.2.3.1.2 Research on Test Validation	32
2.2.4 Reading Ability of Undergraduates in Sri Lanka.....	32
2.3 Theoretical Framework.....	34

2.3.1 Reading	34
2.3.1.1 Reading Comprehension Theories	35
2.3.1.2 Reading Comprehension Models	36
2.3.1.2.1 Reading as a Process	37
2.3.1.2.2 Reading as a Product	38
2.3.1.3 The Nature of Reading	39
2.3.1.3.1 Reading as a Unitary Skill	39
2.3.1.3.2 Reading as a Multidimensional Skill	40
2.3.1.4 Reading in the Second language	42
2.3.1.5 Levels of Reading Comprehension	43
2.3.1.6 Reading skills, sub-skills and strategies	44
2.3.1.6.1 Davis's (1968) Taxonomy	45
2.3.1.6.2 Munby's (1978) Taxonomy	45
2.3.1.6.3 Lunzer's et al. (1979) Taxonomy	47
2.3.1.6.4 Hillock's (1980) Taxonomy	47
2.3.1.6.5 Grabe's (1991) Taxonomy	48
2.3.1.7 Reading Construct	49
2.3.1.7.1 Khalifa and Weir (2009)	51
2.3.1.7.2 Robinson's (1941) SQ3R method	51
2.3.1.8 Academic Reading constructs	52
2.3.2 Assessing Reading	53
2.3.2.1 Text type /Genre/ Purpose	53
2.3.2.2 Test format / Response Type / Type of input/ Item format	55
2.3.2.3 Reading Assessment Scales	60
2.3.2.3.1 ACTFL	60
2.3.2.3.2 TOEFL	61
2.3.2.3.3 DIALANG	62
2.3.2.3.4 DELTA	63
2.3.2.3.5 CEFR Scale of Measurement for reading	64
2.3.2.4 Test Purposes	67
2.3.3 Validation	70
2.3.3.1 Socio-Cognitive Model for Language Test Development and Validation	71
2.3.3.1.1 Context Validation	73
2.3.3.1.2 Cognitive Validation	73
2.3.3.1.3 Scoring Validation	80
2.3.3.1.4 Criterion Validation	80
2.3.3.1.5 Consequential Validation	81
2.4 Measurement Procedure	81
2.4.1 Underlying Principles in Measurement Processes	82
2.4.1.1 Classical Test Theory (CTT) and its Limitations	82
2.4.1.2 Item Response Theory (IRT) and Rasch Measurement Model (RMM)	84
2.4.1.2.1 Characteristics of Rasch Measurement Model	85
2.4.1.3 Conceptual Framework	86
2.4.1.3.1 Limitations of the Previous Method and the proposal of the current method	88

2.4.1.3.2 Application of RMM in validation studies and Socio-cognitive validation framework for Reading	89
2.5 Summary of the Chapter	90
CHAPTER THREE: RESEARCH METHODOLOGY	91
3.1 Introduction.....	91
3.2 Research Design	91
3.3 Research Procedure	92
3.4 Population and Sampling of the Study	94
3.4.1 Sampling Procedure and the Characteristics of the Respondents	94
3.4.2 Sample Size.....	96
3.5 Instrument of the Study	98
3.5.1 Test Development and Adaptation.....	98
3.5.1.1 Selection of the LRN Texts for Item Adaptation.....	99
3.5.1.2 Categorizing Cognitive Processes of Reading	103
3.5.1.3 Test Review	105
3.5.2 Empirical Evaluation.....	105
3.5.2.1 Preliminary Investigation.....	106
3.5.2.1.1 Test Validation.....	106
3.5.2.1.2 Text Inspector Analysis for Readability	119
3.5.2.2 Pilot Study.....	121
3.5.2.2.1 Data Analysis Procedure for Pilot Study	122
3.5.2.3 Modification of the Four Reading Tests Based on Piloting and Experts' Feedback.....	137
3.5.3 Instrument for Final Study	137
3.6 Data Collection	140
3.7 Ethical Considerations	141
3.8 Analysis of Data	141
3.8.1 Rasch Measurement Model Analysis for the Final Instrument.....	142
3.8.2 SPSS Analysis.....	143
3.9 Summary of the Chapter	144
CHAPTER FOUR: RESULTS OF THE STUDY.....	145
4.1 Introduction.....	145
4.2 Preliminary Analysis of the Main Data	146
4.2.1 Screening and Cleaning of Data.....	146
4.2.2 Validity of Reading Tests.....	147
4.2.2.1 Validity of Test Items	148
4.2.2.1.1 Item Fit.....	148
4.2.2.1.2 Item Polarity.....	151
4.2.2.1.3 Unidimensionality of the Items.....	153
4.2.2.2 Construct Validity	154
4.2.2.2.1 Continuum of Increasing Intensity.....	155
4.2.2.2.2 Empirical Scaling of Reading Test	157
4.2.3 The Precision and Reliability of Measurement.....	160
4.2.3.1 Reliability and Separation.....	160
4.2.3.2 Precision of Measures	162
4.2.3.3 Test Targeting	163

4.2.4 Validity of Common Item Linking	163
4.2.5 Validity of Individual Tests	166
4.2.6 Validity of Students' Responses	168
4.2.7 Summary of Acceptability of Reading Tests	171
4.3 Students' Reading Performance Aligned with CEFR Level	171
4.3.1 CEFR Levels of the Tests	171
4.3.2 Grading Scheme of Tests	174
4.3.3 Students' Performance Level	175
4.3.4 Students' Performance Level according to Faculty Background.....	178
4.4 Cognitive Processing in Reading.....	181
4.4.1 Cognitive Processes Achieved by Many Students	183
4.4.2 Cognitive Processes Underachieved by Many Students	187
4.5 Summary of the Key Findings	193
4.6 Summary of the Chapter	195

CHAPTER FIVE: DISCUSSION, RECOMMENDATIONS, AND CONCLUSION 196

5.1 Introduction.....	196
5.2 Overview of the Study	196
5.3 Summary and Discussion of the Findings	197
5.3.1 The Processes of Test Adaptation	197
5.3.2 Validity and Adequacy of the Reading Tests.....	199
5.3.3 Validity of Examinee Responses	200
5.3.4 Construct Definition.....	201
5.3.5 Test Equating Procedures and Validity of Common Item Linking ..	203
5.3.6 Student's Reading Performance Aligned with CEFR Level.....	208
5.3.7 Cognitive Processing and Academic Reading	211
5.4 Implications	215
5.4.1 Theoretical Implications.....	215
5.4.2 Methodical Implications.....	217
5.4.3 Practical Implications.....	219
5.5 Limitations of the Study and Pointers for Further Research	222
5.6 Recommendations.....	224
5.7 Conclusion	226

REFERENCES..... 229

APPENDIX A: FOUR READING TEST PAPERS.....	269
APPENDIX B: LETTER FOR EXPERT ASSISTANCE AND A SAMPLE OF RATER INFORMTION SHEET	302
APPENDIX C I: ITEM OBJECTIVE CONGRUENCE SHEET (TEST 1 AS A SAMPLE).....	304
APPENDIX C II: SAMPLE (TEST 1) DATA AND INDICES OF ITEM OBJECTIVE CONGRUENCE.....	308
APPENDIX D: SAMPLES OF RATED ITEM OBJECTIVE CONGRUENCE SHEETS (RATED BY PROF.TING).....	314
APPENDIX E: APPROVAL LETTER TO COLLECT DATA.....	320
APPENDIX F: PILOT STUDY DATA MATRIX: FIT STATISTICS FOR PILOT STUDY.....	321

APPENDIX G: STATISTICS FOR FINAL DATA.....	331
APPENDIX H: LRN COPYRIGHT PERMISSION	339

LIST OF TABLES

<u>Table No.</u>		<u>Page No.</u>
2.1	CEFR - Overall Reading Comprehension	65
2.2	Componential Matrix	75
2.3	Cognitive Processing at A2 to C2 in Khalifa and Weir's (2009) examples of Cambridge ESOL Main Suite Reading papers	79
3.1	Sample-Size Range for Calibration (Linacre, 2020b)	97
3.2	Overview of Test 1	101
3.3	Overview of Test 2	101
3.4	Overview of Test 3	102
3.5	Overview of Test 4	102
3.6	Cognitive Processing in Reading in Khalifa and Weir (2009)	104
3.7	Descriptions of the SMEs	110
3.8	Sample of IOC Indices for The First Three Common Items of the Tests	113
3.9	Summary of Cognitive Processes of Reading of Each Test According to IOC Indices	115
3.10	Common Items in all Four Tests	119
3.11	Readability Index according to Text Inspector Analysis	120
3.12	Item - Person Reliability of 11 Common Items	124
3.13	Principal Component Analysis of Standardised Residual Variance For Common Items	125
3.14	Item Statistics for Common Items	126
3.15	Summary of Person and Item Reliability of Four Tests of Pilot Study	127
3.16	The PCA of Standardised Residuals for all Four Tests	128
3.17	Person Statistics: Misfit Order	129

3.18	Item and Person Reliability	133
3.19	Dimensionality Map of Concurrent Analysis of All Four Tests	133
3.20	Fit Statistics for Concurrent Analysis	134
3.21	Summary of the Final Instrument	138
3.22	Summary of Mean Score of Individual Tests	139
4.1	Item Fit Statistics – Misfit Order	149
4.2	Summary Table of Frequency of Item Fit within 0.7- 1.3 infit and outfit MNSQ Range	150
4.3	Item Polarity Statistics: Measure Order (Reading Test)	152
4.4	PCA of Standardized Residuals of all Items	153
4.5	Summary of Cognitive Processing of Reading in Each Test Based on Expert Judgment	158
4.6	Reliability of 127 Measured Items	161
4.7	Reliability Indices of Person and Item for the Common Item Calibration	164
4.8	Item Fit Indices for Common Items	164
4.9	Summary of Reliability Indices of all Four Tests	167
4.10	Summary of Fit Statistics and PCA Residuals of all Four Tests	167
4.11	Summary of Person Fit Statistics	168
4.12	Readability Indices of the Selected Passages	172
4.13	Grading system of IELCA Academic Reading Test	175
4.14	Summary of Test Scores in four Tests according to CEFR levels	176
4.15	Descriptive Statistics of Students' Performance	177
4.16	Reading Performance of the Four Faculties	178
4.17	Descriptive Statistics for Cognitive Processing	183
4.18	Ascending order of Item logit measures of Cognitive Processing	187
4.19	Fit Statistics of 127 Individual Items	190
4.20	Summary of the Key Findings	194

LIST OF FIGURES

<u>Figure No.</u>		<u>Page No.</u>
2.1	CEFR Level Illustrative Descriptors (Adopted from Figure 4 of CEFR for Languages: Learning, Teaching, Assessment (Council of Europe, 2001, p.33))	66
2.2	Socio-cognitive Framework for Test Development and Validation (Adopted from Weir's (2005), p. 44)	72
2.3	Khalifa and Weir's (2009) Model of Reading (Adopted from Khalifa & Weir (2009), p.43)	78
2.4	Conceptual Framework	87
3.1	Research Procedure in Graphic View	93
3.2	Networks of Tests (Adopted from Wright & Stone (1979, p. 101)	117
3.3	Linking Procedure Using Common Item Equating (Concurrent Analysis) in the Current Study	118
3.4	Sample of Common Item Equating Data Matrix Configuration	131
3.5	Reading Tests: Person- Item Wright Map	136
4.1	Distribution of Scores among Persons in Test 1	147
4.2	Bubble Chartz	151
4.3	Standardized Residual Variance Scree Plot	154
4.4	Item-ability - Wright Map for all four tests	156
4.5	The Stacks of Items in Test 1	157
4.6	Empirical Scaling of Test Items Based on Cognitive Processes of Reading	159
4.7	Winstep Output Table for Reliability of 902 Measured Persons	162
4.8	Wright Item- Person Map for Common Item Linking	165
4.9	Most Unexpected Responses of the Students	169
4.10	Most Misfitting Students' Response Strings	170

4.11	Mean Item Measure of Cognitive Processing along with CEFR Levels	173
4.12	Boxplot for Inter-Faculty Reading Performance in CEFR-aligned Test	177
4.13	Distribution of Reading Performance of FAC, FMC, FAS, and FE Students on Logit Scale	179
4.14	Wright Person Map for Four Faculties	180
4.15	Wright Item Person Map: Students' Performance on Reading Tests	182
4.16	Distribution of Items Based on Cognitive Processing of Reading	185
4.17	Means of the Item difficulty level of Cognitive Processing and Person	192
5.1	An Overview of the Cognitive Processes of Reading according to Khalifa and Weir (2009) and the Present Study	212

CHAPTER ONE

INTRODUCTION

1.1 INTRODUCTION

This introductory chapter discusses concisely the importance of reading and academic reading skills for learning, and how they are important in English as a Second Language (ESL) classes, and for university students, generally. Avowedly, reading comprehension is integral in English as a medium of instruction (EMI). Assessment of reading ability along the baseline of the Common European Frameworks of Reference (CEFR) is presented, followed by the problem statement, research objectives, rationale, and significance of the study. It also outlines the limitations of the research, operational definitions, as well as overall organization of the study.

1.2 BACKGROUND OF THE STUDY

Reading, in addition to writing, speaking, and listening, is one of the core skills in language mastery. Perfetti (1985) defined reading as the skill of decoding printed words into spoken words. However, Fries (1963) embellished the definition of reading as a process of stimulating, cultivating, and evaluating the techniques of thinking; in fact, he later mentioned that reading is thinking guided by print. Widdowson (1979) stated that reading is the process of getting linguistic information via print. This perspective has been further illustrated by the latest definition provided by Urquhart and Weir (1998), that “Reading is the process of receiving and interpreting information encoded in language from via the medium of print” (p.22).

According to Grabe and Stoller (2011), the above single-sentenced definition has four deficiencies. Firstly, it does not convey the purpose of reading; second, the nature of reading abilities was not emphasized; thirdly, it does not connect reading with the cognitive processes; and fourthly, it does not address the social context in which reading takes place.

Further, reading is viewed as a cognitive process that engages the mind, as well as eye-movement, sub-vocalisation, etc. Since the 1960s, reading has been a major focus of interest among cognitive psychologists (Urquhart & Weir, 1998). They constructed reading models on the premise that reading happens in the human mind.

Reading models are built on the assumption that reading is a process as well as a product. According to Alderson (2000), the process approach emphasizes the interaction between the reader and the text, comprising several stages. The Reading-as-a-Process model is mainly classified into the bottom-up, top-down, and interactive approaches (Birch, 2007; Birch & Fulop, 2020; Urquhart & Weir, 1998). The reader uses cultural and world knowledge and generalized cognitive strategies in the top-down approach to creating meaning for the text by prediction and inferencing. On the other hand, the bottom-up model contains the precise bits of linguistic knowledge of the text from orthographic, phonological, syntactic, and semantic perspectives, which enable the mind to squiggle the page into meaningful symbols (Birch, 2007; Birch & Fulop, 2020). Due to severe criticisms of the aforesaid models, a resultant balanced model, known as the interactive model, combining the best of both approaches, emerged. Stanovich (1980) and Rumelhart (1977), as cited by Urquhart and Weir (1998), stated that in the interactive (a balanced) model, “a pattern is synthesized based on information ‘provided simultaneously from several sources’” (Urquhart & Weir, 1998, p.45).

Urquhart and Weir (1998) characterized reading as a product or componential approach, in which many components are involved in the process of reading comprehension. Hoover and Tunmer (1993) mentioned that the componential model “is to understand reading as a set of theoretically distinct and empirically isolable constituents” (p. 4). Word recognition, language background, world knowledge, and literacy are among the components involved in reading (Hoover & Tunmer, 1993; Urquhart and Weir, 1998). Based on this approach, numerous reading taxonomies consisting of sub-skills of reading emerged (Grabe, 1991; Munby, 1978; Vacca & Vacca, 2008).

Reading comprehension in the first language (L1) is different from that in the second language (L2) (Birch, 2007; Grabe, 2009; Jiang, 2011). Grabe (2009) indicated three major sets of differences: linguistic and processing differences, cognitive and educational differences, and sociocultural and institutional differences; whereas Birch (2007) differentiates the six stages of L1 reading development from three types of L2 reading development procedures, such as incomplete knowledge of English, inferencing, and missing English processing strategies. However, to better understand L2 reading, the role of L1 literacy in the development of L2 reading is essential (Carrell et al., 2000; Hudson, 2007; Wade-Woolley, 1999).

1.2.1 Reading in the Second Language

Reading in L2 is a gateway to enhancing the other skills to be succeeded in a particular language. Anderson (1999) highlights that:

Reading is an essential skill for English as a second/foreign language (ESL/EFL) students; and for many, reading is the most important skill to master. With strengthened reading skills, ESL/EFL readers will make greater progress and attain greater development in all academic areas. (p.1)

Similarly, Mikulecky (2008) mentions that reading is the key to acquiring a second language, which means that reading is the most significant fundamental instruction in all aspects of language learning. Additionally, Carrell et al. (2000) stated, “For many students, reading is by far the most important of the four skills in a second language, particularly in English as a second or foreign language” (p. 1).

Reading is recognized as a receptive skill, according to Aebersold and Field (1997), and has long been considered a prerequisite for learning a foreign language, because it serves as a critical source of input for the development of other skills. Improving one’s reading activity can certainly develop one’s writing and speaking skills. In other words, students who are good readers improve vocabulary, and write more grammatically compared to those who do not read much (Hafiz & Tudor, 1989). Conversely, “The studies are fairly consistent in showing that learners with

inconsequential exposure to the second language have difficulty in reading” (Hudson, 2007, p. 74) also concurred in this regard with other reading researchers.

Brown (2001) stated that reading comprehension is essentially a matter of acquiring adequate, effective comprehension skills for most second language learners who are already literate in a prior language. He suggested that both top-down and bottom-up strategies may need to be emphasized, depending on individual needs and proficiency levels.

1.2.2 Influence of Reading for Academic Success

In higher education, reading is regarded to be one of the essential skills for successful academic study (Hermida, 2009). Howard et al. (2018) mentioned that 83% of faculty members in California institutions of higher education believe that students’ reading skills play a vital role in academic success. Therefore, academic reading is crucial for the L2 learners at tertiary levels while they learn a discipline through English. Academic reading has been defined as “purposeful and critical reading of a range of lengthy academic reading texts for completing the study of specific major subject areas” (Sengupta, 2002, p. 3). Further, this reading draws students into a discourse within their major studies, as well as enhancing their writing and critical thinking skills (Paul & Elder, 2008). Rather than the surface reading approach, deep reading is more effective for academic success at the university level, because university-level reading is different from school-level reading (Hermida, 2009). Internationally, reading is considered to be crucial for higher academic achievement.

To have academic success, a learner needs to be a competent comprehender (Snowling et al., 2010). According to the simple-view formula presented by Gough and Tunmer (1986), reading comprehension (RC) is equal to decoding (D) multiplied by linguistic comprehension (LC), ($RC = D \times LC$). In the simple view, language comprehension becomes reading comprehension when word meaning is decoded or derived from print. Even if a reader has strong language comprehension, if there is difficulty with decoding, there is a possibility that the reader might be a poor comprehender. Kamhi (2007) elaborated that comprehension “is not a skill; it is a

complex of higher-level mental processes that include thinking, reasoning, imagining, and interpreting” (p. 28).

1.2.3 Reading Skill for English Medium Instruction (EMI)

Reading is a needed skill for students to master because information exists in text form in the world (Cimmiyotti, 2013). Much information is heaped in books, websites, magazines, newspapers, notice boards, notes, notices, brochures, leaflets, and sometimes pictures for visual reference for readers. Students must heavily focus on information in text formats to achieve better performance since the educational systems depend more on it. Carrell et al. (1989) highlighted that the ability to read is deliberated as an important feature to comprehend written material and to become successful in higher educational institutions, like universities.

Reading is exceedingly crucial for undergraduate students because they do not depend only on teachers, as the higher education system highly fosters self- or student-centred learning. Hence, they get themselves prepared for the new subjects by reading and understanding diverse sources alone or in groups. Therefore, it is evident that one’s reading ability, especially English-related reading, fosters one’s academic achievement, as was further confirmed by many research studies (Alkialbi, 2015; Anderson, 1999; Bernhardt, 2005; Grabe & Stoller, 2011; Li & Munby, 1996).

At present, English has been a medium of instruction in many countries around the world. According to Rogier (2012), Macaro et al. (2018), and Chalmers (2019), English Medium Instruction (EMI) uses English to teach curriculum subjects to students whose mother tongue or first language is not English. The popularity of EMI in school education around the globe has dramatically increased in recent decades; traditionally, this has been mainly in higher education. To compete in the international education market, universities started to offer courses, modules, and entire degree programmes in English to attract foreign students. To prepare the children to enter such universities, parents demanded the EMI approach in the “secondary”, “primary” and “preschool” curricula (Chalmers, 2019, p. 8).

If EMI is to be practised at the higher education level, students have to read and comprehend enormous amounts of texts to gain knowledge, listen to lectures, interact in the classroom, take notes, present on given topics, and write assignments and final exams in English. Thus, as it is required by many foreign universities for university admission, students must attain the C1 level of the CEFR, which illustrates the ability to use English fluently and flexibly in a wide range of contexts (Cambridge University Press, 2013).

1.2.4 Assessing Reading

Assessing reading is an intricate procedure similar to defining the nature of reading comprehension. Alderson (2000) illustrates that there are various ways of looking at how reading is developed and assessed. Using reading scales with a detailed description of each level, point, or band is one of the ways to assess reading. ACTFL proficiency guidelines, ALTE framework of language tests, Master and Forster scales, DIALANG, and CEFR can-do descriptors are some of such scales. Using language tests with different levels or bands is another way of assessing reading. These include Cambridge ESOL main suite exams like Key English Test (KET), Preliminary English Test (PET), the First Certificate in English (FCE), the Certificate in Advanced English (CAE), Certificate of Proficiency in English (CPE), and TOEFL, International English Language Testing System (IELTS), Learning Resource Network (LRN) ESOL exams; and International English Language Competency Assessment (IELCA).

American Council for the Teaching of Foreign Languages (ACTFL)'s reading definitions focus on text type, reading skill, and task-based performance. These guidelines are commonly used and influential in the USA. The guidelines lack familiarity as they are based on *a priori* definitions of levels and there is no empirical validation (Alderson, 2000).

The Association of Language Testers in Europe (ALTE) has developed a framework of levels, particularly for ALTE member language tests. It presents a general description of what a learner can do at each level before describing each skill separately (ALTE, 2002). According to the ALTE context, text type, language, and

reader's knowledge about the content are needed to be considered when developing reading, while it improved confidence, speed, awareness, length and amount of text, nature of the text, and text practicability (Alderson, 2000)

1.2.5 CEFR

The Common European Framework of Reference (CEFR) is a modified version of ALTE (Council of Europe, 2001a). ALTE's five levels have been aligned with A2 to C2 levels of the CEFR Framework (ALTE, 2002). It has three main groups comprising two stages each. It is intended to provide a common basis for describing "levels of proficiency required by existing standards, tests, and examinations in order to facilitate comparisons between different systems of qualification" (Cambridge University Press, 2013; Council of Europe, 2001, p.21).

Researchers advocate that a university student following the EMI system should be at the C1 level of CEFR (Council of Europe, 2001a; Jiménez-Muñoz, 2014). The Common European Framework of Reference for Languages (CEF or CEFR) is a way of standardizing the levels of language exams in different regions, introduced by the Council of Europe in 1996. Though it was intended to apply to European countries, as the CEFR descriptors have been translated into 40 European languages, including sign language, its influence is unquestionable in language teaching, learning and assessment beyond Europe (Figueras, 2012).

CEFR has been extensively utilized by many organizations and educational institutions as a reference tool for teaching, learning, and assessment for the last decade (North, 2014a; Waluyo, 2019; Wu & Wu, 2007).. In accordance with CEFR, language users are clustered into three main groups: Proficient users (levels C1 & C2), Independent users (levels B1 & B2), and Basic users (levels A1 & A2) (Council of Europe, 2001; Cambridge University Press, 2013). The CEFR levels represent a 'conceptual grid' of illustrative *can-do* descriptors of language competence, which was intended to be applied equally across different European languages since the 1980s (North, 2014b). A comprehensive Swiss research project scaled the levels through empirical Rasch analysis (North & Schneider, 1998).

This CEFR ‘*can-do*’ project is aimed to develop and validate a set of performance-related scales, describing learners’ actual capability in the foreign language (Council of Europe, 2001). Alderson’s (1991) distinction as cited in Council of Europe (2001, p. 244), “between *constructor*, *assessor* and *user*-orientated scales, the ALTE ‘*Can Do*’ statements in their original conception are user-orientated”. They assist communication between stakeholders in the testing process, and in particular the interpretation of test results by non-specialists. These scales of *can-do* descriptors are identified as an unparalleled success, as well as a preferred benchmark for language assessment and published courses worldwide.

As stated by the Council of Europe (2001, p.20), the CEF’s “focus has been upon the nature of language use and the language user and the implications for learning and teaching”. Although, Fulcher (2004) claims that the CEFR scale is designed from the teachers’ perspective, as Alderson (2007) critically pointed out that this scale is offered by language teachers who are neither trained testers nor applied linguists, nevertheless, its usage among several educational institutions is widespread (North, 2014a; Waluyo, 2019; Wu & Wu, 2007).

1.2.6 CEFR Level for reading to achieve academic success

One who performs well in reading inevitably acquires academic success. Those who score the B2, C1, or C2 higher levels of CEFR usually excel in their studies (Cambridge Assessment English, 2019). Language coordinator, Alison Standing, pointed out that students achieving C1 level invest much effort to learn English and show considerable improvement in their language proficiency, which “leads to broad, detailed understanding of English, giving them a strong foundation to manage the tasks they face during their university studies”(Cambridge Assessment English, 2019, p. 8).

Despite the fact that there are several criticisms on the perception that English is one of the sources limiting students’ academic performance, and the idea that the students who are in levels A2 and B1 “find it impossible to cope with the linguistic demands of academic tasks; as a student progresses towards C1-level” simply cannot

be proven (Jiménez-Muñoz, 2014, p.30). Nevertheless, the current study supports the idea that those who are at higher CEFR levels perform better academically.

So, in this study, the researcher will consider CEFR B2 First level as the baseline for measuring the students' reading performance for those who follow EMI (English as a medium of instruction). The Cambridge Assessment English handbook provides evidence that "B2 First is accepted around the world and offers students opportunities for employment, further study and travel" (Cambridge Assessment English, 2019, p. 6).

1.2.7 Current trend of Reading assessment in Sri Lanka

In Sri Lanka, the General Certificate of Education (G.C.E) O/L (Ordinary Level) and G.C.E. A/L (Advanced Level) English Language tests are the higher-level examinations administered nationally to evaluate the achievements of school children. These two paper-based exams are a composite of reading, writing, vocabulary, and grammar activities for a total of 100 marks (NETS, Department of Examination, 2016). 35 marks were given for reading questions in the 2014 G.C.E.O/L English Language Examination (NETS, DP, pp. 12-26), whereas 29.5 and 28 marks were given in the 2018 and 2019 examinations, respectively (NETS, DP). The condition for G.C.E. A/L is correspondingly comparable.

However, these two exams are not counted for the university entrance of undergraduates (Aloysius, 2015; UCAS, 2014). At the university level, for the majority of degree programmes, English is used as a medium of instruction; whereas, in a few degree courses like the human sciences and languages, English is offered as a compulsory or non-credit course. English Language is a compulsory subject for all faculty students of South Eastern University of Sri Lanka at least in the first year of their study (Wazeema & Kareema, 2017), as all these students had been learning English as an L2 from their grade three to G.C.E. Advanced Level (grade thirteen) for eleven years at school level, in compliance with the Sri Lankan primary and secondary school curricula (Wijesekera, 2012).

According to the *Fortieth Annual Report - 2018* of the University Grant Commission Sri Lanka, out of 253,357 candidates who sat for the GCE Advanced Level Examination-2017, 163,160 candidates were eligible to enter any public university in Sri Lanka. However, only 30,550 were selected for the public universities (University Grants Commission, 2020b), which means only 12% of the applicants were selected.

However, at the university level, there is no unified test to assess the English competence of these students. For the first time in the history of university education in Sri Lanka, a separate exam for all four skills was carried out by all thirteen state universities in 2015 under the HETC (Higher Education for Twenty-First Century) project funded by the World Bank (Jayasinghe & Wijethunge, 2015; Ratwatte, 2016). This exam is known as the University Test of English Language (UTEL). Although the process to continue this test for the whole university student population in Sri Lanka is undertaken by the University Grant Commission, it is, however, still at the planning level, and it is expected to be implemented sometime in the near future.

Evaluating the level of reading comprehension is significant as reading is identified as the most important skill. Therefore, assessing the reading performance at the tertiary level in Sri Lanka is crucial in the L2 instructions.

The test material used in the UTEL exam is not open to the public, as the question bank is kept secret. Hence, to evaluate the university students' reading ability, the researcher needs to produce a new instrument. Since the CEFR has exerted tremendous influence on language learning, teaching, and assessment around the world from its inception in 2001, the present research incorporates reading passages from the CEFR-aligned international standardized examination. LRN past papers and sample papers were adapted to develop the test to apply the validity theory proposed by Khalifa and Weir (2009).

Further, in the situational analysis conducted in 2011 among stakeholders, such as the Ministry of Higher Education officials, Sri Lankan university officials, and lecturers to produce course materials and test items to examine the English and IT competencies of the entry-level university students, it was decided to set a minimum band of 5 of the UTEL as benchmark (University Test of the English Language, Sri

Lanka), when the students graduate from the university (Wikramanayake et al., 2012). The UTEL benchmark has 10 levels, from 0 to 9, which are conformed with the CEFR level descriptors (Jayasinghe & Wijethunge, 2015; Kulasingham et al., 2012; Senaratne, 2013).

1.2.8 Reading from Islamic Perspectives

The background of the study is viewed through the Islamic perspective of reading. The first word of the Holy Quran handed down to the last messenger of Allah, the Prophet Muhammed (PBUH) was ‘read’.

Read! In the name of thy Lord, who has created (all that exists). He created man from a clinging substance. Read! And your Lord is the most Generous. Who taught (man) by the pen. He taught man that which he knew not. (Al-Quran, al-Alaq: 1-5)

Islam fervently encourages mankind to seek knowledge for the betterment of life in the world and the world hereafter. Seeking knowledge can be accelerated by practising the necessary skills like reading, writing, sharing knowledge, etc. Prophet Mohammed (PBUH) declared that “Seeking knowledge is compulsory to every Muslim man and woman” (Ibn Majah, Hadith No: 224). Further, one of the Islamic scholars, Khalifa Ali (Ral), mentioned that “A person who keeps himself occupied with books, will never lose his peace of mind”. The above references portray the essence that Islam confers to reading.

1.3 PROBLEM STATEMENT

In light of seven research gaps recognized by Miles (2017), the current research addresses five identified gaps: the knowledge gap, the evidence gap, the theoretical gap, the methodical gap, and the population gap.

One of the most significant abilities for L2 tertiary students studying in English is the ability to read academic materials (Shen, 2013). The CEFR C1 level is a prerequisite to L2 learners while they pursue their higher education in countries where their native language is English, as discussed in section 1.2.3. The minimum requirement for those students for their academic achievement in the ESL or EFL was identified as the B2 level and the upper B1 level, discussed in section 1.2.6. However, there is no watertight decision on the level of reading to cope with academic reading. Since this kind of research has not been studied so far, this study tries to find out the level of reading required for academic success.

Moreover, examining the reading ability of undergraduates of the South Eastern University of Sri Lanka (SEUSL) is a new paradigm in the history of English language teaching and research at SEUSL as well as other universities in Sri Lanka. Exploring the nature of reading in the Sri Lankan context is far from satisfactory, and there is still a large gap in the knowledge that research studies can contribute. Only a few research studies have been carried out to investigate the influence of EMI on the English language proficiency of students in universities (Andrew, 2017). Choosing SEUSL students as the research population undoubtedly solves the issue of the population gap.

There is still little known about the L2 reading process and there is a need to conduct more research on it (Grabe, 1991). Although there were a few studies carried out in the L2 reading context during the past decades, there are still some issues needed to be addressed. In the Sri Lankan educational system, either at school levels or tertiary levels, there is no particular test to assess reading ability individually. Assessing reading is crucial as reading is the most important skill for academic achievement. However, this is not properly done even at university levels in Sri Lanka, besides a single university-wide attempt conducted nationally in 2015. Therefore, to assess reading skills separately as in international proficiency tests like TOEFL, TESOL, IELTS, PET, FCE, CEF, IELCA, and the like, there is a need to prepare reading tests. As AlKialbi (2015) proposed further research in the areas of word-level issues in reading development, main idea comprehension, instructional routines, social-cultural context influences on reading, assessment of reading, etc., the present study aims to develop such tests.

Adopting existing tests is not suitable for the selected population, as the tests' difficulty levels are above the students' ability level. As a result of this, the actual ability level of the students cannot be measured precisely. Developing new tests is also time-consuming. Therefore, concerning cost effects, and the research ethics, Learning Resource Network (LRN)'s selected reading passages and questions were adapted for developing new tests to fulfil the requirement of the current survey. The LRN materials have been validated according to the CEFR framework (Hidri, 2020; Learning Resource Network, 2015). To fill this knowledge gap in evaluating university students' reading ability in terms of the international benchmarking system, this research focuses on the CEFR scale of reading.

In addition, there is not much research relating to the study of reliability and validity of locally-designed proficiency tests in Sri Lankan universities or Sri Lanka overall. In proficiency testing, the transferability of inferences of validity is available to recognized tests, but it is limited to locally-designed tests. Keeping these in mind, it would be important to develop an adaptable method to examine the reliability and validity of locally-designed tests so that they could be applied in other situations as well. This study provides an excellent setting for such studies in the future by providing a presentation of a coherent and flexible methodology to evaluate any test. To sum up, in this way, it provides novelty in knowledge.

The next step is to explain the theoretical gap brought out by this research. Test validation is crucial (Bachman, 2005; Fulcher & Davidson, 2007; McNamara, 2006; Messick, 1989; O'Sullivan & Weir, 2011). In addition, the success of test validity relies on the degree to which the tests fit the intended purpose (Kane, 2012, 2016; Messick, 1989; Weir, 2005). Hence, the model of validation and the theoretical framework which underpins the instrument design, data collection, and evaluation of this study, is the socio-cognitive model of language test development and validation proposed by Khalifa and Weir (2009), from the updates of Weir (2005) and Urquhart and Weir (1998). In the socio-cognitive model, "the abilities to be tested are demonstrated by the *mental* processing of the candidate (the cognitive dimension)", and the model considers "the use of language in performing tasks as a social rather than a purely linguistic phenomenon" (Khalifa & Weir, 2009, p. 4). This validation framework consists of six types of validations, namely, test-taker characteristics,

context validity, cognitive validity, scoring validity, consequential validity, and criterion-related validity. Reading is assessed through these six validation measures.

In assessments to understand the learners' cognitive processes and knowledge structures, empirical evidence is needed to support the theoretical premise that the target cognitive processing intended by a test designer are consistent with the actual processing learners utilize throughout the assessment. Therefore, significant research is invited to provide insightful evidence for Khalifa and Weir's socio-cognitive validation framework to fill the gaps in theory (Bannur et al., 2015; Bax & Chan, 2016; Brunfaut & McCray, 2015; Dunlea, 2015; Khalifa & Weir, 2009; Krishnan, 2011; Weir et al., 2009; Wu, 2011). Further, Khalifa and Weir (2009) mentioned that "There has been limited L2 research to date addressing the cognitive processing" (p, 219) and they urged to carry out research in this area.

Although research studies on cognitive validation focusing on metacognitive activities, such as goal setting at local and global levels (Dabiri & Kashefian-Naeeni, 2021; Moore et al., 2012), and monitoring activities comprising expeditious and careful reading (Aryadoust & Zhang, 2016; Katalayi & Sivasubramaniam, 2013; Krishnan, 2011; Weir et al., 2009, 2012) were carried out, research focusing on the central processing core is lacking.

Further, as reviewed from the previous literature, "no serious studies appear to have been undertaken in which the focus is on the contextual parameters and cognitive processing involved in academic reading" (Weir et al., 2009, p. 100); therefore, Weir et al. strongly appealed for further research in this area. Thus, given the paucity of empirical evidence in support of cognitive processing in the existing literature, this study has the potential to provide fresh light on the extent of difficulty in cognitive processes.

The next concern is the methodological gap. Considering the shortcomings in CTT (Classical Test Theory), the present survey is proposing to employ the RMM (Rasch Measurement Model) of IRT (Item Response Theory), as it is identified as a prospective model for evaluating the insightfulness of the reading ability, especially using the socio-cognitive model (Dunlea, 2015). Many research studies used factor analysis or confirmatory factor analysis to identify the subskills of reading

comprehension (Davis, 1968; Spearitt, 1972). Using Rasch would provide a better understanding of the reliability and validity of the item difficulty, and persons' ability. Further, to analyse the content validation of the experts, this study used the Item Objective Congruence (IOC) method applied by Rovinelli and Hambleton (1977), which is prescribed as one of the best methods to analyse the expert judgment (Berk, 1984; Turner & Carlson, 2003); and the simplified formula presented by Crocker and Algina (1986) for multidimensional objectives, was utilized in this research, concerning the inappropriateness in applying Rovinelli and Hambleton's (1977) single objective formula. The application of these various research methodologies will give new fruitful insights into the present study.

As collecting evidence for the empirical validation strengthens the framework of socio-cognitive validation (Weir, 2005), developing an instrument to assess the reading comprehension level of university students is indeed an influential area of research in the context of Sri Lanka. The adoption of socio-cognitive theory in this present study will hopefully fill the empirical gap and would provide more evidence for the theory.

Concerning all the gaps identified by the researcher, this research will provide better interpretations that are well-grounded theoretically and measured systematically. Further, utilizing an international benchmarking for assessing will help to offer correct directions for teaching, learning, and assessment of reading which will, in turn, expose the weaknesses, strengths, and the way forward for instructions. Hence, the knowledge gap is explicated. It is true that the gap in knowledge may become a problem for educational planners, policymakers, and practitioners like students, teachers, stakeholders, test developers, and administrators. Consequently, this research study will seek to provide informed clarifications for this issue by filling in all these gaps.

1.3.1 Research Objectives

This study seeks to develop local reading tests in alignment with an international standard, namely the Common European Frameworks of Reference (CEFR) and to validate the tests using the Rasch Measurement Model (RMM). Further, this study will examine the performance of university students whose medium of instruction is English (EMI), and attempt to answer questions related to their performance in reading skills. The key objectives of the study are as follows:

1. To adapt and validate reading comprehension tests aligned with the CEFR.
2. To measure the level of reading proficiency of the SEUSL (South Eastern University of Sri Lanka) undergraduates, whose medium of instruction is English.
3. To profile and ascertain the students' cognitive processing in reading.
 - a. To determine which cognitive process is easy to attain.
 - b. To identify which cognitive process is difficult that needs attention for intervention.

1.3.2 Research Questions

1. What are the psychometric properties of the CEFR-aligned reading tests?
2. What is the performance of the students in the CEFR-aligned reading tests?
3. What is the performance level of SEUSL undergraduates who follow the EMI system, in the cognitive processes of English reading?
 - a. In which cognitive processes of reading do the SEUSL students indicate higher achievement?
 - b. In which cognitive processes of reading do the SEUSL students indicate lower achievement?

1.4 RATIONALE FOR THE STUDY

For a language learner, reading is a central skill that he or she must master, as much information is existing in text form in the world (Cimmiyotti, 2013). Especially for an English as a second language (ESL) learner, reading is more significant because of the surfeit of information in written English. Achieving mastery in reading is indeed highly needed for ESL learners, especially for those who are at the tertiary level. Therefore, they must put much effort to develop their reading ability for better achievements in academic, professional, as well as social life, as they will join the job market soon. Since this nature of the study is lacking in the Sri Lankan ESL context, it is crucial to investigate it.

On the other hand, developing reading skills and understanding the cognitive processing in reading, is challenging. Still, there is not a finalized definition of the hierarchy of reading skills. The recent surveys on identifying the hierarchy of reading skills indicated that there is not a strict hierarchical ordering of sub-skills (Badrasawi, 2012; Hudson, 2007; Jusoh, 2018; Rosenshine, 2017). However, Khalifa and Wier (2009) believe that there is a hierarchical order of reading cognitive processes, unlike reading sub-skills. They remark there are low-order skills and high-order skills in cognitive processes. However, there is a need to support their claim, because even reading experts find it difficult to agree on the level, or hierarchical order of item difficulty (Alderson & Lukmani, 1989). So, this study will provide insight into the level of cognitive processes of reading. In order to guide teaching and assessment of reading, it is necessary to have a solid understanding of the diversity of the nature of reading based on a bigger amount of data.

1.5 SIGNIFICANCE OF THE STUDY

The study will also be beneficial to English language learners, teachers, test designers, item writers, educators, and policymakers to gain a broader perspective on language testing and evaluation, exclusively to reading competency. Aligning the local graduates' English performance to an international framework such as the CEFR will certainly enable the stakeholders to measure their language ability perfectly.

Particularly, the students will recognize their strengths or weaknesses in English language skills and make use of the results of the test, which is standardized according to an international framework, to improve themselves. University teachers and material designers will focus on developing the reading skills and their sub-skills to upgrade the students' English proficiencies. This study will further fortify the operation of the UTEL (University Test of English) among Sri Lankan university students. Future research studies in this arena will undoubtedly be promoted by the findings of the current survey.

Collecting evidence to underpin the validation of tests is crucial in language testing, as testing is always concerned with evidence-based validity (Weir, 2005). Methodical as well as more competent test development processes could immensely contribute to the development of the test quality and validity.

1.6 LIMITATIONS OF THE STUDY

There are a few limitations in this research that may affect the generalizability of the study. First of all, this study focuses only on the reading skills based on the socio-cognitive validation framework, which is common to all four skills. Secondly, among the six types of the validity framework, such as: test-taker characteristics, content validity, scoring validity, consequential validity, and criterion-related validity, only cognitive validity is discussed in-depth, especially in terms of the central processing core. Finally, the test is designed based on the CEFR B1, B2, C1, and C2 level passages, but they did not focus on A1 and/or A2 levels, which are comparatively low levels of proficiency for undergraduate students. Overall, the highest level of test difficulty of the tests focuses on the CEFR C1 level. However, despite these limitations, this study attempts to highlight the challenges addressed.

1.7 OPERATIONAL DEFINITION OF TERMS

The terms utilized in this study are defined below:

1. Reading comprehension: Reading comprehension is understanding a written text, which means extracting the required information from the text effectively. In reading comprehension, the students need to read a text, comprehend the relation of one sentence with others within the text, and connect the text with their background knowledge.
2. Reading skills: Reading skills refer to information processing techniques that are automated and applied to a text unconsciously, whereas reading strategies refer to actions selected intentionally to achieve specific aims (Paris et al., 1991). “A reading skill can be described roughly as a cognitive ability which a person is able to use when interacting with written texts” (Urquhart & Weir, 1998, p. 88)
3. Reading test: Assessing the learners’ reading through distinguishing what they can do well and what they find difficult. Assessing reading comprehension is challenging for test developers and teachers (Mckee, 2012).
4. Cognitive processes of reading: Reading is considered a cognitive process (Stauffer, 1967). A reader applies different cognitive processes at each level of comprehending. Cognitive processes are divided into two types: low-order- thinking process and high-order thinking process. Word recognition is at the lower level of reading comprehension, whereas the ability to understand the main ideas and to make inferences are assumed to be higher level processes (Khalifa & Weir, 2009; Urquhart & Weir, 1998).
5. Socio-cognitive model of reading: Weir’s (2005) *Language Testing and Validation: an evidence-based approach* contributed to the new development of a reading model based on the socio-cognitive framework, which considers reading as a mental process accompanied by social experiences.

6. Cognitive validity: The extent to which a test elicits from the test takers cognitive processes that are similar to those they would typically use in a real-life context.
7. The eight central cores (cognitive processes) of cognitive validity are mentioned in Khalifa and Weir (2009):
 - a. Word Recognition (WR): The reader recognizes the word in question or discovers the meaning of a word on their own and matches it to the text. This process occurs at the word level.
 - b. Lexis Access (LA): The reader uses knowledge of (morphology) word meaning or word class to identify synonyms, antonyms, hypernyms, or other related words and matches them in the text. This occurs at the word level.
 - c. Syntactic Parsing (SP): The reader uses grammatical knowledge to establish comprehension to identify answers without logical problems. This can occur at the clause or sentence level.
 - d. Establishing Propositional (core) Meaning (EPM): The reader expeditiously uses knowledge of lexis and grammar to establish the meaning of a sentence at the local level. It is a literal understanding of what is on the page. This occurs at the sentence or clause level.
 - e. Inferencing (I): The reader goes beyond literal or explicitly stated meaning to infer a further significance. The reader can selectively read the paragraphs for the main ideas and implicitly expressed ideas in the text. This can occur at the sentence level, paragraph level, or text level.
 - f. Building a Mental Model (BMM): The reader uses several features of the text to build a larger mental model by recognizing major contrasts in a comparative and contrastive text type. This occurs at a whole text level.

- g. Creating a Text Level Structure (CTLS): The reader uses genre knowledge to identify the text structure and purpose of the whole text by analysing and distinguishing major ideas from supporting details. A trained reader decides how the various sections of the text work together, and which parts of the text are vital to the intent of the author or the audience. This occurs at the text level.
- h. Creating an Inter-Textual Representation (CITR): Understanding text and comparing it across other texts. This occurs beyond the text level.

1.8 ORGANIZATION OF THE STUDY

This study develops a set of CEFR-aligned reading tests targeting B1 to C2 levels, as well as administers the tests among the SEUSL undergraduates, and validates the results of the tests. All these procedures of the study are discussed in five important chapters.

This introductory chapter has provided an outline of the research and background to the study, followed by the problem statement, the purpose of the research, research questions, the rationale for the study, and the significance of the research. In addition, it also explains the limitations of the study and the operational definitions. In Chapter Two, a review of the literature relevant to the current study is given in three main sections. In the first section, the historical background of English language teaching in Sri Lanka and the present trend of the English assessment system are discussed, while the second section reviews the theoretical frameworks of reading and assessing reading. The final section deals with the measurement procedures. In Chapter Three, the methods used in this study are described along with the results of the pilot study. The research findings are provided in Chapter Four. Finally, a review of the research findings, conclusions, implications, and recommendations for future research are offered in Chapter Five.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter reviews different types of literature relevant to the present research. First, it elaborates on ELT in Sri Lanka, issues in ELT, and language testing and evaluation trends and research in Sri Lanka. Then, it focuses on conceptualizing reading ability followed by the assessment of reading, constructs in assessing reading, levels of reading, reading assessment scales, and a special reference to the CEFR scale. A detailed discussion of the underlying socio-cognitive framework is also provided. The last section highlights the Rasch Measurement Model.

2.2 ENGLISH LANGUAGE TEACHING (ELT) IN SRI LANKA

To recognize the scenario of English language teaching (ELT) in Sri Lanka, information about the history of the English language, and ELT in colonial Sri Lanka, the current trend of ELT, and issues in ELT need to be discussed in detail. This discussion will help in understanding the nature of English language testing and evaluation in the Sri Lankan context.

2.2.1 The History of English in Sri Lanka

A brief history of English is pivotal to the understanding of the present study. The first encounter with English in Sri Lanka was when British merchants came to the country in the 1600s. In 1796, the British captured the nation, a Dutch colony then, from the Dutch, and in 1815 the entire country was under the control of the British Raj. In 1948, the country became an independent nation. Since then, English had been the only language of administration, education, and industry in Sri Lanka, until 1956, when Sinhala was declared the sole official language (Attanayake, 2017). In 1987,

however, both Sinhala and Tamil, being the vernacular languages of Sri Lankans, became the official languages, while English was relegated as a link language. English, nevertheless, had been taught as a subject in schools since 1956.

2.2.1.1 English Language Teaching in Colonial Sri Lanka

In 1829, a Royal Commission named the “Colebrooke and Cameron Commission” (CCC) took control of the administration of the Island. The commission proposed to include a proportion of local citizens in the administrative services. The provision, though, was that those who were engaged in that service needed to be proficient in English (Aloysius, 2015; Saunders, 2007). Further, the CCC recommended English as the main language for secondary schools and universities. Hence, local English schools were established, and the Christian missionary schools, which had been running the English Language schools that had previously taught religion in the vernacular, also adopted English as the medium of instruction. All government schools were monitored by the newly formed School Commission of the CCC, which, inadvertently, neglected the Sinhala and Tamil languages.

English education was only accessible to the well-to-do elites in the society, as the English schools were fee-levying schools. Middle-class Sinhalese and Tamil children were sent to Anglo-vernacular schools to prepare them for lower rank posts in the government. The majority of children of the lower socio-economic classes gained free education from their vernacular or traditional schools in rural areas. Because of this situation, English is regarded as the language of the elites, and this situation created a great deal of animosity among the majority of locals, naming English as ‘*Kaduwa*’ in Sinhala, meaning ‘sword’, which segregates the rich from the poor (Fonseka, 2003; Gunasekera, 2005; Parakrama et al., 2021; Walisundara & Hettiarachchi, 2016).

2.2.1.2 Present ELT Trend in Sri Lanka

ELT in Sri Lanka can be divided into two sub-classifications, school education (both primary and secondary), and tertiary education. The first part of this section discusses the development and challenges of the ELT at the school level, whereas the second part deals with the ELT at the university level (higher education/ tertiary level).

2.2.1.2.1 ELT at School Level

The policy of teaching English as a second language to all was implemented in 1991 to promote national unity among Sinhalese, Tamils, Christians, and Muslims, as well as to prepare the young generation to meet the international demands for English communication so that they can be employed in the modern job market (Aloysius, 2015). The National Education Commission realized the need for the improvement of ELT in schools. The 1997 Presidential Task Force on General Education - Sri Lanka, in its publication titled '*General Education Reforms*', proposed the following policy decisions.

- Introduction of Activity Based Oral English (ABOE) programme for Grades One and Two to use simple English for communication, starting in 1999
- Formal teaching of English starting in Grade 3
- Necessary texts and guidebooks development along with supplementary materials and audio cassettes
- English as one of the main subjects for the G.C.E Ordinary level examination, and alternative English syllabuses at Grades Ten and Eleven
- Introduction of General English as a new subject in the G.C.E. Advanced Level subject
- Teacher training and assessing teachers' capabilities to teach General English

(As in 2001 NIE's English Unit policy document cited in Walisundara and Hettiarachchi (2016), and Perera, 2001 cited in Aloysius (2015))

According to this policy, English is taught to children in Grade 1 regardless of the medium of instruction, and it is employed as a tool of communication through so-called Activity Based Oral English teaching. And further, it is being taught from Grades 3 to 13 as a compulsory subject (Brunfaut & Green, 2019; Walisundara & Hettiarachchi, 2016). From time to time, though, there were several educational reforms to make English accessible to all Sri Lankans, and to make English a link language among the different ethnic groups; however, these objectives were not successfully achieved throughout the whole student populace in the country. This then gave rise to the widespread perception that English Language teaching is a failure in the country (Aloysius, 2015; Attanayake, 2017; Gunawardana & Karunarathna, 2017; Walisundara & Hettiarachchi, 2016; Wijesekera, 2012).

2.2.1.2.2 ELT at University Level

Assisting undergraduates to develop their English language skills systematically was introduced in 1960 after the enrolment of ‘svabhāsā’-educated students (those educated in their respective Sinhala or Tamil mother tongues) for English medium degrees in the universities (Parakrama et al., 2021). The history of ELT in the university system reflects the challenges and constraints that it confronted within the university system and among the students. However, in 1986 there was an improvement of ELT in the universities with the establishment of the English Language Teaching Units (ELTU). Further, it has been upgraded to the Department of English Language Teaching (DELT) established in the University of Kelaniya in 2017. Currently, there are around 16 DELTs in all 17 national universities (Parakrama et al.). The DELTs serve various faculties and disciplines.

The so-called DELTs or ELTUs have made three significant contributions to the theory and practice of ELT over the history of more than 60-year evolution of ELT in universities. The key insights and methods reinforced include: (i) the significance of students' first language in learning English; (ii) teaching and validating Sri Lankan English; and (iii) addressing ideological and socio-political attitudes about the colonial and neo-colonial status of English locally and globally (Parakrama et al., 2021).

In addition to this, the DELTs have continued to improve the English language competence of undergraduates in all faculties of all universities, while also developing specialised degree programmes like Teaching English as a Second Language (TESL) to address the national need for qualified and competent English teachers, conducting widely popular external programmes for an unlimited number of non-university learners, conducting research on critical issues on the teaching and learning of English as a second language, and establishing the UTEL tests (Attanayake, 2017; Parakrama et al., 2021; Rameez, 2019; Rathnayake, 2013).

However, there are many challenges that these DELTs face. Treating them as stepmothers indicates inadequate recognition within the university community. Some of the major impediments include: lack of staff, socio-economic influences, and political involvements (Attanayake, 2017; Fonseka, 2003; Gunawardana & Karunarathna, 2017; Parakrama et al., 2021; Rameez, 2019; Walisundara & Hettiarachchi, 2016; Wijesekera, 2012).

Nevertheless, recent studies show that the resistance to English as a weapon used by the colonizers, and afterwards the English-educated elite of Sri Lanka, to maintain power, is waning (Ratwatte, 2016). Further, upgrading higher education is the past and present government policies to make Sri Lanka a Knowledge Hub of Asia. For that, they are working to produce marketable graduates. Improving the English language competency of graduates is one of the key factors to make them stand alone in the global job market. In this endeavour, the recent government policies like AHEAD (Accelerating Higher Education Expansion and Development), HETC (Higher Education for Twenty-First Century), and IRQUE (Improving Relevance and Quality of University Education), place the improvement of the graduates' proficiency level of the English language a very high priority (Dissanayake & Harun, 2012; Parakrama et al., 2021; Rameez, 2019; Umashankar, 2017).

Consequently, both the university administrations as well as the students have started to realize the significance of learning, teaching, and assessing the English language. Furthermore, there is a positive change in the way they treat the recognition of DELTs in the university environment.

Given all of these considerations, ELT in universities alone cannot be regarded as a need for young Sri Lankan undergraduates to complete their academic duties; rather, it is virtually a life-long capacity that sets their future life aspirations as educated residents of the country, as highlighted by Jayasinghe and Wijethunge (2015). It should be treated as “a life-skill devoid of hegemony” as Ratwatte (2016, p. 102) pointed out.

2.2.2 Issues in ELT in Sri Lanka

Since the ELT in Sri Lanka is considered a failure, the issues in ELT should be discussed further. The reasons behind the failure of ELT in Sri Lanka were thoroughly discussed by many intellectuals. Aloysius (2015, p. 9) mentioned in his thesis that:

Many studies (Fernando and Mallawa, 2003; Hettiarachchi, 2010; Karunaratne, 2008; Perera, 2006; Perera et al., 2010; Perera, 2001; Wijesekera, 2012) have dealt with various problems and challenges related to the failure of ELT in Sri Lanka.

The literature shows that the effects of English education at the school level significantly affect ELT in tertiary levels or higher education. Although one of the main objectives of Attanayake’s (2017) research was to get rid of the blame on school English education for the failure of the ELT in Sri Lankan universities, she strongly believed that English language teaching in Sri Lanka was an utter failure. She emphasised that:

Despite having studied English for nearly 10 years during their school careers and being among the best of their generation to have passed the highly competitive university entrance examination, undergraduates face difficulties in achieving the English language proficiency demanded of them by employers. The result of teaching English to students throughout their academic life, commencing from primary school and culminating in the university, has so far resulted in complete failure. (Attanayake, 2017, pp. 1–2).

Although the English language is taught from grade three to G.C.E Advanced Level as a subject in schools, a pass in A/L English is not necessary to enter university (Aloysius, 2015; Navaz, 2016). This is another factor that demotivates the students' interest in learning English. Further, the “minimum requirements for university admission” do not take into consideration fluency in the English language, in which the majority of degree courses are taught (University Grants Commission, 2020a, p. 10).

Inadequate resources and lack of motivation among the students towards learning the language are some of the reasons for this failure, as identified by Parakrama et al. (2021). Issues related to students, problems confronted in the teaching and learning environment, issues in textbooks, learning materials, and curriculum, problems teachers face, and pitfalls related to socio-economic and political dilemmas, are among the main findings of the PhD study on ‘Problems of English teaching in Sri Lanka: how they affect teaching efficacy’, conducted by Aloysius (2015). As was mentioned in the previous section, the findings of Attanayake (2017), Azeera et al. (2016), Fonseka (2003) Gunawardana and Karunarathna (2017), Parakrama et al. (2021), Rameez (2019), Walisundara and Hettiarachchi (2016), and Wijesekera (2012) confirmed the reasons for the foundering of ELT in Sri Lanka.

2.2.3 English Language Testing and Evaluation in Sri Lanka

There have been just a few research studies in the Sri Lankan context in the area of language testing and evaluation so far (Brunfaut & Green, 2019). A new-fangled research survey on *English Language Assessment in Sri Lanka* was published under the *Transform* project of the British Council and Ministry of Education, Sri Lanka in 2019. This research spotlighted mainly the General Certificate of Education (GCE) examinations in the English Language, namely, the G.C.E O-Level and A-Level examinations. These are identified as high-stakes tests due to their exam-oriented nature. According to Brunfaut and Green (2019), the responsibilities shared by three different entities like the National Institute of Education (NIE) to produce the curriculum, the Department of Examinations (DoE) to conduct the public

examinations (Grade 5 scholarship, GCE O- and A-Level), and Educational Publications for the writing, publication, and distribution of textbooks, is a significant factor that influences the productive assessment in Sri Lanka.

The National Education Commission (NEC), in their 2016 proposals for General Education in Sri Lanka (National Education Commission, 2016), identified the following common issues in the general education assessment system in Sri Lanka:

1. The heavy exam-orientated nature of the educational practice.
2. The quality of the national examination papers.
3. The priorities for memory-based knowledge and lower-order skills in the exams.
4. The underutilization of assessment results for educational policymaking.
5. The under-exploitation of assessment results (formative and summative) to inform language learning, teaching, and remediation.
6. The under-implementation of assessment results for classroom lamination.
7. The lack of assessment skills among teachers.

These issues are critically effective in the English language teaching, learning, and assessment process as well. In addition to these weaknesses, Brunfaut and Green (2019)'s findings identified some more major weaknesses of the present assessment system as follows:

1. A lack of testing of listening and speaking abilities.
2. Assessments have not been understood as measures for diagnostic purposes, but they have been used merely for record purposes, without any application.
3. Lack of language assessment expertise and training among teachers and other stakeholders.
4. A misalignment of the curriculum and the assessments.

Therefore, incorporating the problems identified in the 2016 NEC proposal, and their findings, Brunfaut and Green came up with the following recommendations for the assessment of the English language in Sri Lanka.

1. Establishing a ‘full circle’ in The English Language education (teaching-learning-assessment).
2. Close collaboration between Departments: To ensure such a ‘full circle’, it is vital that the institutions responsible for the curriculum (NIE), the textbooks (EPD), the exams (DoE), and teacher training (NCE), work together.
3. Enhancing the development of learners’ English listening and speaking skills.
4. Improving the quality of English language assessment.
5. Developing stakeholders’ language assessment literacy.
6. Addressing systemic factors.

(Brunfaut & Green, 2019, p. 38,39)

So, a close collaboration among the stakeholders and training them on language assessment principles should be carried out to better evaluate English language proficiency at any level including both schools as well as universities.

To form unity in the English language assessment among university students, the UTEL exams were first created in the late 1990s and piloted in 2000 (Ratwatte, 2001). The UTEL is a nationwide assessment produced by the Ministry of Higher Education in partnership with the HETC project, which is available to university students in Sri Lanka (Jayasinghe & Wijethunge, 2015; Kulasingham et al., 2012; Ratwatte, 2016; Senaratne, 2013). The UTEL consists of two online components focused on reading and listening skills. Productive skills like speaking and writing were created to be tested in universities employing traditional testing methodologies (Senaratne, 2013). The UTEL benchmarking is aligned with the six levels of the

CEFR scales that span the benchmarks from 0 to 9, consisting of 10 levels (Jayasinghe & Wijethunge, 2015; Kulasingham et al., 2012; Senaratne, 2013).

During the academic year 2013/14, all 15 state institutions offered UTEL assessments both online and offline for the first time in 2015, and around 13,000 students from all 15 universities took the test. Only 11.26 % of pupils (N=1499) received a band score of 5 or above in all four areas, indicating inadequate performance (Ratwatte, 2016). It is noted that a minimum UTEL band score of 5 is recommended for successful university graduation.

2.2.3.1 Previous Research in Language Testing in Sri Lanka

Research studies on the washback effect and a study on test validation have been identified by the researcher as literature from the Sri Lankan perspective.

2.2.3.1.1 Research Studies on Washback

Alderson and Wall (1993) studied the washback effect, the influence of testing on the teaching and learning of the G.C.E. O/L (Ordinary Level) English Language test in Sri Lanka for the first time, using the classroom observation method. They deliberated the positive and negative properties of washback in terms of the content of teaching, instructional methods and technique of teaching, and ways of assessing. The results of the project, after two rounds of classroom observation, indicated that washback occurred, to some extent, in instructional content in both positive and negative forms and in ways of assessing; however, there is no evidence of washback in methodology. Further, Wall's projects in 1996, 1999, and 2005 were carried out to bring out the washback influence within the Sri Lankan context addressing a change in the G.C.E O/ L English examination.

A research study on the washback effects of speaking assessment of a newly developed English as a Life Skill Programme was performed by Umashankar (2017) on students', teachers', teacher trainers', and policy makers' perspectives. The study

found that there was an influence of tests on teaching and learning at the school level and on teaching practices.

2.2.3.1.2 Research on Test Validation

Only a trickle of studies has shed light on the assessment of language in higher education in Sri Lanka. Among them, a survey was conducted to validate the English Placement Test (EPT), administered in 1999 at the University of Colombo, Sri Lanka, for the students of the Faculty of Arts. This test was evaluated in terms of test reliability, construct validity, authenticity, inter-activeness, impact, and practicality, using Bachman and Palmer's (1996) framework (Abeywickrama, 2000). The findings of the study showed that the administration of the EPT (99) as a placement instrument, allocated students well in different levels, and the test possessed construct validity in terms of grammar, reading, and writing, whereas the test ignored the listening and speaking skills. Furthermore, high authenticity, interactiveness, and practicality were observed in the administration of EPT in 1999.

Nevertheless, there is no particular experiment that studied the in-depth nature of testing and evaluation, particularly on the different language skills of ESL undergraduates. While exploring the challenges and needs of those students to learn English, more than 84% of teachers concurred that reading skill is more essential for the academic success of undergraduates who follow English as a medium of instruction (De Silva & Devendra, 2014). Keeping these factors in mind, it is prudent to explore the reading comprehension ability of students in Sri Lankan universities to attain academic success, as it has not been illustrated heretofore at the university level.

2.2.4 Reading Ability of Undergraduates in Sri Lanka

There is no clear-cut evidence to separately portray the reading ability of the Sri Lankan undergraduates in the literature. The G.C.E O/L and A/L English language test papers include a certain proportion of reading items. Further, the ESL testing papers in most universities do not have separate test papers for reading skills.

However, the results of the UTEL examination conducted in 2015 nationwide, which implicate reading ability individually, are not open to everyone; therefore, it is difficult to ascertain the level of undergraduates' reading skills.

However, a few studies might help to determine the level of reading ability of these students. Among such studies, the research conducted at the Faculty of Management, Social Sciences and Humanities (FMSH) of General Sir John Kotelawala Defence University (KDU), showed that the undergraduates have successfully performed in reading skills (87%) at the predicted ILOs in UTEL-A benchmark 5 (Jayasinghe & Wijethunge, 2015).

Another study was carried out to evaluate and improve the basic reading comprehension abilities of first-year engineering undergraduates enrolled in an EAP course at the Sri Lanka Institute of Information Technology (SLIIT) by Dissanayake (2018). Although the results were not aligned with any benchmarking, the basic reading skills such as skimming, scanning, and vocabulary, were assessed in the test and the results indicated poor attainment in reading skills.

Nevertheless, a graduate should possess at least a UTEL band score of level 5 by the time he/she graduates (Abeywickrama, 2020; Jayasinghe & Wijethunge, 2015; Ratwatte, 2016; Wikramanayake et al., 2012). The UTEL, which is a locally accepted, and internationally compatible benchmark for measuring English language competencies (Jayasinghe & Wijethunge, 2015; Kulasingham et al., 2012), has been utilised to monitor and evaluate the English language skills of undergraduates. Therefore, a UTEL band score of level 5 can be considered equivalent to the CEFR upper B1 level or the First B2 level. As per the aforesaid discussion, this level is required for the successful attainment of students' reading skills too. These levels of reading skills are targeted at the universities in their ESL teaching instructions as per the direction of the University Grant Commission, Sri Lanka.

2.3 THEORETICAL FRAMEWORK

Several theories, frameworks and models related to reading, assessing reading, and test validation are investigated and discussed in the following sections.

2.3.1 Reading

Reading is a complex multifaceted construct (Carrell et al., 2000; Grabe, 1991; Koda, 2005; Mckee, 2012). It has always been “a key element in university study” (Moore et al., 2012, p. 4). Taylor (2009), for example, regarded that in the university, whether writing essays, discussing ideas in tutorials and seminars, interacting with the content of lectures or enquiring into the connections between the book and the syllabus, most students view these as originating from a “conversation” with one’s readings in a discipline (Taylor, p 54). Reading is a crucial skill for academic success. Carrell et al. (2000, p. 1) reiterated the same point that “reading is by far the most important of the four skills in a second language”, and reading is the main reason why students learn the language. Aebersold and Field (1997) stated that reading, as a receptive skill, has long been regarded as a prerequisite for foreign language acquisition since it functions as an essential source of input for other skills to develop. If a person’s reading skill is enhanced, it is expected to improve their listening, speaking, and writing skills.

Primarily, English is used as a medium of instruction in higher education (Kedzierski, 2016; Kirkpatrick, 2011). For academic purposes, English is widely used in teaching and learning in universities, colleges, or tertiary educational institutions. To master English, reading is essential (Carrell et al., 2000). Similar views have been shared by Grabe and Stroller (2011) that, in the setting of English as an L2, reading continues to be exceedingly important. English is pervasively the language of science, technology, and advanced research; students are urged to read at a high level of proficiency to achieve notable personal, occupational, and professional goals.

Knowledge is gained through reading. Thus, it is the most fundamental of the four language skills (Carrell et al., 2000; Grabe, 2009; Li & Wilhelm, 2008; Mermelstein, 2015). Bachman, in his preface to *Assessing Reading*, discloses that:

Reading, through which we can access worlds of ideas and feeling, as well as the knowledge of the ages and visions of the future, is at once the most extensively researched and the most enigmatic of the so-called language skills. (Alderson, 2000, p. x)

In other words, reading is very crucial in order to fathom worldly matters, and as such, it is often considered to be the most significant skill.

Reading is necessary for building vocabulary and is indispensable for long-lasting education and improvement in first and second language skills. Thus, mastery of reading is obligatory in school to ensure successful learning in any subject (Nuttall, 1996). If a child is poor in reading ability, his performance in school life is inhibited (National Reading Panel, 2000). Similarly, having a good command of reading is pivotal for adolescents, too, as Moore et al., cited in Vacca (2002) emphasised: “Adolescents entering the adult world in the 21st century will read and write more than at any other time in human history... In a complex and sometimes even dangerous world, their ability to read will be crucial. Continual instruction beyond the early grades is needed” (p. 7).

In this connection, reading can be regarded as systematically mapping the visual language skill onto the spoken language skill (Mann, 1984). Therefore, learning to read is a complex matter, and that is why some children perform well whereas others find it difficult. On the other hand, it can be argued that reading disability may arise at any level from visual sensitivity to general cognition.

2.3.1.1 Reading Comprehension Theories

Grabe (2009) illustrated that “reading is a strategic process in that a number of the skills and processes are needed on the part of the reader to anticipate text information, select key information, organize and mentally summarize information, monitor comprehension, repair comprehension breakdowns, and match comprehension output to reader goals” (p. 15). In the late 1990s, the RAND Reading Study Group (RRSG) defined reading comprehension “as the process of simultaneously extracting and

constructing meaning through interaction and involvement with written language” (Snow, 2002, p. xiii).

Li and Munby (1996, p. 199) stated that “reading comprehension is not just understanding words, sentences, or even texts, but involves a complex integration of the reader's prior knowledge, language proficiency and their metacognitive strategies”.

Veeravagu et al. (2010, p. 206) defined reading comprehension as:

a thinking process by which a reader selects facts, information or ideas from printed materials; determines the meanings the author intended to transmit; decides how they relate to previous knowledge and judges their appropriateness and worth for meeting the learner’s own needs and objectives.

However, reading comprehension skills are much more complex than this definition suggests (Grabe & Stoller, 2011; Mckee, 2012). This is further confirmed by Pearson and Johnson (1978), mentioning that reading comprehension involves multiple levels whereby each level has several cognitive demands on readers. However, reading comprehension is viewed through three different theories. The first is the independent skills theory, which mentions that reading comprehension includes different processes that can be learnt independently from each other in any order. The second is the global skills theory, depicting reading comprehension as a single or general unitary process, which, after being learned, will enable learners to answer any kind of comprehension questions about a given passage (Chapman, 1969, pp. 6–7). The third is the hierarchical skills theory, which asserts that reading skills can be arranged into levels according to the complexity of the behavior necessary to learn each skill (Chapman, p.9).

2.3.1.2 Reading Comprehension Models

As mentioned in the earlier sections, reading skills appear to be complex in nature. There are different reading comprehension models purporting to explain its complexity, attesting to the distressingly complicated nature of this phenomenon. As a result, a plethora of reading skill models have been offered and conceptualised as

refinements of previous ones in terms of components (product) and interrelationships (process) (Urquhart & Weir, 1998).

2.3.1.2.1 Reading as a Process

Alderson (2000, p. 3) mentioned that the process of reading is “the interaction between a reader and the text”. He elaborated that understanding the process of reading helps in understanding the nature of reading. The bottom up, the top down, and the interactive approaches are the commonly known approaches according to

Urquhart and Weir (1998) and Grabe (1991); additionally, however, the interactive-compensatory model, the situation model, the construction-integration model, and the structure building model, are also identified as reading models (Davoudi & Moghadam, 2015).

Goodman (1967), as cited in Urquhart and Weir (1998), proposed the top-down or conceptually driven processing approach, which is one of the most widely used reading skill models. Reading, in his opinion, is a psycholinguistic guessing game, since the readers' expectations and prior knowledge have a significant impact on lower-level processes like orthographic and phonological processing, as well as word recognition ability. This approach is focused on higher level cognitive processes. On the other hand, the bottom-up approach focuses on the lower-order processes. This model presumes that reading is a hierarchical and sequential process that begins with the perception of individual phonemes and progresses through words, phrases, sentences, and finally the entire piece of discourse (Rieben & Perfetti, 2013).

These two techniques, on the other hand, have drawn a lot of criticism, primarily because of their simplistic conceptions of reading. Further, these approaches considered reading to be a linear activity involving only one direction, while studies have shown that it is a more complex and difficult process. Rumelhart (1977) introduced the interactive model, whereby reading is regarded as a process that incorporates both bottom-up and top-down approaches. In practice, information gathered by the eyes is visually registered before being transferred to the pattern synthesiser. Simultaneously, a large amount of information concerning semantic,

syntactic, and pragmatic concepts is retrieved from long-term memory and stored in working memory. However, this model was found to be unpredictable by Alderson (2000).

2.3.1.2.2 Reading as a Product

Alderson (2000) pinpointed that the result of the reading process is a product. The product approach is often known as the componential model, comprising the components involved in the product of reading. Reading comprehension as a process differs from reading comprehension as a product in the sense that, as a process, it is a vague and ill-defined entity. After all, it entails the mental process that occurs when readers read a text, whereas, as a product, it is less nebulous because it deals with measuring the nature of the process and evaluating responses to specific test items. Inferences regarding the reading processes of readers are drawn from there (Pearson & Johnson, 1978). Urquhart and Weir (1998) studied the components involved in reading and identified word recognition, vocabulary, meta-cognition, and background knowledge as the commonly recognized components.

Grabe (1991, p. 382-3) noted that “A ‘reading components’ perspective is an appropriate research direction to the extent that such an approach leads to important insights into the reading process...the component skills approach, at least in the broad sense outlined here, is indeed a useful approach”.

Although the reading taxonomies developed by Munby (1978) and Vacca and Vacca (2008) viewed reading as a product approach, Alderson (2000, p.5), however, mentioned that this approach is regarded as “unfashionable in recent years as research efforts have concentrated on understanding the reading process, and as teachers of reading have endeavoured to improve the way in which their students approach text”.

2.3.1.3 The Nature of Reading

Several writers (Carrell et al., 2000; Grabe, 2009) emphasised the importance of reading in English for L2 and academic purposes, despite the fact that most earlier research focused on reading for L1 learners. Reading is not something that can be observed directly. There is a lot of disagreement over the nature of reading (Hubley, 2012). Although the nature of the reading construct is given more attention than the mechanism of reading assessment (Urquhart & Weir, 1998), there is a lack of agreement in the literature on how to define the nature of reading: whether it is an indivisible, unitary ability, or a multi-divisible skill.

...there is a considerable degree of controversy in the theory of reading over whether it is possible to label separate skills of reading. Thus, it is unclear (a) whether separable skills exist, and (b) what such skills might consist of and how they might be classified (as well as acquired, taught and tested). (Alderson, 2000, p. 10).

Reading skills are currently divided into two categories: reading as an 'indivisible' or 'unitary' skill and reading as a 'multi-divisible' skill (Alderson, 2005; Weir & Porter, 1994). However, according to Weir and Porter (1994), a 'bi-divisible' view of reading is also conceivable. Because they mentioned that 'vocabulary' appears to be a separate component from reading comprehension, as indicated by several quantitative studies.

2.3.1.3.1 Reading as a Unitary Skill

Reading as a unitary skill means that reading does not have clear separable and identifiable sub-skills or components. Thus it is a single global construct, which is unidimensional (Alderson, 2005; Alderson & Lukmani, 1989; Bachman, 1990; Lunzer et al., 1979; Weir & Porter, 1994).

Several studies proved that reading is a unitary skill. For example, Lunzer et al. (1979), did a study on elementary English native speakers and discovered that there is little evidence to support the idea that reading ability may be divided into several sub-skills.

Furthermore, another study performed by van Steensel et al. (2013) involving seventh-grade low-achieving kids, using CFA, identified a similar trend in which just one underlying skill was discovered. In another study, Rosenshine (2017), for example, attempted to explain whether statements about the prevalence of sub-skills in L1 reading may be substantiated. He concluded that reading is a unitary skill after reviewing numerous sources, including elementary reading textbooks, authoritative sources of reading skills, factor analytic research, and items developed for children.

Alderson's (2005) research on 718 readers from various European countries, applying the DIALANG test, found that there was just one component that accounted for between 68 and 74 percent of the variance in reading. The findings indicated that reading is a single ability.

In a recent research involving eight Turkish ESL undergraduates, applying the eye-tracking movement approach to investigate the nature of reading, Kahraman (2019, p. 206) rejected the “multi-divisible view of L2 reading construct”. All these studies supported the unitary nature of reading skills.

2.3.1.3.2 Reading as a Multidimensional Skill

Reading research has also backed up the idea that reading is a multi-divisible construct with distinct and identifiable sub-skills (Farhadi & Hessami, 2005; Grabe, 1991; Hughes, 1989; Khalifa & Weir, 2009; Kim, 2020; Munby, 1978; Weir et al., 1990). Weir and Porter (1994) identified a bi-divisible view of reading skills. Similarly, Urquhart and Weir (1998) mentioned that as 'vocabulary' could be a separate component from reading comprehension, a bi-divisible reading perspective appears to be acceptable. However, many studies observed that reading is a multi-divisible skill.

Hughes (1989, p.116) distinguished between "macro skills" and "micro skills" of reading comprehension. The term "macro skills" refers to the ability to grasp the main points of the text. Scanning, skimming, detecting stages of an argument, and identifying instances offered in support of an argument are all examples of these talents. “Micro skills”, on the other hand, pertain to the ability to recognise and

analyse the linguistic elements of the text. Micro skills are taught as enablers to help macro skills improve. Micro skills include guessing the meaning of unknown words using context, detecting pronoun referents, and grasping the relationship between parts of the text by recognising indicators in discourse.

Another study comprising 1606 Iranian learners, both English majors and non-English majors, was undertaken to improve the construct validity of L2 reading comprehension exams, and to examine the multi-divisible nature of reading by Farhady and Hessamy (2005). The findings revealed that L2 reading ability comprised several latent qualities or macro-skills. Although this study was conducted with EFL students, the results were similar to those obtained with ESL students.

Andrich and Godfrey (1978) investigated the existence of multi-divisible ability of reading by administering the Davis' Reading Skills Test, Form D. The findings of 188 subjects, spanning from 9th grade to first-year tertiary students, revealed that 76 of the 96 questions on the test are consistent with the Rasch latent trait model and those items were evaluating several sub-skills of reading which has a hierarchical order.

Alderson and Lukmani (1989), for example, studied non-native speakers to discern between lower and higher order skills. Teachers at Bombay University were given a 41-item reading exam to rate the items as lower, middle or higher order abilities. Only 14 items achieved concurrence. Furthermore, when these 14 items were evaluated on 100 Bombay University students, the analyses revealed that the item difficulty estimations did not match the expected sequence of the skills involved. Higher order items performed slightly better for the weaker pupils than lower order items, and vice versa. Although reading skills have multiple divisible skills, the hierarchy of these skills cannot be clearly defined so far.

2.3.1.4 Reading in the Second language

Reading in L2 is a gateway to enhancing the other skills to be succeeded in a particular language. Anderson (1999, p.1) highlighted that:

Reading is an essential skill for English as a second/foreign language (ESL/EFL) students; and for many, reading is the most important skill to master. With strengthened reading skills, ESL/EFL readers will make greater progress and attain greater development in all academic areas.

Similarly, Mikulecky (2008) mentioned that reading is the key to acquiring a second language, which means that reading is the most significant fundamental instruction in all aspects of language learning. Carrell (1989), as well, affirmed that “For many students, reading is by far the most important of the four skills in a second language, particularly in English as a second or foreign language.” (p. 1).

Grabe (2009) illustrated that one of the most significant skills required of people in multilingual and international situations is the ability to read in an L2. It is also one of the hardest skills to hone to a high level of mastery. A modern understanding of reading necessitates consideration of six fundamental issues like different reading objectives, reading definition requirements, processes that underpin reading as a separate skill, institutional and social context influence on L2 reading, recognizing challenges important to L2 reading training and determining unique features of L2 reading (as opposed to L1 reading), and improving instruction and student learning by applying L2 research implications.

According to Aebersold and Field (1997) as cited in Hoang (2016) “Reading, as a receptive skill, has long been regarded as a prerequisite for foreign language acquisition since it functions as an essential source of input for other skills to develop” (p. 5). It is obvious that improving one’s reading activity can develop their writing and speaking skills. In other words, students who are good readers can write well, and improve their vocabulary, and linguistic skills (Hafiz & Tudor, 1989). On the contrary, Hudson (2007, p. 74) remarked that “The studies are fairly consistent in showing that learners with very little exposure to the second language have difficulty in reading.”

Learning to read a second language includes two, or more languages as its phrasing suggests, and in developing L2 reading both L1 and L2 characteristics are explored under individual differences (Koda, 2005). Further, this idea is elaborated by Brown (2001) who stated that “For most second language learners who are already literate in a previous language, reading comprehension is primarily a matter of developing appropriate, efficient comprehension strategies” (p. 291). He suggested that both top-down and bottom-up strategies may need to be emphasized depending on individual needs and proficiency levels.

2.3.1.5 Levels of Reading Comprehension

From some of the above definitions, it can be understood that reading comprehension relates to the understanding and thinking process involved in getting the message that the writer wants to share with the reader from the reading materials. Thus, reading comprehension is developed at multiple levels. A reader exerts different cognitive processes at each level of comprehension (Pearson & Johnson, 1978). Word recognition is at the lower level of reading comprehension whereas the ability to understand the main ideas and to make inferences are assumed to be higher level skills (Hudson, 2007; Khalifa & Weir, 2009; Urquhart & Weir, 1998).

The levels of reading comprehension have been identified differently by many researchers. Hosenfeld (1977) researched twenty ESL successful and twenty unsuccessful readers using the main-meaning line, and word-solving strategies, and observed that poor readers focused more on solving the unknown words, while successful readers kept the meaning of the entire passage in mind. Alderson (1991) mentioned that there are three levels of comprehension, namely, understanding main ideas, understanding direct statements, and drawing inferences. Basaraba et al. (2013), meanwhile, listed literal comprehension, inferential comprehension, and evaluative comprehension as the three levels, and among them, literal is less challenging than the others.

2.3.1.6 Reading skills, sub-skills and strategies

Sometimes the term “skills” can be interchangeably used by scholars for “levels of understanding”. “A reading skill can be described roughly as a cognitive ability which a person is able to use when interacting with written texts” (Urquhart & Weir, 1998, p. 88), which is similar to Pearson and Johnson’s (1978) understanding of levels of reading.

Notably, providing the relationship between skills and strategies is significantly crucial at this juncture. Both academic literature and instructional materials have significant terminological confusion when it comes to the terms “skill” and “strategy”. For instance, “inferring” is a skill for Davis (1968), and Khalifa and Weir (2009), but it is a strategy for Olshavsky (1977), and Grabe and Stoller (2011). Nuttall (1985) and Grabe (1991) mentioned that skills and strategies are synonymous, while they are used interchangeably in Davies and Whitney (1981), Taylor et al. (1986), and Maingay (1983) as cited in Williams and Moran (1989).

Urquhart and Weir (1998) cited several publications issued between 1966 and 1996 that mention “skills” and “strategies” in their title in Eric’s index, or abstract. They stated that the term “strategy” became popular in the 1980s. Skills refer to information processing techniques that are automated and applied to a text unconsciously, whereas strategies refer to actions selected intentionally to achieve specific goals (Paris et al., 1991). Koda (2005, p. 205) defined strategies as “deliberate, goal/problem-oriented, and reader-initiated/controlled”. Further, on pages 96-98 Urquhart and Weir (1998) clearly explained the differences between these two terms, which correspond to the description given by Koda and Paris et al.

However, the distinction provided by Williams and Moran (1989, p. 223) adapting Olshavsky (1977), is highly significant to note here, that “a skill is an acquired ability, which has been automatized and operates largely subconsciously, whereas a strategy is a conscious procedure carried out in order to solve a problem”.

While researching “skills” further, Alderson (2000, p.9) mentioned that “reading skills or abilities” were identified by researchers “to test the different levels of understanding of the passages”. Vincent (1985, cited in Urquhart & Weir, 1998)

stated that skills are recommended as necessary for structuring reading syllabi. Many lists of skills and taxonomies have been developed in such a fashion (Alderson, 2000; Hedgcock & Ferris, 2009; Urquhart & Weir, 1998; Williams & Moran, 1989). As Alderson (2000, p. 11) noted that the taxonomies “are potentially very powerful frameworks for test construction and will doubtless continue to be used”. Though these taxonomies are arguably provisional, the following taxonomies, which were empirically tested, are fairly justifiable according to Williams and Moran (1989) and Urquhart and Weir (1998).

2.3.1.6.1 Davis's (1968) Taxonomy

1. Identifying word meaning
2. Drawing inferences
3. Identifying the writer's technique and recognising the mood of the passage
4. Finding answers to questions

Although Alderson (2000, p. 9) mentioned that Davis' taxonomy has “eight skills”, actually “Davis identified four skills” (Williams & Moran, 1989, p. 223), and the fourth skill in his taxonomy is criticised by Urquhart and Weir (1998), and Williams and Moran (1989). However, these skills were the base shared by other researchers in the future (Alderson).

2.3.1.6.2 Munby's (1978) Taxonomy

Munby's taxonomy of language has a great influence on L2, especially on second language education curriculum, materials design, and language test design (Alderson, 2000; Alderson & Lukmani, 1989; Hudson, 2007). His taxonomy focuses on all language elements, such as reading, writing, speaking, and listening, and it includes a large range of skills and sub-skills that can be used as examples in a skill-based curriculum. Munby's (1978) micro-skill taxonomy includes two hundred and sixty sub-skills in fifty-four groups; however, this taxonomy did not support the hierarchical

approach of reading comprehension (Urquhart & Weir, 1998). Though there are some critics of this taxonomy, it is still applied by some international testing agencies. The following 19 ‘microskills’ are crucial to be noted.

1. Recognizing the script of a language
2. Deducing meaning and use of unfamiliar lexical items
3. Understanding explicitly stated information
4. Understanding information when not explicitly state
5. Understanding conceptual meaning
6. Understanding the communicative value of sentences and utterances
7. Understanding relations within the sentence
8. Understanding relations between parts of a text through lexical cohesion devices
9. Understanding cohesion between parts of a text through grammatical cohesion devices
10. Interpreting text by going outside it
11. Recognizing indicators in discourse
12. Identifying the main point of information in a piece of discourse
13. Distinguishing the main idea from supporting details
14. Extracting salient points to summarize
15. Selective extraction of relevant points from a text
16. Basic reference skills
17. Skimming
18. Scanning to locate specifically required information
19. Trans-coding information presented in diagrammatic display

2.3.1.6.3 Lunzer's et al. (1979) Taxonomy

1. Word meaning
2. Words in context
3. Literal comprehension
4. Drawing inferences from single strings
5. Drawing inferences from multiple strings
6. Interpretation of metaphor
7. Finding salient or main ideas
8. Forming judgements

Lunzer et al. (1979) administered a test to 257 English primary school children. While no two lists of reading skills are similar, a cursory review shows that the skills can be divided into two categories: “language related”, and “reason related”. (Williams & Moran, 1989, p. 223). Urquhart and Weir (1998) justified the fourth and fifth skills in this taxonomy and claimed that this taxonomy is hierarchically arranged.

2.3.1.6.4 Hillock's (1980) Taxonomy

Hillock (1980) as cited in Hillocks and Ludlow (1984), applied two levels of comprehension in his research, namely inferential and literal levels. “All literal level skills require identification of information that appears explicitly in the text” (Hillocks & Ludlow, p.8).

1. Literal level of comprehension:
 - a. Basic Stated Information (BSI)
 - b. Key Detail (KD)
 - c. Stated Relationship (SR)

2. Inferential level of comprehension:
 - a. Simple Implied Relationship (SIR)
 - b. Complex Implied Relationship (CIR)
 - c. Author's Generalisation (AG)
 - d. Structural Generalisation (SG)

This taxonomy has been empirically validated by Tian (1991) and Hillocks and Ludlow (1984).

2.3.1.6.5 Grabe's (1991) Taxonomy

According to Grabe (1991), many researchers have categorised the reading process into a collection of component skills and knowledge areas, to better understand this dynamic process. He believes that many good readers have been reported to automatically become engaged while applying this interactive process. He developed the following six-component reading process:

1. Automatic recognition skills
2. Vocabulary and structural knowledge
3. Formal discourse structure knowledge
4. Content/world background knowledge
5. Synthesis and evaluation skills/strategies
6. Meta-cognitive knowledge and skills monitoring

Alderson (2000) highlights the metacognitive skills applied by Grabe consisting of recognition of more important information in the text; using context to solve a misinterpretation; adjusting reading rate; skimming; previewing headings, pictures, and summaries; finding specific information; using a dictionary; formulating questions about the information; using word formation and affix information to guess

word meanings; taking notes; underlining; summarizing information; and so on. Nevertheless, Urquhart and Weir (1998) criticised Grabe's taxonomy using very general categories, while William and Moran (1989) selected this as one of the justifiable taxonomies.

Though there are many skills lists and taxonomies, as Hedgcock and Ferris warned, "they should not be used wholesale as prescriptions for instructional design" (Hedgcock & Ferris, 2009, p. 38); furthermore, there are many questions to ask about the taxonomies (Urquhart & Weir, 1998). Many of these taxonomies are outside empirical validations. However, these taxonomies may be used in test designing as they are powerful frameworks for test development (Alderson, 2000).

2.3.1.7 Reading Construct

The term "construct" coupled with reading emerged after the 2000s. The Scopus database search engine provided only 26 documents for the search term "reading construct" (as seen on 02.04.2021). Before 2000, only one research study was found focusing on "Reading and television" by Roberts et al. (1984), while all the other studies were carried out after 2000. Fulcher and Davidson (2007) differentiated the terms 'model', 'framework', and 'construct' as well; they mentioned that 'model' is the larger unit of all three, and 'construct' is the description of the components of a model (Fulcher & Davidson, p. 36). Alderson (2000, p. 118) stated that "A construct is a psychological concept, which derives from a theory of the ability to be tested. Constructs are the main components of the theory, and the relationship between these components is also specified by the theory". For Messick (1989) constructs were defined as the underlined psychological abilities produced in an experimental setting.

The constructs of reading are based on a model of reading and the factors influencing reading that are relevant to the assessment of reading constructs (Alderson, 2000). Alderson et al. (1995) noted that some theories of reading state that there are various constructs involved in reading and those constructs vary from one another. In other words, skimming, scanning, synthesising, and evaluating skills are parts of a theoretical construct of reading, and they are different skills when assessing,

and they must be operationalized differently. Messick (1996, p. 252) added that the validity of tests can be influenced by an inadequate sampling of the construct or “construct-irrelevant difficulty” or variance. This means that improper constructs can have a negative washback from the test to the teaching and learning, as teachers might ignore important constructs if they are not included in the test.

Alderson (2000) argued that the understanding of reading is essential to the development of assessment instruments and emphasized the importance of adopting a model of reading while constructing such instruments. He focused on reader and text variables and emphasised that any variable that impacts the reading process or its product should be considered when validating a test design. So the construct of reading is influenced by text type, item format, text language, text content/topic, as well as the reader’s schemata and background knowledge, physical, and psychological characteristics, linguistic capabilities, and interests/motivation, which are the main constructs to consider while designing a test (Alderson, 2000; Carrell et al., 2000; Hudson, 2007; Khalifa & Weir, 2009; Koda, 2005; Urquhart & Weir, 1998)

In this regard, test specifications are pivotal in displaying the theoretical framework underlying the test. They explain the constructs of the test and the relationship among the constructs; in other words, they provide the link between the theoretical and operational definitions. However, designing test specifications is not an easy matter. In showing the impact of test specifications on test design, Alderson (2000) exemplified the test specifications of DIALANG, FCE, IELTS as well as CEFR, and mentioned that researchers are still uncertain of which variables affect construct validity. In addition, the influence of the assessment scales is too conspicuous in the reading construct. Commonly, scales are more productive in measuring the performance skills like speaking and listening. It is predicted that these scales can help in the assessment of reading as well, because they can be related to actual performances, salient features, and interlanguage development processes (Alderson, 2000; North, 2014).

North and Schneider (1998) were the first to relate the construct of reading to communicative language ability. They attempted to integrate four concepts of reading constructs, namely strategic, linguistic, discourse, and sociolinguistic competence.

Alderson (2000, p 120) suggested adopting “a particular model of reading” to construct test designs, hence, it is crucial to examine a particular model. For this purpose, Khalifa and Weir’s (2009) model of reading is illustrated in detail in the later sections.

2.3.1.7.1 Khalifa and Weir (2009)

Weir’s (2005) socio-cognitive framework, designed for all four skills, was modified by Khalifa and Weir (Khalifa & Weir, 2009) to examine reading skill. They introduced a reading model, which includes both low and high-order cognitive processes of reading. Specifically, eight cognitive processes were introduced, such as Word Recognition (WR), Lexical Access (LA), Syntactic Parsing (SP), Establishing Propositional Meaning (EPM), Inferencing (I), Building a Mental Model (BMM), Creating a Text Level Structure (CTLS), and Creating an inter-textual representation (CITR). Khalifa and Weir’s model for reading has been empirically validated by some researchers (Bannur et al., 2015; Dunlea, 2015; Wu, 2011). However, collecting more evidence for empirical validation is highly crucial as Weir (2005) urged further research on his framework. As this framework has been validated by adequate research findings, this study chooses this framework to develop its test materials.

2.3.1.7.2 Robinson’s (1941) SQ3R method

SQ3R was one of the successful reading methods introduced by Robinson (1941). The abbreviation: SQ3R stands for Survey, Questions, Read, Recite, and Review. Through this method, a broad idea of what we read can be easily understood (Thamburaj et al., 2021). “Survey” gives the overall idea of the text. The reader needs to create a “Question” to be able to answer as they read. Then, to get the answer to the question, the reader “Reads” the selected passage. “Recite” is the most important part, which recalls the answers to the questions in each section of the reading passage. “Review” is the process of going back to the text again to re-examine the answer concerning all relevant information in the text (Johns & McNamara, 1980, p. 705). Although Thamburaj et al. (2021) pinpointed that this method focuses highly on the cognitive

process of readers, it is a rather appropriate method for textbook and assignment reading (Huber, 2004). It was further proven in a survey conducted among Tamil L1 learners to show improvement in reading the Form 4 Tamil textbook (Thamburaj et al., 2021). Moreover, since there were no details of Robinson's experiment and no empirical evidence, this model is viewed as an opinion rather than research (Johns & McNamara, 1980).

2.3.1.8 Academic Reading constructs

Achieving a higher level of ability would lead to better comprehension resulting in academic success. Unlike simple reading, academic reading necessitates in-depth comprehension, which is frequently linked to the need to complete certain cognitive and procedural activities, such as taking a test, writing a paper, or presenting a speech (Li & Munby, 1996). As practically all authentic writings are created for native speaking readers, the fundamental challenge in academic reading for most second language readers will simply be the gap between what they know and what native speakers know. Adult academic second language readers, especially those with extensive understanding of the language, nonetheless, have identification problems that interfere with their attempts to comprehend the texts they must read, despite their higher-level skills

In order to have a clearer understanding of learners' academic reading, it is important to learn their conceptions. There have been very few attempts to investigate conceptions, perceptions, or perspectives on academic English reading. The studies are, for example, Hooley et al. (2013) and Ohata and Fukao (2014). Hooley et al. (2013) investigated high school students' academic reading perceptions and their link to reading proficiency. They investigated academic reading perceptions in terms of class reading, teacher support, students' understanding, etc. However, the findings concern perceptions of high school reading, which might not be as academic as university-level reading. Furthermore, Ohata and Fukao (2014) also examine conceptions of academic reading. Although the participants were college students, the main focus of their study tended to be on how such conceptions are constructed and developed.

There is a close relationship between EAP (English for Academic Purposes) competency and academic achievement, especially in the Sri Lankan context (Dissanayake & Harun, 2012). Having this in mind, the present study was then aimed at investigating ESL learners' conceptions of academic reading in terms of how they conceived academic reading, what difficulties they had, and what they needed for university reading.

2.3.2 Assessing Reading

Reading comprehension is a complex multifaceted process, and assessing it is also a challenging effort. Assessing reading has been a problem to test developers, material designers, and teachers (Mckee, 2012). Many researchers have focused on the variables affecting the testing of reading comprehension. A better understanding of reading is a prerequisite for its assessment. Assessment helps teachers determine whether the instruction given is resulting in satisfactory student progress. It assists the teachers to recognise the academic levels of the students and the levels they need to achieve.

Alderson (2000), in his book *Assessing Reading*, focused the research on the assessment of reading in chapter three. He discussed variables affecting reading, and related test methods: their validity, reliability, and factors affecting their use as one of the major areas of assessing reading research. The difficulty of the reading test is dependent on both passage and item difficulty (Alderson). Based on the reading definitions, models, taxonomies, etc., test developers, educators, psycholinguists, applied linguists, and researchers design reading tests to measure what they intend to measure in the test.

2.3.2.1 Text type /Genre/ Purpose

Khalifa and Weir (2009) included “test-taker characteristics”, and “context validity” as two different features out of their six validation types of the socio-cognitive validation framework. According to Alderson (2000), the reader variable, as well as

the text variable, are two crucial factors influencing the assessment of reading constructs, which he clearly explained in chapter two. “Overall text purpose” mentioned by Khalifa and Weir (2009), “Text type and genre” (Alderson, 2000, p. 63), or “Text types” (Urquhart & Weir, 1998, p. 83), is an element; included under “Context validity”, or “Text variables” (respectively, discussed by the authors of the above two citations), affects the difficulty of reading assessment (Weige (2002) cited in Khalifa and Weir (2009, p. 105) provides five purposes of reading, for instance: “*referential* (intended to inform), *conative* (intended to persuade or convince), *emotive* (intended to convey feelings or emotions), *poetic* (intended to entertain, delight, please), and *phatic* (intended to keep in touch)”. Cambridge ESOL Main Suite practices mainly referential texts for all CEFR level tests while including a proportion of poetic, emotive, and conative texts according to its level.

Nonetheless, according to Weige (2002, p.62) cited in Khalifa and Weir 2019, “genre” refers to:

the expected form of communicative function of the written product; for example, a letter, an essay, or a laboratory report. The rhetorical task is broadly defined as one of the traditional discourse models of narration, description, exposition, and argument/persuasion, as specified in the prompt, while the pattern of exposition (Hale et al. 1996) refers to subcategories of expositions or specific instructions to test takers to make comprehensions, outline causes and effect and so on.

However, generally, in reading assessment research, based on the structure, style, purpose, and discourse mode, texts are broadly divided into narrative and expository types (Weaver & Kintsch, 1991). Expository texts provide information about a particular topic. As can be seen in the narrative texts, there is no specified structure followed in expository writing. However, narrative texts are written in the form of stories using temporal sequence, using the past tense, and making use of common everyday vocabulary (Medina & Pilonieta, 2006). “One interesting feature of narrative texts, in particular, is that they appear to include visualisation in the reader as part of the reading process” (Alderson, 2000, p. 64).

There have been research studies done on the influence of both text types on L1 reading and L2 reading. In the context of L1 reading, Cervetti et al. (2009) claim that the text type influences the reading performance, expressly expository passages were able to gain more accurate answers while Wilson's findings suggest that comprehension was not motivated by text type; however, reading ability, and text topic knowledge matter. In the L2 context, many studies came out with the idea that expository texts are more challenging than narrative texts (Alderson, 2000; Berkowitz & Taylor, 1981; Eason et al., 2012; Ebibi, 2014).

Kobayashi (2002) looked at the relationship between student test performance and test type and test format. The findings of the study show that test format and test type had a significant effect on students' performance as well, as well-structured texts made it easier to differentiate between students with different levels of proficiency. He concluded that reading performance is influenced by text types or text organization (Kobayashi, 2009). Eason et al.'s (2012) results did not show a significant variation in students' reading performance; it indicated that expository texts require more high level cognitive skills. Alderson (2000, p. 64) mentioned that "There is a long tradition of research into the differences between expository and narrative texts. The general conclusion is that expository texts are harder to process than narrative texts". This current research focuses only on expository texts, as expository texts mainly deal with information, argumentation, exposition, and description, which are needed for academic success, in coping with lectures, assignments, presentations, and evaluations.

2.3.2.2 Test format / Response Type / Type of input/ Item format

The main purpose of reading is comprehension. Comprehension is assessed through the student's ability to recall the details of what they have read (Allington, 2001). As comprehension is an unobserved behaviour, assessing reading comprehension can be made through the careful application of testing techniques. Alderson (2000, p: 202), more or less, conflated the terms "test method", "test technique" and "test format", as the testing literature does not mention the possible differences between them. Test

format (Alderson, 2000) or item format (Jusoh, 2018), or response method (Khalifa & Weir, 2009) are interchangeably used.

In the literature, it has been identified that test-takers with different abilities and skills may be affected by a test that is different from the ones commonly used. Further, test formats may be fair to some of the test-takers while some formats are challenging (Kunnan, 2013). A good test should be fair to all and should not confuse other students because of different test formats (Elder, 1998). Masoumi and Sadeghi (2020), Kastner and Stangl (2011), Shohamy (1984), and Wolf (1993), claim that the item format serves as a potential contributor to reading comprehension difficulty. “Selected response” (SR) or “constructed response” (CR) are identified as the item formats (Khalifa & Weir, 2009, p.83).

Selected Response Items

In SR, “the candidate chooses the answer from a set of options provided at the word, phrase, sentence or paragraph level”, whereas in CR “candidates have to produce the answer themselves” (Khalifa & Weir, 2009, p.83). Multiple Choice Questions (MCQ), true/false items, right/wrong/ doesn't say item, and gapped text, are included in the classification of SR by Khalifa and Weir (2009). In MCQ, there is a stem that represents the question as a problem to be solved with given alternative responses. The correct response is known as the “key”, and the wrong responses are “distractors” (Kastner & Stangl, 2011, p. 265). In an MCQ, there may be one key with two to seven distractors (Ebel & Frisbie, 1972). Munby (1968) provided a detailed illustration of this responding method on pages xiv to xxii.

Constructed Response

A “Constructed response format” (CR) includes “short answer questions (SAQ)”, “cloze” and “gap filling”, “information transfer”, and “reading into writing” types (Khalifa & Weir, 2009, pp.87-91). One of the most common arguments for using CR in exams is that it checks a deeper understanding of the content (Bacon, 2003; Rogers

& Harley, 1999). Furthermore, CR tests are the best format for making practical decisions and representing fluctuating social values (Katz et al., 2000), and motivating students to evaluate problems critically (Rotfeld, 1998).

SR vs CR

Even though Khalifa and Weir (2009) apply both CR and SR to examine reading ability, claiming that CR is more suitable to assess higher-order thinking skills, this response format has been criticised by many scholars, as only comparatively few questions can be included in this response format; in other words, all taught material is not covered (Ventouras et al., 2010).

As the CR format involves some aspects of writing skills, it can influence the measurement of the intended reading construct. Another concern that Zeidner (1987) mentioned is that students with low writing abilities are underprivileged, even though their content knowledge is superior. Powell and Gillespie (1990) and Downing and Haladyna (2006) criticised that the marking of CR exams has additional disadvantages despite defined scoring criteria; grading appears to be more subjective and inconsistent. Furthermore, marking CR exams takes time (Ventouras et al., 2010), and computerized assessment of these answers still remains challenging.

On the other hand, SR items, especially MCQs, are cost-effective (Bennett et al., 1990) and easy to mark, as Powell and Gillespie (1990, p. 1) mentioned, “scoring is much easier”. Exams are graded uniformly, so there are no ranking biases, which eliminates the need for cross-marking (Farthing et al., 1998). To handle large-scale examinations under pressure, educational institutions can benefit more because of this unbiased scoring system (Roediger & Marsh, 2005).

However, “Selected-response tests require much more time to create” (Powell & Gillespie, 1990, p. 1). Furthermore, covering a variety of topics within a short time is the utmost strength of MCQ (Bennett et al., 1990; Popham, 2000). As Powell and Gillespie (1990, p. 1) noted: “One major advantage of these tests is for measuring knowledge of specific facts”. Moreover, the writing speeds of different students do not

impact their reading performance (Farthing et al., 1998). When exams are performed online or using computers, students benefit from sitting for the exams from remote locations, and they can get timely feedback, which facilitates their understanding and learning processes (Weiss et al., 2006). Comparing the benefits of CR and MCQ, the present study focuses on the utility of the MCQ type.

Previous studies on SR and CR

To investigate the impact of item format on reading performance, several studies have been carried out using various methodologies to compare the CR and SR types of item formats.

MCQ easier

Masoumi and Sadeghi (2020) examined the influence of test format on test performance and checked the function of gender by comparing MCQ and CR in vocabulary tests in an EFL setting. Using descriptive analysis, the study revealed that MCQ tests were easier than CR, and males scored better than females in all versions of MCQ tests while females outperformed males in CR. Similar research was conducted by Nixon and Kennedy (2002), in which they compared students' scores in stem-equivalent MCQ and CR economics exams, and found that the students performed better in MCQ, and the effect of gender was insignificant. According to Famularo (2007), there were major differences between MCQ and CR products, with MCQ tests being much easier than their CR equivalents.

No significant difference between SR and CR formats

Hickson et al. (2012) investigated the degree to which grades based exclusively on CR differ from that of MCQ in economics classes. Although much effort went into constructing the CR, there was no significant difference in scoring, as they remarked that "the instructors of these classes made conscientious efforts to write CR questions

that assessed higher levels of learning (Bloom, 1956). Despite this, we find relatively little difference in grade outcomes (Hickson et al., 2012, p. 200)”.

In the 1995 and 1999 Trends in International Mathematics and Science Study (TIMSS), Hastedt and Sibberns (2005) compared MCQ and CR test formats as well but found only slight variations in MCQ and CR ratings. They suggested that “using both multiple-choice items and constructed response items seems to be appropriate in international studies because this assures that all students are treated equally and fairly” (Hastedt & Sibberns, p. 159).

Using Rasch analysis, Shaibah and van der Vleuten (2013) confirmed that MCQs are valid while comparing the differences in scoring of MCQ and CR items. They suggested that the results of this study provided empirical evidence that the SRF (MCQ) response format is a valid method and can be used as an alternative to the traditional FRF steeplechase examination (Shaibah & van der Vleuten, p. 149). Moreover, their results showed that students scored better in recall MCQ tests.

In a standardised exam for both reading and mathematics, Hollingworth et al. (2007) looked at the relationship between MCQ and CR using confirmatory factor analysis. They discovered that there was mixed support in their analysis. Their results backed up other research that suggested higher-order thinking can be manipulated using both formats.

Furthermore, to compare both item formats, a quasi-experimental analysis was performed on university students in the United States by Hancock (1994). However, according to the results, there was no difference in the formats. Similarly, in research conducted by Samson (1983) to compare the performance of MCQ, open-ended, and summary tasks, there have not been significant differences identified among these types. Many respectable tests in the United States use the MCQ format (American Marketing Association, 2001) as it is popular in testing economics (Nixon & Kennedy, 2002). As the previous studies highlight, there is no significant difference in the performance, and both CR and MCQ can test high order thinking skills. The present study chooses solely MCQ items, considering the cost factor.

2.3.2.3 Reading Assessment Scales

There are various approaches to looking at how reading is developed and tested (Alderson, 2000). Reading scales with full descriptions of each point, level, or band on the scale is one such method. ACTFL, TOEFL, DIALANG, DELTA, and CEFR are some of the outstanding scales in the assessment of reading. Using language tests with different levels is another method.

2.3.2.3.1 ACTFL

The American Council on the Teaching of Foreign Languages (ACTFL) aims at improving and expanding the learning and teaching of all languages from elementary to university levels.

The ACTFL Proficiency Guidelines are a set of guidelines for determining a foreign language speaker's proficiency. The ACTFL Oral Proficiency Interview is the most widely used oral proficiency test in North America (Lazaraton, 2002), and it is widely used in schools and universities in the United States (Ulrich, 2021). The guidelines are divided into different degrees of proficiency:

1. novice – divided into three levels: low (NL), mid (NM), and high (NH);
2. intermediate – divided into three levels: low (IL), mid (IM), and high (IH);
3. advanced – divided into three levels: low (AL), mid (AM), and high (AH);
4. superior (S)
5. distinguished (D)

In terms of domains, functions, contexts/content, text type, language control, vocabulary, communication strategies, and cultural awareness, the ACTFL Performance Descriptors are defined in three different subsets of communications skills with their own more generalised grading scales in all the following modes of communication:

1. Interpersonal (Novice, Intermediate, and Advanced)
2. Interpretative (Novice, Intermediate, and Advanced)
3. Presentational (Novice, Intermediate, and Advanced)

However, this scale has been criticised for its levels for not being empirically validated. Further, the research carried out by Lee and Musumeci (1988) and Allen et al. (1988) cited in Alderson (2000, p. 280), indicated the same, mentioning “failing to discover any significant difference between texts across different levels of learners”.

2.3.2.3.2 TOEFL

Test of English as a Foreign Language (TOEFL) is a standardised test used to assess the English language competence of non-native English speakers to enrol them in universities in English-speaking countries. The reading section of the test is designed to represent the types of reading that occur in university level academic contexts in keeping with this goal (Enright et al., 2000; Qian, 2002). From the standpoint of reader purpose (Enright et al., 2000), four types of reading activities have been identified:

1. reading to locate information, or search reading
2. reading for basic comprehension
3. reading to learn
4. reading to integrate information across numerous texts

The most difficult of the four categories is reading to integrate information (Goldman, 1997; Perfetti, 1997 as cited in Qian, 2002). This type asks the reader to combine knowledge from a variety of sources, including prose, diagrams, charts, and other kinds of presentations. However, due to several reasons, the TOEFL test has been criticised. The first complaint is that it has a significant, if not overwhelming, cultural bias because it appears that this test was made for Americans by Americans

(Traynor, 1985). Another major criticism of the TOEFL is that it is very much a blunt instrument, as there are many MCQs which allow guessing (Traynor).

2.3.2.3.3 DIALANG

DIALANG is a web-based system that supports 14 European languages and is based on the CEFR. It is not tied to any particular curriculum. The goal of the test is to provide diagnostic information on their vocabulary, structure, reading, and listening skills. It is a semi-adaptive exam, which means that instead of item-level adaptivity, it consists of three tests that cater to three different levels of proficiency, with each test-taker being assigned to the test that best matches his or her assessed proficiency. The competence level of a test-taker can be predicted using a Vocabulary Size Placement Test, allowing the appropriate test level to be assigned (Alderson, 2005).

Reading in DIALANG is viewed in terms of purpose and orientation. While reading for information is the primary focus, it also includes aesthetic, critical, and reflective reading. Descriptive, narrative, expository, argumentative, and educational texts are examples of text types (Alderson, 2000). Only three skills are tested in the DIALANG prototype:

1. identifying the main idea
2. identifying specific details
3. making inferences

In five ways, DIALANG proved to be a diagnostic test. First, it includes four tests: Listening, Reading, Vocabulary, and Structure. After taking all four exams, an individual will be able to determine his measurable degree of proficiency in each of the competencies (Kektsidou & Tsagari, 2019). He will also get a full description of what learners at his level of competency can perform for each of the abilities. Second, a test-taker has the option to activate the feedback button while taking the test, which will notify him of the correctness or incorrectness of his responses after each response is submitted. Third, a test taker's performance on various subskill components is

displayed. Fourth, a test-taker receives quick feedback and guidance on how to enhance his ability after each test. Finally, at the start of each test, the system performs a self-assessment. The goal of this self-assessment system is to improve the washback effect of DIALANG by encouraging the test-taker's awareness and self-directed language learning. The system notifies a test-taker at the end of the test whether he accurately self-assessed himself, and if there is a disparity between self-evaluation and measured assessment, possible explanations are provided.

2.3.2.3.4 DELTA

The Diagnostic English Language Tracking Assessment (DELTA) tests a range of reading subskills across various text types addressing different topics (Harding et al., 2015). This test is designed online for diagnosing Hong Kong tertiary level students' all four language skills, including reading based on Munby's (1978) taxonomy and Bachman and Palmer's (1996) framework (Urmston et al., 2013). This test consists of eight reading sub-skills which include: (a) identifying specific information, (b) interpreting a word or phrase as used by the writer, (c) understanding main and supporting ideas, (d) understanding information and making inferences, (e) understanding an argument made by the writer, (f) interpreting an attitude or intention of the writer, (g) understanding grammatical relationships of words or phrases across a text, and (h) identifying text type.

Another test similarly named the *Diagnostic English Language Test of Australia (DELTA)* was designed by McQueen and Alduous in 1994. This test, too, evaluates listening, reading, speaking, and writing skills. This test battery was published by the Australian Council for Educational Research (ACER). The goal is to provide a complete diagnostic account of non-English speaking background (NESB) pupils' English skills as they attend Australian secondary schools in Years 10 and 11 (Giri, 2005; O'Neill, 2009).

The DELTA reading test includes texts varying in complexity and length and texts relating to the demands of everyday school life and school subject areas. It is similarly contextualised to the learner group, to make it as communicative as feasible. The diagnostic map evaluates whether learners can read to complete both easier tasks like finding particular information on a timetable, and more complex ones like inferring the meaning of a word from context or understanding comparison (O'Neill, 2009). However, this test is designed for school-level students only.

2.3.2.3.5 CEFR Scale of Measurement for reading

Among the scales of reading measurement, CEFR scales are identified to be the best scale to assess the reading comprehension level expected of undergraduates for academic success. Nevertheless, CEFR has been empirically validated by many studies, for example, a wide-ranging Swiss research project scaled the levels through empirical Rasch analysis (North & Schneider, 1998), and further Alderson (2002), North (2014a), Waluyo (2019), Wu and Wu (2007), Dunlea (2015), as well as Huy and Humid (2015), validated the framework. As its scales of *can-do* descriptors are identified as an unparalleled success, as well as a preferred benchmark for language assessment and published courses worldwide, the present study uses the CEFR framework as a scale to measure reading performance.

Table 2.1 CEFR - Overall Reading Comprehension

CEFR Scale	ABILITY
C2	<p>Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings.</p> <p>Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.</p>
C1	<p>Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.</p>
B2	<p>Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.</p>
B1	<p>Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.</p>
A2	<p>Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency every day or job-related language</p> <p>Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.</p>
A1	<p>Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.</p>

(Adopted from Structured overview of all CEFR scales (Council of Europe, 2001b, p. 10))

As can be seen from Figure 2.1 illustrating Figure 4 of CEFR (Council of Europe, 2001, p.33), ten layers starting from 1 to 10 can be applied to ten levels of UTEL benchmarking. Compared to Figure 3 and Figure 5 of CEFR (Council of Europe, p.32), Figure 4 is more relevant to the current survey as it deals with a learning environment focusing much on *Independent Users*. According to a webpage

(www.ielts.org/about-ielts/ielts-in-cefr-scale), the chart which maps the IELTS band scores and the CEFR level descriptors, an IELTS band score of 5 is considered to be borderline for B1 and B2. However, the fact is that the IELTS band scores have not been rationalised with the CEFR level points very precisely, since the IELTS band score descriptors preceded (ielts.org, n.d.).

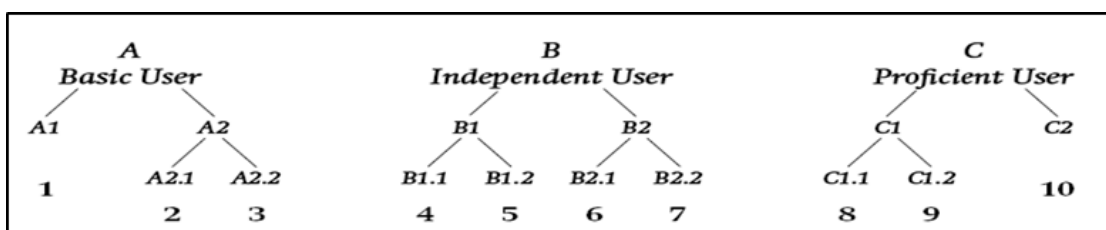


Figure 2.1 CEFR Level Illustrative Descriptors (Adopted from Figure 4 of CEFR for Languages: Learning, Teaching, Assessment (Council of Europe, 2001, p.33))

2.3.2.3.5.1 Profiling University Students' Reading Performance Aligned with CEFR

A few previous research studies have been identified to support the claim that a minimum requirement to achieve academic success in an EMI teaching or learning environment in the scenario of L2 is B2 or B1. Wu (2011) adapted B1 as the baseline for Taiwanese university students to establish the validity of the reading test of GEPT (General English Proficiency Test) in alignment with CEFR. In this context, English has been a second language (L2) for Taiwanese students, while outside the classroom they access their first language (L1).

A survey was carried out in Taiwan in 2018 at Walailak University to investigate the English proficiency of Thai EFL learners on their CEFR levels. Contrary to Wu's (2011) L2 situation, this study focused on 2248 Thai EFL (English as a Foreign Language) students. WU-TEP (Walailak University – Test of English Proficiency) was handled by the researcher, which is a locally designed comprehensive standardized test framed by the Classical Test Theory (CTT) and CEFR. Significantly, the results of the analyses exposed that a majority of the students were at A1 and A2 levels, being basic users of CEFR (Waluyo, 2019).

In Spain over 75 universities offer English medium degree programmes. For instance, at the University of Alcalá, a B1 level of CEFR is demonstrated at the exit level of graduation for many degree programs. However, a B2 level is required for those who take “English specialization” under Education (Laborda et al., 2017, p. 5).

On the other hand, the evidence for a minimum CEFR level required for foreign students in an English-speaking country was revealed through a study carried out by Carlsen (2018). She researched the relationship between academic success and Norwegian foreign students’ language proficiency, as measured by a CEFR-based university entrance test. The findings of the research revealed that the B2 level is the minimum requirement for the students to manage the linguistic demands they face in higher education, regardless of discipline and faculty.

On the whole, many higher educational institutions in Europe, as well as English-speaking countries, demand the C1 level as it provides the English proficiency the students need to succeed in their “undergraduate and postgraduate programmes” (Cambridge Assessment English, 2019, p.8). Some institutes there offer chances to those who have “B2 First qualifications” (Cambridge Assessment English, p.6). However, the prerequisite level is a little lower at B1 or B2 First in countries like Taiwan, Thailand and Spain where English prevails as an L2 or an FL (Foreign Language).

In this regard, in the context of Sri Lanka, the prominent ELT professionals agreed to have a UTEL band score of 5, which means an upper level of B1 as the exit level English qualification for the undergraduates who pass out from the university level. Therefore, the present survey considers the B2 First level (5 and above) as the baseline for the reading ability of the students who continue their studies in the EMI system for academic success.

2.3.2.4 Test Purposes

In the educational process, assessing learners is an indispensable function. Students’ assessments are unique in terms of what they test, who uses the results, and how the findings are used. A test’s purpose is a huge predictor of how the test will be created

and validated (Bachman & Palmer, 1996; Chapelle et al., 2003). Therefore, assessments could be classified into four categories based on their purposes: diagnostic, progress, achievement, and proficiency assessment (Hedgcock & Ferris, 2009).

Diagnostic Test

Diagnostic tests are tests that identify examinees' strengths and weaknesses (Harding et al., 2015); they are useful for instruction and learning (Alderson, 2005); however, due to the large number of items required to represent the detailed breakdown of the skills to be tested, and the scores must be analysed and reported in a way that provides information on those strengths and weaknesses, they are relatively difficult and time-consuming to construct.

Placement Test

Before the start of lessons, placement tests are performed to determine a student's language skill level so that an adviser and the student may sit down and decide the best course for the student based on the test results. A class that is too difficult for the student will not benefit their education, and a class that is too easy for them would be annoying. The exam results do aid in the selection of a course that will push the student.

In Malaysia, in most cases, although universities began their placement process by administering a tailored placement test, they abandoned this practice due to administrative constraints and the efficacy of the placement system (Zubairi, 2001). Because placement tests are required of all incoming students, they must be administered at the start of each academic year to avoid disrupting registration and participation in other academic courses. The time between taking the placement test and receiving the results is usually relatively short. Simultaneously, with such a large intake each year, there is always the issue of insufficient personnel resources to

accomplish the grading, double-marking, and calculation of results. These issues are common in Sri Lanka, too, according to the observation of the researcher.

Achievement Test

Any measurement technique or instrument whose objective is to estimate an examinee's level of attainment of specified knowledge or abilities is referred to as achievement testing (Elliott et al., 2011). Aside from this fundamental objective, achievement exams differ in terms of the intended score inference and application. The most common conclusions are either absolute levels of performance on the specified material or relative standing in comparison to other examinees. These tests are usually administered at the end of the course to measure the skills or knowledge that the learners gained.

Proficiency test

Proficiency tests determine linguistic skills based on what is required for a certain purpose, such as English for engineers, English for secretaries, English for car mechanics, and so on.

Performance Test

In performance tests, the examinee must show that what they have learnt has been applied in a real or simulated situation. The test stimuli and desired responses, or both, are designed to make the test situation realistic (Wesche, 1987).

According to Jones (1985), as cited in Wesche (1987), the difference between performance tests and other tests is in the “degree to which testing procedures approximate the reality in which the actual task would be performed” (Jones, 1985:16), so that “by observing examinees using the language within the context of a specific task, it is possible to predict how well they can perform under real conditions”

(p. 17). As a result, one of the most compelling reasons for using performance assessments is their ability to predict outcomes.

There are three major types of performance tests discussed in the literature: direct assessment, work sample, and simulation technique. In direct assessment, the examinee is placed in a realistic setting and his or her performance is assessed. The situation is not changed in this case to present certain duties. Work sample tests may or may not occur in the real setting and the situation is changed by the examiner to increase the efficiency of the test. In the simulation technique, the entire testing environment is performed, however, the results are still valid to reflect what are thought to be the most important characteristics of the situational context of actual use (Wesche, 1987). For example, role-playing is the most used simulation technique in language testing. Using the direct assessment method, the present study seeks to find out the performance level of university students' reading ability.

2.3.3 Validation

Validity is a broad measure of how well empirical data and theoretical rationales endorse the adequacy and appropriateness of interpretations and decisions based on test scores or other methods of assessment (Messick, 1989). Over time, the validation of a test has been established in a variety of ways by various scholars. A valid exam, according to Kelly (1927), tests what it is supposed to measure. Validity is a property of the meaning of the test scores, not of the test or assessment itself. Thus, it is the inferences drawn from test scores or other measures that should be validated, not the test or observation system per se (Messick, 1989).

The American Psychological Association (1985) recognized four forms of validity in 1954: material, predictive, concurrent, and construct validity. The degree to which the items used in a test are chosen from a universe of items and are indicative of the content expected to be tested is referred to as content validity. Whereas predictive validity is defined as a test's ability to predict a person's future success, and it is measured by comparing the results of one test with those of another. Concurrent validity is similar to predictive validity in that it is concerned with the degree of

correlation with another test, with the exception that the criterion test is administered at about the same time. Due to practical considerations, concurrent validity is necessary when substituting a test for an already existing standard one. Later, predictive and concurrent validity were merged into a single type as criterion-related validity (E. V. J. Smith, 2001). The degree to which a test reflects the underlying concept it intends to evaluate is referred to as construct validity.

As Cronbach (1980) mentioned, “all validation is one” (p. 99), and all these above-mentioned types became one type, and by “one” he meant construct validity. Messick (1989) corroborated it, stating the unified nature of validity and extending the definition of construct validity: “six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external and consequential aspects of construct validity” (p. 248). The American Psychological Association (1985) also accepted the unified view concept of validity.

Although Messick’s (1987) framework has been validated by many researchers (Baghaei & Amrahi, 2011; Bornstein, 1996; Guerrero, 2000), it has been widely criticised as well. McNamara (2006, p. 31) mentioned that “Messick’s writing on test consequences has informed debate on ethics, impact, accountability, and washback in language testing in the work of several researchers”. Messick encountered validity theory in the area of values. According to McNamara, his works missed focusing on the perspective of the social construction of language test constructs. Therefore, the present research identifies this gap and adapts Khalifa and Weir’s (2009) socio-cognitive validation framework for the validation of reading test as it highlights the social aspects on test validation.

2.3.3.1 Socio-Cognitive Model for Language Test Development and Validation

The main features of the socio-cognitive framework

The model includes six components such as test-taker characteristics, context validity, cognitive validity, scoring validity, consequential validity, and criterion validity. The first three components must be attended by the test developer before the test event

which is identified as *a priori* validation and the last three components should be concentrated as *a posteriori* validation after the test occurred.

Components relating to the *Test taker* are connected to both *Cognitive validity* and *Context validity* because “these individual characteristics will directly impact the way the individuals process the test task set up by the context validity box. Obviously, the tasks themselves will also be constructed with the overall test population and the target use situation clearly in mind as well as with concern for their theory-based (cognitive) validity” (Weir, 2005, p. 51).

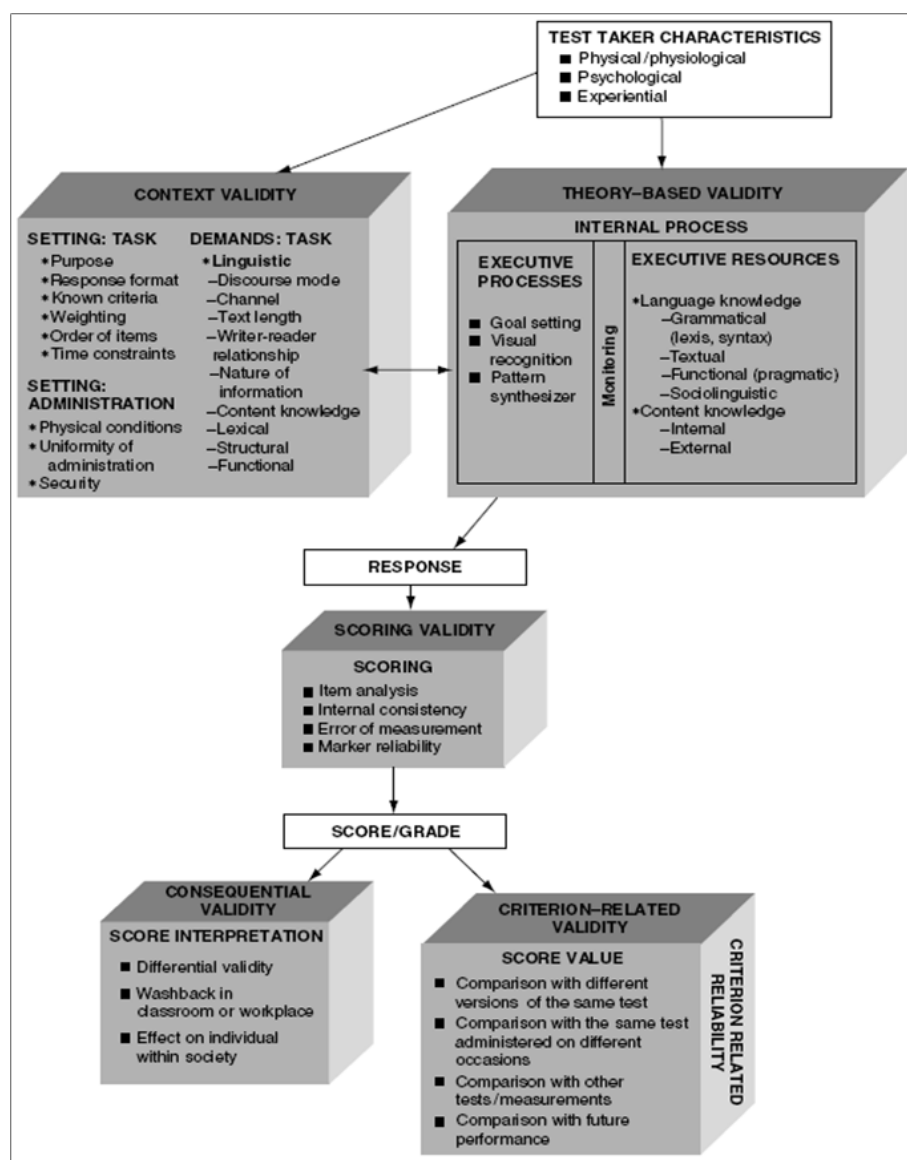


Figure 2.2 Socio-cognitive Framework for Test Development and Validation (Adopted from Weir’s (2005), p. 44)

2.3.3.1.1 Context Validation

The component of *Context validity* concerns the contextual parameters of a task. The test content should be suitable and comprehensive for the test takers. This component is often socially or externally determined in terms of task setting, test setting or administration, and task-specified input and expected output. There are fifteen components in Weir's (2005) context validity. Khalifa and Weir (2009) highlighted the significance of the "Order of items" (p. 82) from the task setting category and "content knowledge" (p. 82) from the linguistic demands category. For them, the order of items in which students are required to provide answers to reading comprehension questions should match the order in which the relevant reading material is presented. Because students apply the technique of chronological order when they read comprehension passages, when these students are exposed to exercises using a random order of responses, this order may delay comprehension and disintegrate the reliability of the test. Further, they highlighted that the students' background or content knowledge promotes their reading comprehension and enhances the authenticity of context validity.

2.3.3.1.2 Cognitive Validation

The cognitive processing approach is concerned with what readers do while they read different types of comprehension passages in real life (Khalifa & Weir, 2008, 2009). "This approach appears to offer the most tenable and productive theoretical basis for establishing the construct validity of test instruments" (Khalifa & Weir, 2009, p. 34). Comparing the factorial approach, the reading subskills approach, and the cognitive processing approach, Khalifa and Weir finally created a functional reading model as can be seen in Figure 2.2. It has three main components: meta cognitive activity, central core, and knowledge base (Brunfaut & McCray, 2015; Khalifa & Weir, 2009).

2.3.3.1.2.1 Metacognitive activity

Metacognition is broadly known as thinking about thinking. Goal setting and goal checking are primarily considered metacognitive activities under cognitive validation. The type of reading to employ when faced with different texts is determined by the metacognitive process of “goal setter” which links the central core (cognitive processes). In other words, the goal setter is critical in the decision taken on the purpose of the reading to select the most suitable strategies and determine what information they need to focus on in the text (Urquhart & Weir, 1998). “Local” and “global” are the two levels under goal setting. The term “local” comprehension means the understanding of propositions at the micro-level, which means at the sentence or clause level, whereas “global” refers to the macro-level referring to the entire structure of the text.

The goal checking is applied to each of the levels of cognitive processing (in the central core) according to the goal setter’s instructions. “Careful” or “expeditious” are the two “types of reading” according to Khalifa and Weir (2009) (Brunfaut & McCray, 2015).

As stipulated by Moore et al. The “componential matrix” formed by Weir and Urquhart’s two dimensions has the advantage of being a more dynamic model, one that is capable of generating a range of reading modes (Moore et al., 2012). The model is given in Table 2.2. However, as supported by Alderson (2000), Khalifa and Weir (2009) simply differentiate the two levels of “global” and “local” based on the purpose of understanding information within the sentence and beyond the sentence.

Table 2.2 Componential Matrix

	Global level (macro-structure)	Local level (micro-structure)
Careful Reading	<ul style="list-style-type: none"> • Establishing accurate comprehension of explicitly stated main ideas across sentences • Building a macro-structure on the basis of the information received • Making propositional inferences • Establishing how ideas and details related to each other within a whole text or comprehending overall text • Establishing how ideas and details relate to each other across texts or comprehending overall texts 	<ul style="list-style-type: none"> • Establishing accurate comprehension of explicitly stated main ideas or supporting details at clause and sentence level. • Identifying lexicons • Local inferencing is needed to build a mental model at sentence level • Understanding syntax
Expeditious Reading	<ul style="list-style-type: none"> • ‘skimming’ quickly to establish discourse topic and main ideas, or gist or macro structure of text, or relevance to needs • ‘search reading’ speedily to locate and understand information relevant to predetermined topics or needs 	<ul style="list-style-type: none"> • ‘scanning’ to locate specific points of information within the sentence or clause. • Looking for specific words/phrases, figures/percentages, names, dates of particular events or specific items • ‘search reading’ speedily within the sentence to gain certain key ideas

Adapted from Urquhart and Weir (1998, pp. 100-104) and Khalifa and Weir (2009, pp. 56-61)

2.3.3.1.2.2 Central processing core

Khalifa and Weir's (2009) model developed by Weir (2005), and Urquhart and Weir (1998), is designed on the assumption of a multi-componential approach to reading and the assessment of reading. There are eight cognitive processes elaborated by the framework as the central core of cognitive validation, which a reader needs in order to mature to gain a complete understanding of reading comprehension. The cognitive processes of reading are:

Word Recognition (WR), Lexical Access (LA), Syntactic Parsing (SP), Establishing Propositional Meaning (EPM), Inferencing (I), Building a Mental Model (BMM), Creating a Text Level Structure (CTLS), Creating an Inter-Textual Representation (CITR) (Refer to Figure 2.3).

Though the question of the hierarchy of the cognitive processing complexity in reading is difficult to answer (Badrasawi, 2012; Hudson, 2007; Jusoh, 2018), related to Bax, Khalifa and Weir's cognitive processes in reading supports a hierarchy order:

Khalifa and Weir's model describes cognitive processing in reading in terms of different levels of complexity, with, for example, lexical processing as the least complex, and intertextual reading as the most. Khalifa and Weir's model is therefore particularly valuable in that it operationalizes the concept of cognitive processing in reading (Bax, 2013, p.443).

The first four sub-skills presented by Khalifa and Weir (2009) are identified as low-level processes (low-order thinking skills), whereas the rest are known as high-level processes (high-order skills) (Bax, 2013; Bax & Chan, 2016; Brunfaut & McCray, 2015).

1. Word Recognition (WR): The reader identifies the same word in question or determines a word meaning independently and matches it in the text. This occurs at the word level.
2. Lexical Access (LA): The reader matches a synonym, antonym, hypernym, or another related word in the text using their knowledge of word meaning or word class (morphology). This occurs at the word level.

3. Syntactic Parsing (SP): The reader employs grammatical knowledge to determine comprehension and identify answers that are free of logical errors. This can occur at the clause or sentence level.
4. Establishing Propositional(core) meaning (EPM): The reader quickly establishes the meaning of a sentence at the local level by applying lexis and grammatical knowledge. It's a literal interpretation of the text. This occurs at the sentence or clause level.
5. Inferencing (I): To infer a deeper meaning, the reader looks beyond the literal or explicitly stated meaning. The reader can skim through the paragraphs looking for main concepts and notions that are just not explicitly stated. This can occur at the sentence level, paragraph level, or text level.
6. Building a Mental Model (BMM): By noticing important contrasts in a comparative and contrastive text type, the reader employs various elements of the text to form a wider mental model. This occurs at a whole text level.
7. Creating a Text Level Structure (CTLS): By analysing and differentiating primary ideas from supporting details, the reader applies genre knowledge to determine the text structure and purpose of the entire work. A skilled reader determines how the various sections of the text interact and which parts of the text are critical to the author's or audience's intention. This occurs at the text level.
8. Creating an Inter-Textual Representation (CITR): Understanding text and comparing it to other texts is important at this level. This occurs beyond the text level.

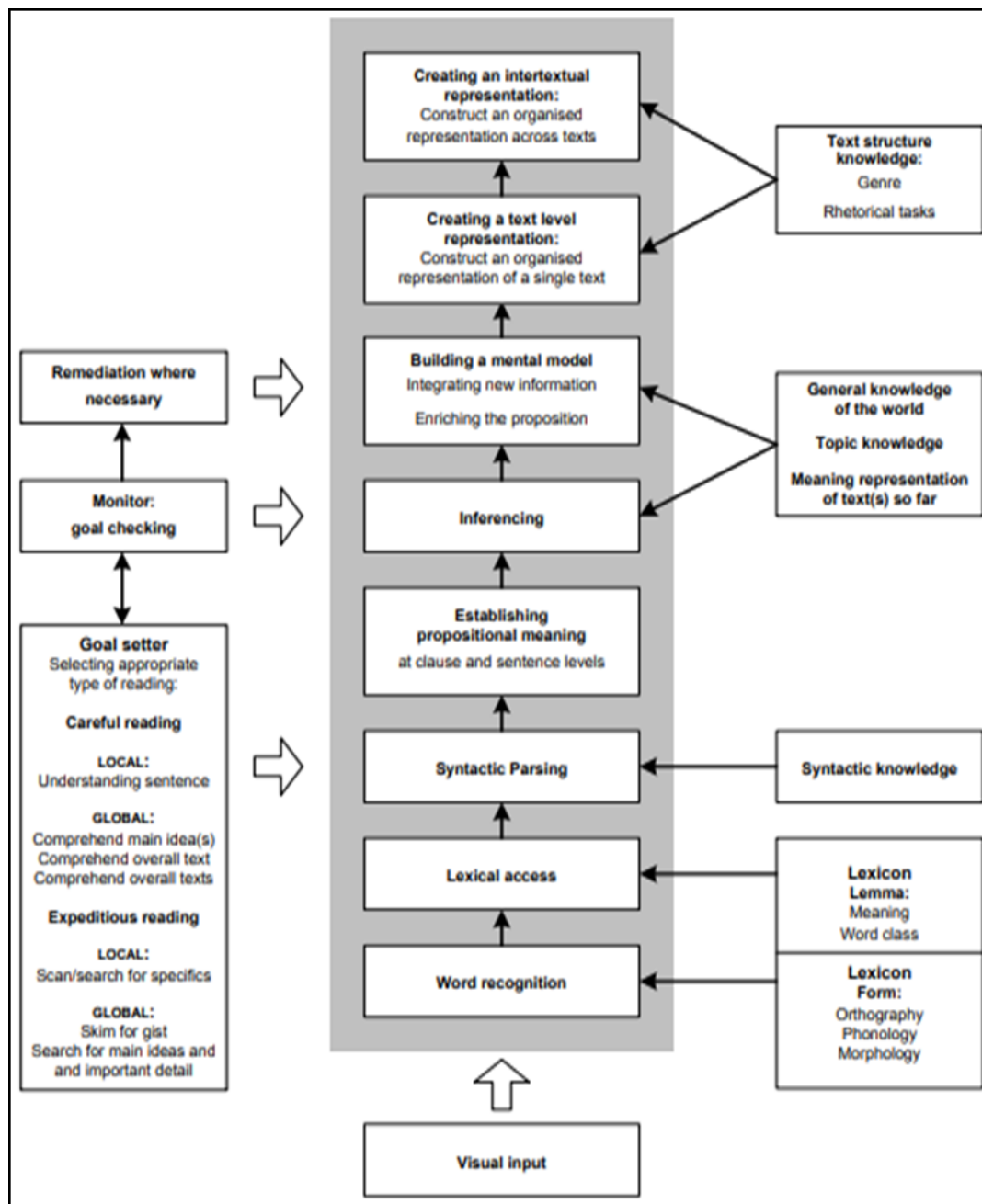


Figure 2.3 Khalifa and Weir's (2009) Model of Reading (Adopted from Khalifa & Weir (2009), p.43)

2.3.3.1.2.3 Knowledge base

The right-hand column of Figure 2.3 indicates the knowledge basis required to complete the task in the real-world context. Test-taker's lexical and syntactic knowledge is needed for the lower-level processes whereas text structure knowledge (genre, rhetorical tasks), general knowledge of the world, topic knowledge, and

meaning representation of the text is required for high-level processes (Brunfaut & McCray, 2015). This column is identified by monitoring by Owen (2016). The efficacy of the cognitive reading processes, which begin at the orthographic, phonological, and morphological levels in response to text-based input material, is directly influenced by the metacognitive and knowledge-base columns. Consequently, the model includes both bottom-up and top-down processes (Owen).

In Khalifa and Weir's (2009) survey, the Cambridge ESOL Main Suite Reading paper tasks do not focus much on the most complex cognitive processing in reading as it was shown in Table 2.3 CEFR A2 level to C2 level were studied in their survey.

Table 2.3 Cognitive Processing at A2 to C2 in Khalifa and Weir's (2009) examples of Cambridge ESOL Main Suite Reading papers

	KET A2	PET B1	FCE B2	CAE C1	CPE C2
Word recognition	√	√	√	√	√
Lexical access	√	√	√	√	√
Parsing	√	√	√	√	√
Establishing propositional meaning	√	√	√	√	√
Inferencing	(√)	√	√	√	√
Building a mental model	(√)	√	√	√	√
Creating a text level structure	■	■	■	√	√
Creating an organised representation of several texts	■	■	■	■	√

√ indicates a whole task or numerous items in a task(s) elicit this type of processing

(√) indicates only limited coverage of this type of processing

shading indicates this type of processing does not occur at all in a paper

(Adopted from Khalifa and Weir (2009, p. 70)

Therefore, as Mumin (2011) demanded to revisit chapter four, the Cognitive Validity of *Examining reading* can be viewed critically to enhance linguistic analyses of cognitive processes.

2.3.3.1.3 Scoring Validation

The ‘symbolic’ relationship between cognitive validity, context validity, and scoring validity creates what is known as “construct validity”, as Khalifa and Weir (2009) cited in Geranpayeh (2013, p. 242).

Scoring validity is concerned with all aspects of the testing process that can impact on the reliability of test scores. It accounts for the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in marking, are as free as possible from measurement error, are stable over time, and engender confidence as reliable decision-making indicators.

(Khalifa and Weir, 2009, p.143)

Scoring validity is vital because if we do not rely on students' scores, it does not matter if the tasks are potentially valid in terms of both cognitive and contextual parameters. The understanding that the test-taker has on the scoring criteria of the test is likely to influence their decisions about what to focus on and what not to focus on in their performance, and hence where to focus their attentional resources.

2.3.3.1.4 Criterion Validation

Criterion validity assesses how effectively a test can predict a specific outcome, or how closely the results match those of another test. Weir (2005, p. 35) defined it “a predominantly quantitative and *a posteriori* concept, concerned with the extent to which test scores correlate with a suitable external criterion of performance with established properties”. The comparison of the external criterion of performance and the test scores on the test that are to be validated may be considered either predictive or concurrent in nature. Predictive validity is comparing test results to an external assessment of the same candidate's performance after they have completed the test whereas concurrent validity is always observed by “comparing scores from a given test with some other measure of the same ability of the test takers taken at the same time as the test” (Shaw & Weir, 2007, p. 229).

Test scores, teacher's ratings, test takers' self-assessments, and real-life academic results are among the most widely utilised external measurements. According to Khalifa and Weir (2009, p. 190), criterion-related validity is examined using three parameters: a) “cross-test comparability”, b) “equivalence with different versions of the same test”, and c) “comparability with external standards”.

2.3.3.1.5 Consequential Validation

Messick (1989) highlighted that the appropriateness, usefulness, and meaningfulness of score-based inferences are determined by the external social consequences of the testing. The three aspects are evaluated when considering the consequential validity of the test as mentioned by Khalifa and Weir (2009, p. 169): (1) “Impact on institutions and society”, (2) “Washback on individuals in classroom/workplace”, and (3) “Avoidance of test bias”.

Washback is a term that relates to the impact of a test on teaching and learning (Alderson & Wall, 1993; Messick, 1996; Thaidan, 2015). Although consequential validity addresses the larger social influence of the exam, washback is an aspect of the consequential validity of the socio-cognitive framework. In language testing, the importance of positive washback is emphasised in the literature.

2.4 MEASUREMENT PROCEDURE

According to Mundrake (2000, p. 45), "Assessment, testing, and evaluation are terms used to describe the outcomes of the educational process". Measurement is another term that is often combined with assessment (Ghaicha, 2016). It really is the process of assigning a numerical value to the traits or dimensions of a student's performance while measuring ability or aptitude in such a way that maintains the student's performance quality (Bachman, 2004). It is particularly essential in the educational system because it is used to assess students' learning. Because there are so many stakes in the outcome of the measurement, it is critical that the method utilised is

valid, focusing on both the quality and quantity of a variable (Thorndike cited in Crocker and Algina, 1986).

2.4.1 Underlying Principles in Measurement Processes

There are three general steps in the process of measurement in psychology and education (Bachman, 1990; Crocker & Algina, 1986). The first one is to specify what is to measure. The next step is to locate or create a set of procedures that will isolate and display the attribute of interest. Unlike physical dimensions like length and height, which are easy to measure, psychological constructs are more difficult to quantify since they are defined in an overt, subjective manner. In psychological and educational testing, normally tests or batteries of tests are utilized as instruments. The third step in the measurement process is to develop a set of criteria or procedures for converting observations into quantitative statements of amount and degree.

Furthermore, Crocker and Algina (1986) added another step including the testing of the instrument before it is used. An instrument should be tested on a sample that is not going to be selected for real data. Ambiguous or conflicting conclusions would result if a pre-test were not conducted. However, this is a short account of the theoretical structure of the assessment theory or measurement procedures. Moreover, the limitations of the Classical Test Theory as well as the strengths of the Rasch MM are discussed here in recognition of their relevance to this study. Khalifa and Weir (2009, p. 144) mentioned that “Test items are usually analysed using classical test theory (CTT) or Item Response Theory (IRT) based statistics”.

2.4.1.1 Classical Test Theory (CTT) and its Limitations

The earliest theory of measuring is the classical test theory (CTT), which has been a dominant theory in the area of measurement (Suen, 2012). Its main goal is to assess the reliability of observed test scores; therefore, CTT is known as the classical reliability theory, or sometimes it is known as the true-score theory (Krabbe, 2017). Further, it converts a key component of qualitative or quantitative data into a

collection of real numbers (DeVellis, 2006). Regardless of its widespread use, due to many constraints, CTT has been criticised by many experts in the measurement field, like Keeves (1990), Hambleton et al. (1991), Hambleton and Jones (1993), Wright (1999), and Linacre (2003).

First, CTT is focused on the quality of the test but not on the individual items. Therefore, it is impossible to determine the performance of the examinee on a particular item; it also has another defect, that all the patterns of responses are accepted as valid even if they are extremely implausible.

The second constraint is that irrespective of the item difficulty level, CTT largely uses total scores to determine a person's ability. This means that a person's ability to answer the items is determined solely by the number of items properly answered and that both challenging and simple items are equally weighted (Hambleton et al., 1991). Therefore, using true (raw) scores are problematic. There are four issues with utilising raw scores in measurement: they count unequal size as equal, unequal size categories as equal, missing responses as failures, and incoherent responses as valid (Hambleton et al.).

Third, in CTT, test quality is mostly determined by reliability, which is calculated using a formula like Cronbach's alpha. The correlation between test scores on similar tests is sometimes referred to as reliability in CTT (Hambleton et al., 1991). The difficulty of achieving this parallel condition is the main criticism pointed towards CCT in this case. Furthermore, all examinees are supposed to have the same standard error of measurement. This assumption is incorrect since test results are not the equally exact system of measurement for examinees of varying abilities. As a result, the assumption of equal measurement errors for all examinees is implausible.

The other limitation is that CTT is generally descriptive in nature, and it is sample- or group-dependent. Hambleton et al. (1991) mentioned that the results of a test will differ if it is administered to various samples. This occurs because CTT describes data from a single test administration, its reliability, and the difficulty of particular questions, which is based on the percentage of test-takers who correctly answered them. Depending on the samples is an issue for test developers because they

must always attempt to design tests by putting them on a sample of students who represent the population for which the test is meant (Bond & Fox, 2015).

Finally, CTT is a deterministic theory in nature. As a result, Rasch (1980) criticised deterministic models for being limited in their ability to explain measurement defects, which are common in most natural occurrences

2.4.1.2 Item Response Theory (IRT) and Rasch Measurement Model (RMM)

Item Response Theory (IRT) is a testing theory based on the association between test takers' levels of performance on an overall measure of the ability that the item was supposed to measure and their performance on the test item. IRT has rapidly become popular in the measurement field due to its nature of solving practical measurement issues and providing theoretically justifiable assessment policies. It is used in many standardized tests, for instance, the Scholastic Aptitude Test, the Armed Services Vocational Aptitude Battery, Graduate Record Examination, etc. There are several IRT models available in different ranges of psychological areas. The Rasch measurement model (MM), the Andrich model, and the Masters models are to name a few (Embretson & Reise, 2013). Nevertheless, what Boone et al. (2014) mentioned about Rasch is noted here as they stated that the Rasch MM differs from the IRT model in that the IRT model is modified to match the data, whereas the Rasch MM is evaluated based on how well the data fits the model.

The Rasch MM differs from the IRT model in that the IRT model is modified to match the data, whereas when Rasch is employed, the data is evaluated to see how well it fits the model (Boone et al., 2014). "Rasch analysis considers" only one parameter, that is, item difficulty, whereas other models consider one or more parameters, like item discrimination or guessing factor (Khalifa & Weir, 2009, p. 146).

2.4.1.2.1 Characteristics of Rasch Measurement Model

George Rasch, a Danish mathematician, created the Rasch MM which was expanded by Benjamin Wright of the University of Chicago, who elaborated on this in his books *Best Test Design* (Wright & Stone, 1979) and *Rating Scale Analysis* (Wright & Masters, 1982). Rasch and Wright identified the inefficiency of the raw test data as specified by many scholars in Section 2.4.1.1. They noticed these raw data are ordinal and highlighted that these ordinal data can violate the assumptions of test statistics by providing three possible problems.

1. The inability to express how an individual student performed on a test (or their attitude expressed on a survey) with respect to its items.
2. The inability to track students' development over time in detail with a single scale.
3. Lack of quality control in terms of data quality and instrument functioning.

The Rasch MM prescribed by the Rasch family has become popular around the world to facilitate precise measurement instrument development. It is the only system of accurate measurement of latent traits that can be applied to both the physical and psychological sciences (Granger & Linacre, 2008). It is a mathematical function that relates the chance of getting the right answer on an item to the difference between a person's ability and an item's difficulty (Bond & Fox, 2015; Boone, 2016; Rasch, 1980).

The theoretical notion of the Rasch model is based on two important premises:

1. More capable individuals are more likely to successfully answer all of the items.
2. Easy items are more likely to be correctly replied to or attained by all individuals.

To apply both IRT and Rasch MM, it is necessary for data to fulfil the following assumptions: unidimensionality of the construct; item independence; similar

item discrimination indices; and sufficient allocation of time so that the test-takers can answer all items (Hambleton et al., 1991). Unidimensionality means that the test items focus on a single ability or construct (Bond & Fox, 2015; Hambleton et al., 1991). Item independence refers to the fact that the response to one item does not influence the response to another item in the same test (Hambleton et al., 1991; Khalifa & Weir, 2009). Another essential characteristic of RMM is its objectivity which means that the separation of item difficulty and person ability provides objectivity in RMM because person measurements are independent of the items being measured, and vice versa (Granger, 2008).

Rasch is more than just using the Rasch model in a software program; it also entails considering what it means to measure. When Rasch MM is used, multiple pieces of evidence are frequently considered and assessed to establish a conclusion, such as the overall performance of an instrument (Neumann et al., 2014). Rasch MM offers a wide range of strategies for assessing instrument performance. Evaluating the performance of an instrument using Rasch MM is very similar to the processes of developing laboratory equipment by biologists, physicists, and other scientists (Neumann et al.).

2.4.1.3 Conceptual Framework

A comprehensive literature review of existing studies and theories on the present topic, which is discussed in this chapter, is used to develop a conceptual framework. A conceptual framework is a collection of linked concepts organised into a network or plane that collectively provide a thorough knowledge of a situation. A textual or visual depiction of how variables and concepts are supposed to interact is given in Figure 2.4. In this conceptual framework, each idea has an ontological or epistemic purpose.

Figure 2.4 explains the visual description of the conceptual framework. Based on the limitations of the previous research, the present study highlights the concepts needed to interact in this study. The following sections 2.4.1.3.1 and 2.4.1.3.2 explain the variables and concepts included in the present study.

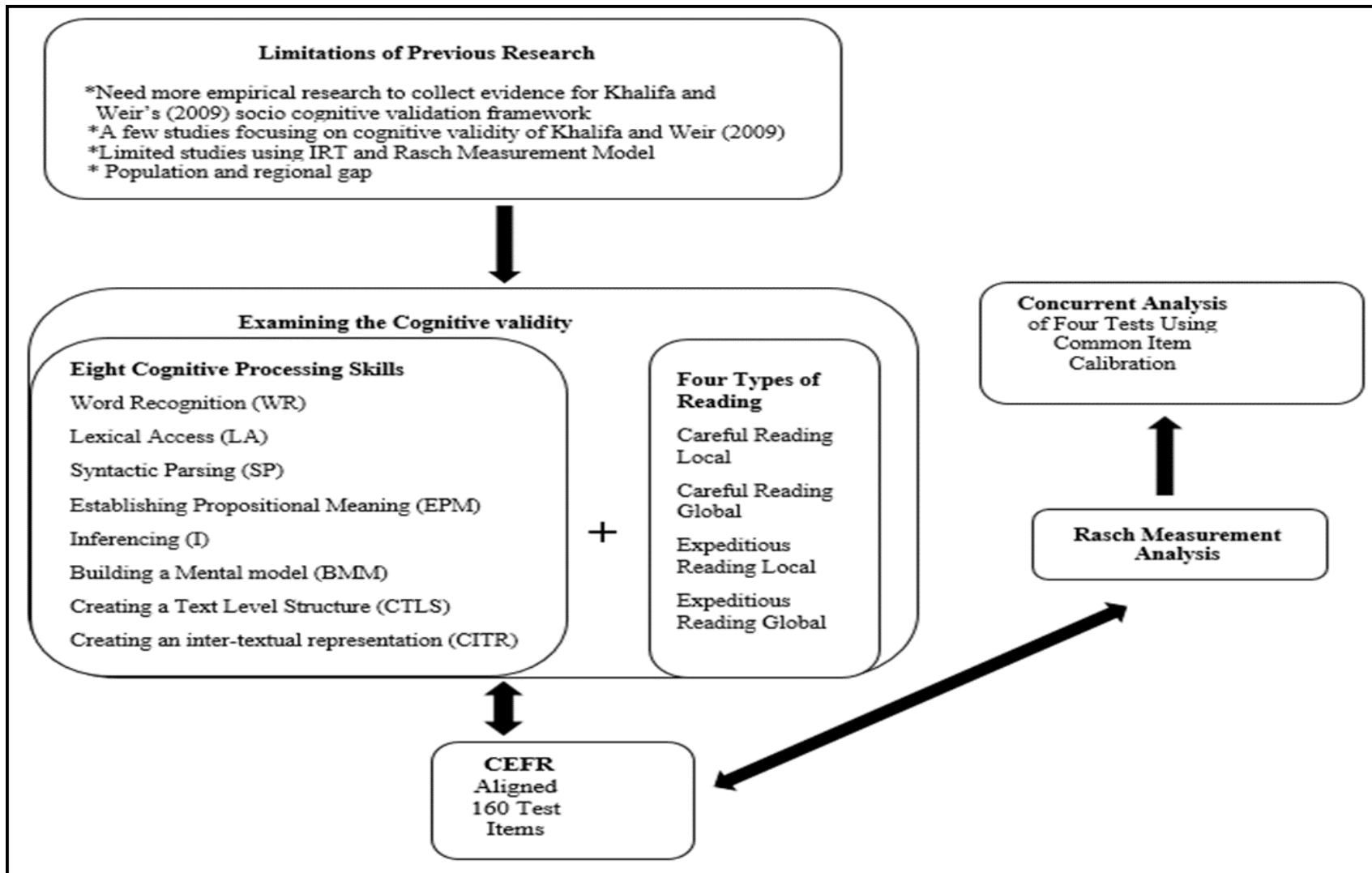


Figure 2.4 Conceptual Framework

2.4.1.3.1 Limitations of the Previous Method and the proposal of the current method

Conventional approaches to determining item difficulty by depending on expert assessment have resulted in a variety of issues. One issue is that there is no guarantee that the items tested will be of the difficulty that the experts have specified. This can be demonstrated by examining the outcomes of numerous studies (Bachman & Palmer, 1996) that suggest complications in determining the difficulty of items. Despite its shortcomings in this system, it is nonetheless extensively used when attempting to classify the skills that an item tests as well as the level of difficulty (Bramley & Wilson, 2016; Wu, 2011). Because of this over-reliance on expert judgments, the findings of studies using this method have been inconsistent. Although the practice is beneficial, it should be backed up by empirical evidence.

The use of factor analysis is another prominent method of evaluating the ordering of reading skills. Despite its widespread use, factor analysis has its own set of restrictions. If the method and perspective of factor analysis are not coherent with the rules controlling the development of the data, statistical summaries achieved from the factor analysis from replication to replication may easily give misleading conclusions (Andrich & Godfrey, 1978).

Another flaw is that most techniques are incapable of handling a high number of items for each cognitive process of interest. Because the main goal of this study is to find out the performance of the students of different faculties, therefore, there must be different types of tests including several items spanning all cognitive processes, to produce accurate and conclusive results. Longer testing, on the other hand, usually have an impact on the validity of the results. Nonetheless, the Rasch analysis offers a unique quality that allows many items to be assessed without lengthening the test using a linking mechanism. Common item linking of Rasch MM can help sort out this issue.

The study proposes the use of IRT and the use of Rasch Measurement Model. Kobayashi (2009) strongly suggests that IRT is capable enough to be used to investigate reading performance. Further, the Rasch MM is believed to be flexible and promising, due to the limits provided by previously stated methodologies in the investigation of the reading performance of university students. Baghaei and Amrahi (2011) and Aryadoust and Zhang (2016) advocated for the use of the Rasch Measurement Model in measuring reading comprehension because of its robustness.

2.4.1.3.2 Application of RMM in validation studies and Socio-cognitive validation framework for Reading

Rasch analysis is primarily used to satisfy the construct validity requirements (Messick, 1996). Item fit statistics and unidimensionality map are other features of Rasch in the validation of the construct. Further, external form construct validity is measured by person separation or stratum (Wright & Stone, 2004)

Rasch has not defined a method for determining the consequential factor of validity. However, issues such as item discrimination, differential item functioning (DIF), or a close inspection of the person-item map, which shows the amount of knowledge from which action decisions are made, can provide useful evidence to decide on the consequential aspect of constructs (Baghaei & Amrahi, 2011). Therefore, DIF can be utilised to identify the consequential validity.

This model allows the basis for the construction of the measurement scales that is necessary for item banking and linking procedures. In item banking, items of the same difficulty are calibrated as well as they can be linked together using common persons or items (Bond & Fox, 2015; Khalifa & Weir, 2009; Wright & Stone, 1979). In operationalizing the reading construct, Khalifa and Weir (2009) applied the Rasch MM in *Examining Reading*. Similarly, “Cambridge ESOL uses the one parameter Rasch model” to calibrate the items (Khalifa & Weir, 2009, p. 147).

In the area of TESOL, the use of the Rasch measuring model to assess students or validate examinations and surveys has grown more widespread (Baghaei & Amrahi, 2011; Karlin & Karlin, 2018; Wu, 2011). Although Baghaei and Amrahi (2011), claimed that all tests and surveys are multidimensional to some extent, in language assessment it is inevitable for the test designers, teachers, and researchers to evaluate a single construct. However, the Rasch MM can determine how much multidimensionality is present in a test, and it is up to the test-maker to decide if this level of multidimensionality is acceptable (Baghaei & Amrahi, 2011; Karlin & Karlin, 2018). Based on the aforementioned advantages of the Rasch MM, the present study applies it in its appraisal.

The goal of this study was to apply Khalifa and Weir's (2009) socio-cognitive validation paradigm to reading examinations, which is currently limited to the cognitive validity of the test. In terms of test development and validation, the framework is thought to have “direct relevance and value to an operational language testing/assessment context” and “to be both theoretically sound and practically useful” (Taylor, 2011, p. 2). Despite its widespread use in test validation research, the current use of the framework is confined to the central processing core of cognitive validity and for independent reading tests like the present tests.

2.5 SUMMARY OF THE CHAPTER

The ELT and language testing in Sri Lanka, concepts involved in reading, assessing reading, as well as studies connected to the investigation of Khalifa and Weir's (2009) reading model, have been thoroughly explored in this chapter. This chapter has also emphasised the advantages of the Rasch MM in resolving measurement issues. The next chapter explains the methods implied in the study.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 INTRODUCTION

This chapter discusses the stages involved in the research design and research procedure in detail. The selection of the reading model, construct definition, instrumentation, content validation of tests by experts, analysis of content validation using multidimensional objective measures of IOC, sampling procedure, piloting the test, and the analysis of the results are presented accordingly.

3.2 RESEARCH DESIGN

Good research is driven by the paradigm in which it is conducted. Although educational research procedures are multimethod and multidisciplinary (Gay & Airasian, 2000), based on the objectives of the current research, it converges more on quantitative methods applying statistical analysis. Out of seven types of research mentioned by Kor and Teoh (2009), typically descriptive research expounds on the characteristics of the population, situation, or phenomenon of the research being studied. Since the present study intends to investigate the nature of reading skills and to describe the L2 learners' skills empirically, the most suitable research design for this study is descriptive (Bickman & Rog, 1998). Particular research methodologies and data collection procedures such as tests, surveys, observations, and self-reports are commonly exploited in this research design (Fraenkel & Wallen, 2006; Gay & Airasian, 2000; Kor & Teoh, 2009). In this research, to identify the cognitive processing in reading the content validation involving a panel of experts was carried out both qualitatively and quantitatively. The L2 learners' reading performance was measured through the pilot study which collected data in the form of responses to CEFR-aligned reading tests, adopted, and adapted from LRN.

Tests that entail written answers are typically used to measure achievements in the subject matter in the educational settings (Keeves, 1990), for one of the purposes of the test is “to measure the language proficiency” of the students (Hughes, 1989, p. 7); in this case, the tests explore the reading ability. Thus, the main tool employed to collect information regarding the cognitive processes in reading was in the form of English reading tests in the present study.

3.3 RESEARCH PROCEDURE

This section explains the two stages of how this research study was carried out as discussed earlier. The first stage included the logical analysis of the expert judgment and the second involved the empirical evaluation of the data collection as shown in Figure 3.1.

Figure 3.1 explains the phases involved in the logical analysis and empirical evaluation. Analysis of reading theories, assessment of reading, defining reading skills, sub-skills, and construct, understanding of reading taxonomies, and reading models are some procedures involved in logical analysis. Selecting Khalifa and Weir’s (2009) socio-cognitive validation framework, selection of CEFR-aligned texts along with their items, and analysis of items based on Khalifa and Weir’s cognitive processes of reading are, too, among the logical analysis phase. Empirical evaluation of the study included content validation procedures, pilot study, analysis of the data received from the pilot test using the Rasch Measurement Model, and refinement of the items for the final data collection.

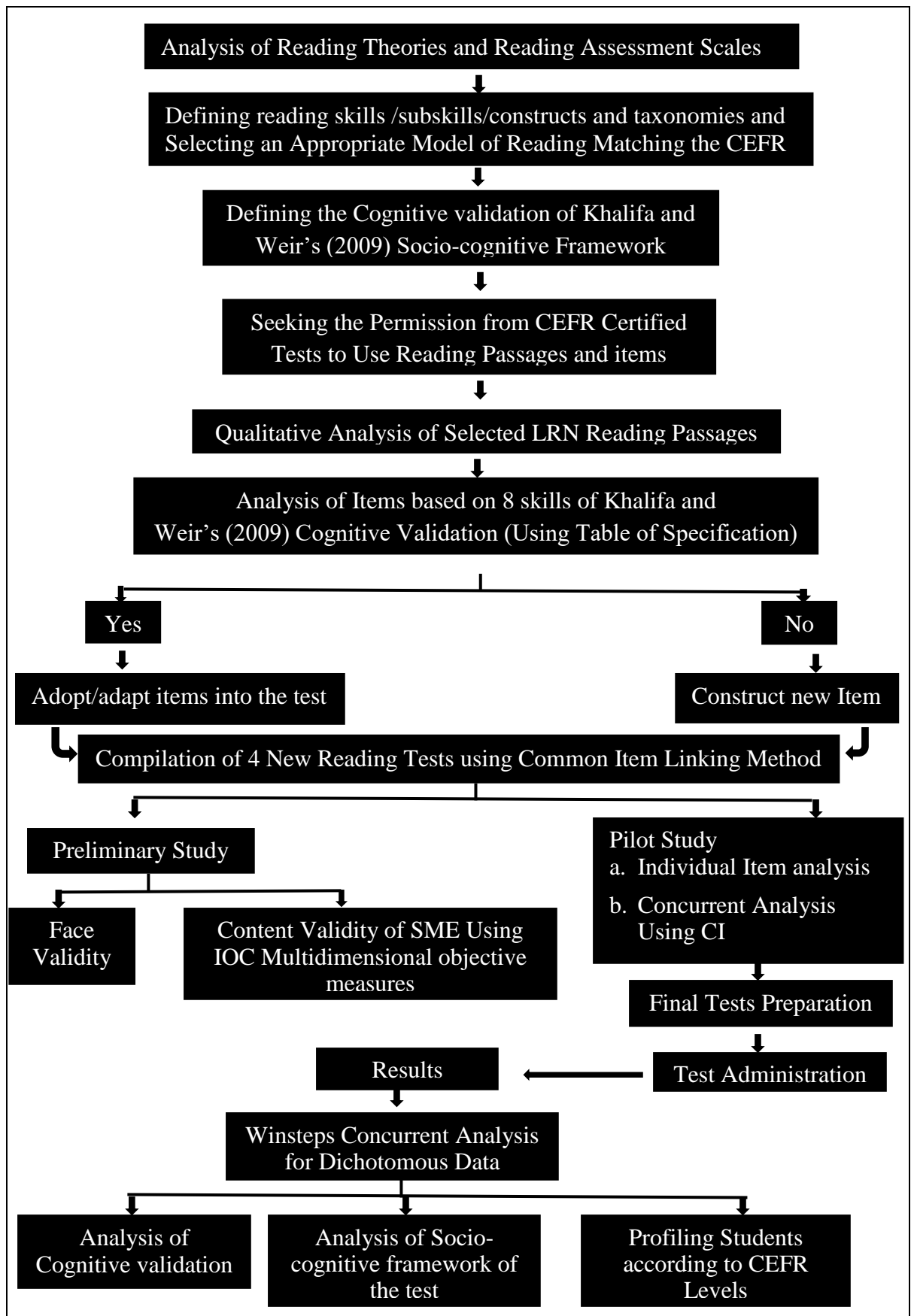


Figure 3.1 Research Procedure in Graphic View

3.4 POPULATION AND SAMPLING OF THE STUDY

The population of the current study belongs to the South Eastern University of Sri Lanka, which is one of the national universities in Sri Lanka established in 1995 (Kareema, 2013). It expands its services to six faculties, which include Arts & Culture, Management & Commerce, Islamic Studies & Arabic Language, Applied Sciences, Engineering, and Technology. Besides the Facultes of Arts & Culture, and Islamic Studies & Arabic Language, all the other four faculties offer English Medium Instruction (EMI); however, certain special degree programmes in these two above-mentioned faculties are also offered in the English Medium. All students in these programmes are offered credit-carrying English language courses presented by the Department of English Language Teaching (DELT).

Presently, around 4800 students are studying at the university (www.seu.ac.lk/vcoffice/) and half of them follow EMI. For undergraduates, reading is the most significant skill for better academic performance (Hermida, 2009), thus this study focuses on this population to examine the level of reading ability of these students according to the CEFR levels.

3.4.1 Sampling Procedure and the Characteristics of the Respondents

The sampling technique is central to any research work (Bryman, 2016). This study deals with the undergraduates of the South Eastern University of Sri Lanka who opt for the EMI programme for their degree in four different faculties, namely (i) Faculty of Arts and Culture (FAC), (ii) Faculty of Management and Commerce (FMC), (iii) Faculty of Engineering (FE), and (iv) Faculty of Applied Sciences (FAS) The first two faculties are categorized as human science faculties, which include the Humanities, Education, Management, and Social Sciences (HEMS), whereas the remaining two are listed as pure science faculties comprising the Sciences, Technology, Engineering, and Mathematics (STEM) (Ministry of Higher Education and Highways and World Bank, 2018).

In line with the objective of this study, the sampling technique used is the purposive sampling mode, in order to ensure the representativeness of the sample to the population. Even though it is statistically unrepresentative of the larger population, researchers believe this mode of sampling matches the profile of the people they need to meet (Lavrakas, 2008). Though this sampling procedure is very popular in qualitative research methods (Creswell, 2012), in this research this method is presumed to be more fitting to collect data in terms of students' ability and their academic achievement, as the DELT has the practice of maintaining ability level grouping, and certain faculties offer only EMI instruction. In the Sri Lankan university system, the English language is taught to all university students. However, some students follow the EMI system because certain courses are taught only in English, whereas some of them follow their degree programmes in their mother tongue.

Each academic year, the Department of English Language Teaching (DELTA) of SEUSL administers a placement test in the English language to all new student entrants. Based on their performance in that test, they are divided into ability-level groups within their respective faculties. This classification enables the researcher to select the appropriate samples for the study more efficiently since many of these groups share similar characteristics among their members. Therefore, among the many procedures of the purposive sampling method, this study focused on homogeneous sampling, whereby "the researcher purposefully samples individuals or sites based on membership in a subgroup that has defining characteristics" (Creswell, 2012, p. 208).

Contrary to the view that purposeful sampling has low-level reliability and inability to generalize the findings, the present study applies the Rasch MM, which is a very flexible approach that allows for certain requirements (Ingebo, 1997); as such, this sampling procedure is adequate to reflect the entire population, in order to ascertain generalizable decisions regarding the reading performance of the EMI students.

Another reason for selecting this method is to profile different subgroups of students with different levels of reading skills. It is believed that the English language ability of students of the Faculty of Applied Sciences and the Faculty of Engineering is superior to that of students of the Faculty of Management and Commerce and the

Faculty of Arts and Culture, respectively (Rathnayake, 2013). Similarly, urban students display higher ability than their rural counterparts. To expand the spread of scores in the tests in Rasch analysis, it is crucial to choose different language ability students to examine different CEFR levels, as expected in reading.

3.4.2 Sample Size

According to Fraenkel et al. (2012), the large sample size is vital to guarantee that relevant subgroups are represented in a good manner, and it is more likely to characterize the total population. On the other hand, in the Rasch measurement model, it is possible to obtain useful results with a small sample size. According to the sample-free item analysis procedure mentioned by Wright and Panchapakesan (1969), the significance of the sample's ability or size becomes insignificant. Further, this idea is highlighted by many scholars in the measurement field that Rasch is strong enough to identify missing data (Bond & Fox, 2015; Granger & Linacre, 2008; Linacre, 2020b; Wright & Stone, 1979). Nevertheless, Granger and Linacre (2008) believe that the most reliable analysis comes from at least 50 - 100 samples.

...it is necessary that the sample size is sufficient to assure that items and categories are representative of the construct/dimension being measured. The intent of Rasch modeling is to create a measure that is "sample free" meaning being independent of the sample from which it was derived. Thus, while a minimum sample size is not a requirement to perform Rasch analysis; the interpretation is most reliable with at least 50 - 100 subjects. While RA can handle missing data by estimating the value of a missing item,... (Granger & Linacre, 2008, p. 9).

Likewise, the present study opted to use a minimum of 100 samples for each test, as suggested by Linacre (2020b, p. 3). He mentioned that the results are "more believable" when the sample size is "closer" to 100. His guidelines for maintaining useful measurement calibration stability are stated in Table 3.1 as follows:

Table 3.1 Sample-Size Range for Calibration (Linacre, 2020b)

Item calibrations stable within	Confidence	Minimum sample size range (best to poor targeting)	Size for most purposes
± 1 logit	95%	16 -- 36	30
± 1 logit	99%	27 -- 61	50
± ½ logit	95%	64 -- 144	100
± ½ logit	99%	108 -- 243	150
Definitive or High Stakes	99&+ (Items)	250—20 * test Length	250
Adverse Circumstances	Robust	450 upwards	500

However, the minimum number of respondents for a study may vary depending on what the study is trying to achieve. For instance, Wright and Stone (1979) highlighted two aspects, such as test takers and the items determining the sample size. They proposed a minimum of 20 items and 200 examinees for a particular test; the current study used 40 items among a minimum of 100 examinees for an individual test.

In the present study, four different reading tests were conducted among the chosen four faculties: Test 1 was conducted among 208 FAC students; Test 2 was carried out among 247 FMC students; Test 3 was conducted among 268 FAS students; and Test 4 was administered among 179 FE students in the final data collection. A total of 902 students participated in this study, which is a large sample size representing relevant subgroups and different ability level students.

3.5 INSTRUMENT OF THE STUDY

This section presents the steps involved in developing the instrument for the study. First, a careful selection of reading texts along with items was carried out before adapting the 13 texts and their accompanying items to construct four different final tests. Then, the tests were validated through empirical evaluation processes involving preliminary investigation and piloting.

3.5.1 Test Development and Adaptation

Test development is compared to a kind of architectural activity (Fulcher, 2010). “The test design cycle” by Fulcher (2010, p. 94) is a good example that elaborates on the endless process of test designing. The purpose of a test is a very strong determinant of how to develop and validate the test (Bachman & Palmer, 1996; Chapelle et al., 2003). Three separate objectives were differentiated by Chapelle et al. (2003): use, infer, and effect. Focusing on these three purposes and teaching, learning, and assessing conditions prevailing among the target population, the present study relied on test adaptation due to save time and expenses.

Test adaptation is a process by which a test is conducted from a source language and/or culture into one or more languages and/ or cultures. There are several guidelines presented by Hambleton (1996) for test adaptation. According to these guidelines first, it is an important matter to select a standardized reading assessment scale, and the research ethics involved in utilizing authentic test materials, the researcher endeavoured to adopt and adapt test materials from various test-providing agencies.

After a considerable waiting period, the Learning Resource Network (LRN), a CEFR-aligned testing agency, finally granted their permission to utilize their reading test materials. A test was expected to have 40 items and in all 4 tests, a minimum of 160 items were needed. Creating more items can facilitate to maintain an effective instrument, because at the end of content validation, some items may need to be removed or modified. Thus, out of 162 items created for the study around, 7 items

were created by the researcher, whereas the other 155 items were originally taken from different CEFR levels LRN reading texts and their accompanying items. Almost all the texts selected in the tests were designed for the purpose of reading for orientation or information, which requires deep reading, whereas reading for pleasure (entertainment/extensive), which involves shallow comprehension (Zhang & Duke, 2008), is not targeted, since the target population of the tests belongs to ESL adult learners who are motivated towards academic achievement. The test items were developed based on Khalifa and Weir's (2009) socio-cognitive frameworks of reference for reading, founded on a table of specifications recommended by Alderson (2000) and Fulcher (2010). The table of specifications developed for this study (see Appendix B) was designed to focus on the parameters of cognitive processing, types of reading, and item format (test method).

3.5.1.1 Selection of the LRN Texts for Item Adaptation

Reading test materials were taken from the LRN, which is a globally-recognized awarding organization accredited by the Ofqual in England and by the British Council, the UK, since 2011. It is a member of leading international testing associations, including the ALTE, EALTA, IATEFL, and a corporate member of English UK (www.lrn-global.org, n.d.-b). It issues the international ISOL examinations for CEFR A1, A2, B1, B2, C1, and C2 levels, IELCA (multi-level test), CAB, etc. According to Hidri (2020, p. 742),

The International English Language Competency Assessment (*IELCA*) is an exam recognized by international public and private academic and professional institutions, such as *Bangor University, De Montfort University Leicester, University of East London, Maritime & Coastguard Agency, UK, Italian Ministry of Education, Malta Qualifications Recognition Information Centre, Conferencia de Rectores de las Universidades Españolas, College of Europe, Toyota, and DHL.*

Thirteen texts with their accompanying items were either chosen or adapted from the LRN, which has achieved accreditation and externally validated licenses like ISO 9001, ISO 14001, and ISO 27001, and its materials are already validated by the international assessment authorities' vetting processes. The present study adopted reading materials from this resource because of its accessibility, global recognition, assured quality, and flexibility.

Tables 3.2 to 3.5 below explain the summary of the tests in tabulated view for easy understanding. Under 'Content' text numbers were included. It means there are four texts in each testlet. Each text is selected according to the CEFR level, and the texts have varied difficulty levels according to the CEFR. The second column indicates the CEFR level of the text and the 'Source' where the texts have been taken from. Further, the title of the text is given in the third column, followed by the number of types of the test method (item format) included in each text. For example, text 4 in Test 1 has two types of item formats, such as multiple matching, and true, false, or not given type of items. The fifth column explains the response method.

The originality of the item is given by the terms 'original' or 'self-constructed' in the sixth column, the term 'original' meaning that the items were taken from the LRN resources; however, the term 'self-constructed' meant that the items were created by the researcher. Only 5 items were finally constructed by the researcher in all four tests, after the content were validated by experts. The item difficulty (or the CEFR level) of those self-constructed items follows the item difficulty level of the texts they belong to. The number of items in each text is followed by the total number of items in each test. There are around 9 to 11 items developed based on each text, and the total number of items in each test is 40. The expected amount of time to complete the task under test conditions is given in the final column.

Table 3.2 Overview of Test 1

Content	CEFR level & Source	Title	Type	Response method	Original/ Self-constructed	No. of Items	Total Items	Time (mins)
Text 1	B1(Fill in the gap) - LRN Sample paper	ICE- CREAM	1	MCQ 3 options	Original	10	10	10
Text 2	C1 - LRN Sample paper	Jet Lag	1	MCQ 3 options	Original	9	9	40
Text 3	C1- 2016/January past paper	Holidays	1	MCQ 3 options	Original	11	11	15
Text 4	IELCA – Multi level- 2016 past paper General English	Smartphone, technology	1	Multiple Matching	Original	4	10	20
			2	True/ False/NG	Original + Self constructed	5+1		

Table 3.3 Overview of Test 2

Content	CEFR level & Source	Title	Type	Response method	Original/ Self-constructed	No. of Items	Total Items	Time (mins)
Text 1	B1(Fill in the gap) - LRN Sample paper	Educational Programmes for Adults	1	MCQ 3 options	Original	10	10	10
Text 2	C1 - LRN Sample paper	The Tradition of Coffee Drinking	1	MCQ 3 options	Original	9	9	40
Text 3	C1- 2016/January past paper	Holidays	1	MCQ 3 options	Original	11	11	15
Text 4	IELCA – Multi level- 2016 past paper General English	London Olympics	1	Multiple Matching	Original	4	10	20
			2	True/ False/NG	Original + Self constructed	5+1		

Table 3.4 Overview of Test 3

Content	CEFR level & Source	Title	Type	Response method	Original/ Self-constructed	No. of Items	Total Items	Time (mins)
Text 1	B1(Fill in the gap) - LRN Sample paper	Playing Outdoors	1	MCQ 3 options	Original	10	10	10
Text 2	C1 - LRN Sample paper	Supersonic Flight	1	MCQ 3 options	Original	9	9	40
Text 3	C1- 2016/January past paper	Holidays	1	MCQ 3 options	Original	11	11	15
Text 4	IELCA – Multi Level- Sample paper – Academic English	Sir William Empson	1	Multiple Matching	Original	4	10	20
			2	True/ False/NG	Original + Self constructed	5+1		

Table 3.5 Overview of Test 4

Content	CEFR level & Source	Title	Type	Response method	Original/ Self-constructed	No. of Items	Total Items	Time (mins)
Text 1	B1(Fill in the gap) - LRN Sample paper	Having Friends	1	MCQ 3 options	Original	10	10	10
Text 2	C1 - LRN Sample paper	The Magic of the Cinema	1	MCQ 3 options	Original+ Self constructed	8+1	9	40
Text 3	C1- 2016/January past paper	Holidays	1	MCQ 3 options	Original	11	11	15
Text 4	IELCA – Multi level- 2016 past paper General English	Architect Renzo Piano/ The Shard	1	Multiple Matching	Original	5	10	20
			2	True/ False/NG	Original + Self constructed	4+1		

The length of the text, linguistic complexity, readability, topic of interest to the adult learners, and text type, were parameters considered when choosing appropriate texts in the test development. Even though there are several reading texts available on the ‘www.lrnnglobal.org’ site, only the above-selected texts were selected for the study, after a thorough scrutiny of numerous texts.

3.5.1.2 Categorizing Cognitive Processes of Reading

Reading sub-skills that are generally tested by established assessments have been investigated in the literature discussed in Chapter Two. Such assessments comprise the TOEFL iBT exam, the Diagnostic English Language Tracking Evaluation (DELTA), the Australian Diagnostic English Language Test (DELTA), the DIALANG test, and the Online English Assessment Framework (OEAS). Further, a number of taxonomies were also discussed in the previous chapter. However, the present study employs an internationally renowned modern framework for assessing reading suggested by Khalifa and Weir (2009). As discussed in the earlier chapter, the reading model based on the socio-cognitive validation framework is a well-established model worldwide, exclusively employed by Cambridge Assessment English.

Khalifa and Weir address controversial issues concerning the structure and content of Cambridge English for Speakers of Other Languages (ESOL) exams. They focus on the dynamic relationships between students’ cognitive skills, the test’s scoring criteria, and its assessment tasks that exhibit a wide variety of different contexts. In order to create reliable scoring assessments, they fundamentally contended that test developers must clearly explicate and attentively examine how reading comprehension expectations align with students’ learning needs (Mumin, 2011). Concerning the effectiveness and timeframe, the present research highlights the students’ cognitive validity and scoring validity in depth.

The eight socio-cognitive processes outlined by Khalifa and Weir (2009) were discussed broadly in Chapter Two, section 2.3.3. However, to refresh the reader, a brief description is given in Table 3.6.

Table 3.6 Cognitive Processing in Reading in Khalifa and Weir (2009)

Word Recognition (WR)	The reader identifies the same word in question or determines a word meaning independently and matches it in the text. This occurs at the word level.
Lexis Access (LA)	The reader uses knowledge of (morphology) word meaning or word class to identify synonym, antonym, hypernym, or other related words and matches it in the text. This occurs at the word level.
Syntactic Parsing (SP)	The reader uses grammatical knowledge to establish comprehension to identify answers without logical problems. This can occur at the clause or sentence level.
Establishing Propositional (core) Meaning (EPM)	The Reader expeditiously uses knowledge of lexis and grammar to establish the meaning of a sentence at the local level. It is a literal understanding of what is on the page. This occurs at the sentence or clause level.
Inferencing (I)	The reader goes beyond literal or explicitly stated meaning to infer a further significance. The reader can selectively read the paragraphs for main ideas and implicitly expressed ideas in the text. This can occur at the sentence level, paragraph level, or text level.
Building a Mental Model (BMM)	The reader uses several features of the text to build a larger mental model by recognizing major contrasts in a comparative and contrastive text type. This occurs at a whole text level.
Creating a Text Level Structure (CTLS)	The reader uses genre knowledge to identify the text structure and purpose of the whole text by analysing and distinguishing major ideas from supporting details. A trained reader decides how the various sections of the text work together, and which parts of the text are vital to the intent of the author or the audience. This occurs at the text level.
Creating an Inter-Textual Representation (CITR)	Understanding text and compare it across other texts. This occurs beyond the text level.

It is noted that the last two skills were not fully tested even in the “Cambridge ESOL” examinations (Khalifa and Weir, 2009, p. 70); with this in mind we can look into the items of the present study concerning item difficulty and test taker’s performance.

3.5.1.3 Test Review

The subject teachers’ judgments are the best way to determine the level of language difficulty of the texts (Weir et al., 2000). Further, the item writers must consult the subject specialists in order to interpret the test as “insider” readers (Fulcher, 1997, p. 132). Therefore, the selected texts along with the questions were validated by two experts, who are PhD holders gaining practical experience in researching reading and language assessment before those items were finally validated by experts. One of them was a professor at the Kulliyyah of Education in IIUM, while the other was a senior lecturer at SEUSL, having around 33 years of experience in teaching English.

These experts have validated the texts and items concerning text length, text difficulty, text type, the appropriateness of stem and distracters, and the amount of time needed for students to complete the test. They were also asked to verify the suitability of the texts for the intended samples, besides assessing the clarity and simplicity of the test instructions. As mentioned by Fulcher (2003, p. 390), “Terms used in instructions should be clear, simple and consistent”. Further, guided reading is considered a good mechanism to produce independent strategic readers, as highlighted by Ford and Opitz (2011), cited in Blything et al. (2020).

3.5.2 Empirical Evaluation

Preliminary investigation of the tests and piloting them enabled the researcher to collect empirical evidence.

3.5.2.1 Preliminary Investigation

Test validation, as well as identification of readability indices, were performed before the instrument was tested through piloting.

3.5.2.1.1 Test Validation

The tests are validated in three phases, using face validity, content validity, and a common item linking procedure.

3.5.2.1.1.1 Face Validity

The selected 13 texts along with their accompanying 155 original items and 7 self-constructed items were reviewed and validated by two experts, along with the researcher. In the process of reviewing, the items were examined to ensure their relevance and accountability. Further, they were studied to determine the relevance of the cognitive processes tested, based on what they require students to do (whether they require students to recognize the word, identify the lexical category, identify the grammatical knowledge, establish a propositional meaning, infer the idea, build a mental model, create a textual structure, or create inter-textual representations). The following two questions were asked in determining the suitability of the items:

- i. Does the item examine the specified cognitive process required from the readers to answer it?
- ii. Do the stem, key, and distractors function well in the development of an item?

Only the items that were considered suitable were selected. Further, the test materials taken from the CEFR-accredited LRN tests, which have already been validated by a team of experts, were finalised accordingly. Once all the items were included, the whole set of tests was given to experts for content validation.

3.5.2.1.1.2 Content Validation

Together with face validity, content validity (CV) is the minimum quality requirement for instrument development at the item development stage (Halek et al., 2017). Content validity means “the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose” (Haynes et al., 1995, p. 238). In simple terms, it can be stated that a test should be able to measure what it intends to measure, as highlighted by Turner and Carlson: “An important component in test development is providing evidence that the items created are measuring the content or construct they are defined to measure” (Turner & Carlson, 2003, p. 164) and Hughes (1989).

One method of accomplishing content validity includes the judgment of a board of subject specialists thinking about the significance of individual items within an instrument. Lawshe's technique, proposed in 1975, has been generally used to set up and evaluate content legitimacy in different fields, comprising health care, education, organizational development, personnel psychology, and market research (Ayre & Scally, 2014). Crocker and Algina (1986) added three more steps to Lawshe's method, suggesting four main steps to identify the content validity of an instrument. The CV processes consist of defining the performance domain of interest, selecting a trained panel of experts in the content area, offering a structured framework for the process of rating or matching items to the performance domain, and gathering and summarising data from the rating or matching process.

Though CV is crucial in measurement, its procedures are not well discussed in many research studies (Crocker, 2003). Content validation cannot be written within a paragraph suggesting that the items are good enough according to expert judgment, as discussed in many research surveys. Therefore, the present study selected a joint approach to validate the content. In this regard, both qualitative and quantitative methods were employed to validate the content, as per Creswell (2012).

3.5.2.1.1.2.1 Qualitative Method

Expert judgment is a qualitative act applied to the investigation of language testing research, among other aspects like “introspective and retrospective verbal reports, observations, questionnaires, and interviews, as well as text analysis, conversational analysis, and discourse analysis” (Bachman, 2000, p. 7). Content validation (CV) is done through expert judgment; in other words, it can be obtained through the involvement of a group of subject matter experts (SMEs) (Creswell, 2012; Crocker & Algina, 1986). Moreover, CV is crucial for an assessment tool, when its scores are used as evidence in making decisions to select the examinee for the enrolment of an educational or occupational prospect, or promotion (Wilson et al., 2012).

The test designers begin with a definition of the content or ability domain through the format of a ‘table of specifications’ from which they decide on items and test tasks to develop tests. Therefore, the consideration of test “content” is thus an important part of both test development and test use, and an analysis of the content of a test is highly significant for “validation” (Bachman, 1990, p. 244). Because the judgment of the experts is utilized to confirm the appropriateness and specifications of the item. Furthermore, to define the construct, it is essential “to engage the expertise of a subject matter specialist in the design and the development of the language test” (Bachman & Palmer, 1996, p. 96).

On the other hand, some scholars do not concur with the practice of CV. Compared to the extant literature for CV, the literature against CV is quite sparse. In realising the significance of CV in test development, Turner and Carlson (2003) cited Hambleton (1978), who more specifically identified content validity as being item validity. That is why CV is widely utilized in language testing research (Alderson et al., 1995). With this consideration, the present research also employs CV for the tests. The request for the expert judgment was sent to thirty experts who have ample experience in English language teaching, language testing, assessment of reading, and development of test materials around the world, considering the recommendations suggested by Crocker et al. (1988) in appointing the expert panel. Of these, twelve experts consented to participate in the validation process, free of charge.

The design of the study includes four different reading testlets. Each testlet has four CEFR-aligned texts belonging to the Learning Resource Network (LRN). All texts have a maximum of nine, ten, or eleven questions. Each of the questions was rated under Khalifa and Weir's (2009) socio-cognitive processes of reading. Altogether there were around 162 questions, originally taken from the LRN, and a few questions self-constructed by the researcher. As there were too many texts and questions in total, it was decided to get the content validation separately for each test, so that the experts who volunteered to do the validation, are not unnecessarily inconvenienced by having to validate the whole lot. A summary of the present research, a letter from the Post Graduate (P.G.) Office of the Faculty of Education, IIUM, requesting the appointment of an expert panel, a testlet with the key, and the Item Objective Congruence (IOC) rating sheet, were attached with the first email correspondence to the experts. Subsequently, only twelve experts expressed their willingness to participate in the validation process.

A test was validated by a minimum of three experts, as feedback from at least three judges for each task is recommended for better rater performances (Crocker et al., 1988; Fulcher, 1997). Following the above recommendation, a dozen experts were actively involved in the validation process. Since the present research was conducted during the COVID-19 pandemic, the researcher utilized the full capacity available through distance education. During the duration of the validation process, a minimum of fifteen to twenty email communications were conducted between the researcher and some of the raters. Alderson et al. (1995, p. 63) mentioned that “they must take each item as if they were the students”; thus, it took them a considerable amount of time to give their judgment. With some raters, the researcher had online meetings to clarify the doubts that arose while validating. The researcher is most grateful for the services provided by the expert panels. Table 3.7 below shows some brief information on the experts participating in the CV.

Table 3.7 Descriptions of the SMEs

Demographic information (Variables)		N	%
Affiliation	IIUM	3	25.0
	UniMaS	1	8.3
	UniSZA	1	8.3
	UTM	1	8.3
	University of Bedfordshire - CRELLA-UK	1	8.3
	Uni of Kelaniya -SL	2	16.7
	Uni of Colombo- SL	1	8.3
	SEUSL-SL	2	16.7
Qualification	Post-Doctoral	1	8.3
	PhD.	9	75
	PhD reading	1	8.3
	M.A reading	1	8.3
Teaching & Language Testing experience	>=30 years	2	16.7
	20-29 yrs	3	25
	10-19 yrs	3	25
	0-9 yrs	4	33.3
Gender	Male	1	8.3
	Female	11	91.7

Through this content validation of the above expert panel, the researcher may answer the question raised by Haynes, et al. (1995), whether the test measure what it intends to measure. The results of the content validation are discussed in the following sections, which include detailed information on the quantitative content validation procedure.

3.5.2.1.1.2.2 Quantitative Approach

For Haynes et al. (1995, pp.238-239) “*the degree to which*” refers to the fact that content validity is a quantitative-based judgment. Further, from the quantitative perspective, establishing content validity does not involve a rigid procedure. Several methods are available, but a common practice adopted by most researchers is to use a method whereby the agreements and disagreements among the judges are compared. From there, the items are harmonised by deciding whether to retain, remove or revise them. In the case of this study, once the experts’ opinion was carefully reviewed and necessary changes were made, the items were subjected to further analysis to quantify the judgments of the experts. Crocker et al. (1989) recommended two methods of validation. One method is to utilise approaches like the percentage of items or the index of relevance to examine the overall fit between the test and the curriculum. Another approach is to assess how well individual items fit into a content domain.

Item objective congruence, validity index, and content validity ratio are examples of techniques that fall under this second category. The item objective congruence (Rovinelli & Hambleton, 1977) method was chosen for this study because it allowed for the quantitative evaluation of individual items.

3.5.2.1.1.2.2.1 Item Objective Congruence (IOC) and the Analysis of its Results

One of the popular methods to analyse content validation is Item Objective Congruence (IOC) introduced by Rovinelli and Hambleton (1977); for them, the IOC is a technique that enables content validity to be quantified. This approach provides ample information in test development, providing evidence that the items assess the content of the construct they are supposed to measure. Turner and Carlson (2003) cited Berk (1984) that he believed “that an evaluation of the match between items and objectives is the most important assessment during the content validation stage” (p.164).

The majority of experts selected in this study were familiar with the construct and the socio-cognitive processes of reading; for the rest, the specified cognitive processes were explained. They rated each item according to its objective and the types of reading required to answer the question. An item objective congruence index was established from their ratings to assess the fitness of each item against its intended objective. The IOC index measurement is based on the degree to which an item measures (or does not measure) a particular objective. Some of the experts in this study rated more than one objective for some items; therefore, the formula proposed by Martuza (1977) or Rovenelli and Hambleton (1977) on the assumption that there is only one valid objective being measured by each item, is not appropriate to produce a reliable cut-score. Consequently, the multidimensional item formula simplified by Crocker and Algina (1986) in their *Introduction to Classical and Modern Test Theory*, was utilized to evaluate the congruence between an item and a set of objectives.

The equation for the adjusted index is as follows:

$$I'_{ik} = \frac{(N)\mu_k - (N - p)\mu_l}{2N - p}$$

where I'_{ik} is the index of item-objective congruence for item i on a set of objectives k , N = the number of objectives, p = the number of valid objectives, μ_k = the judges' mean rating of item i on the valid objectives k , and μ_l = the judges' mean rating of item i on the invalid objectives l . (Crocker and Algina (1986) as cited in Turner and Carlson, 2003, p. 169)

Using this formula, the items of this survey were analysed to receive the accepted congruence, as shown in Table 3.8, as an example for comparing interpretations for the first three items of the common items used in all four tests. The detailed judges' ratings for all 162 items are attached in Appendix C (II). This output also consists of the item number, the index value for the valid objective, and the average ratings of the three experts on each objective.

Table 3.8 Sample of IOC Indices for The First Three Common Items of the Tests

Item	<i>Index of Item– Objective Congruence</i>	Objectives							
		1	2	3	4	5	6	7	8
20	0.956	-1.00	-1.00	-0.33	1.00	-1.00	-1.00	-1.00	-1.00
21	0.556	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	-1.00	-1.00
22	0.556	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00

In the current research, raters were briefed on the rating procedure, and assigned a rating from 1 to 3 for each item for its objectives, based on Khalifa and Weir’s (2009) cognitive processing in reading and type of reading, as follows:

1. that the item definitely measures the objective
2. uncertain whether the item measures the objective
3. that the item does not measure the objective

They were not informed “which item is meant to be matched with which objective”, as recommended by Osterfind (1997, p. 259). Hence, the judges freely measured the items. Once they completed the task, which took them some time, they emailed the filled rating sheet back to the researcher. Further, these ratings were then coded by the researcher according to the criteria suggested by Turner and Carlson (2003, pp. 164-165).

The range of the index score for an item is –1 to 1, where a value of 1 indicates that all experts agree that the item is clearly measuring that objective and is clearly not measuring any other objective. A value of –1 for the “valid” objective would indicate that the experts believe the item is measuring all objectives that it was not defined to measure and is not measuring the hypothesized objective.

Following this, in Table 3.8, Item 20 indicates that all experts agreed that the item is measuring Objective 4 and not measuring Objectives 1,2,5,6,7, and 8, and also that the item has a high IOC value of 0.956. However, one of three experts believed that the item is a measure of Objective 3. Similarly, Item 21 has a valid IOC of 0.556, which is an accepted value according to Brown (2005) and Takwin et al. (2018), although Pengruck et al. (2019) used IOC indices of 0.60 to 1.00 as accepted values. Brown (2005) mentioned that if the index of the IOC falls between 0.5 and 1.00, it means that the item is acceptable, but if the IOC falls below 0.5, it means that the item is not fitting and should be reviewed or removed. Further, this view was affirmed by the designers of the IOC, Rovinelli and Hambleton (1977, pp. 15–16) that “if an item is to be a perfect match to an objective, while the others were not able to make a decision, the computed value of the index would be 0.50”. It means that Item 21 was agreed by two experts that it is clearly measuring Objective 2 and not Objectives 1,5,6,7, and 8, whilst one expert agreed that this item can measure Objectives 3 and 4. That is why the average rating of all three experts for Objectives 3 and 4 has a congruence of -0.33 for Item 21. If any of the experts are uncertain of the objective, whether the item measures it or not, then the congruence value indicates 0.0.

Out of 162 total items, only 4 items were identified with low IOC indices of less than 0.5 (See Appendix C (II)). This resulted from a scenario in which judges either rated more than one objective or were not certain of the valid cognitive processing in reading for those items. Nevertheless, this is normal as some questions may be categorised as measuring more than one cognitive category, while there are more objectives, as there are eight in the present study (Kim, 1996). However, based on the experts’ remarks, and concerning the uniformity of the four tests among four different faculties, two items were removed after the experts’ judgment before they were administered in the pilot tests. Table 3.8 illustrates the summary of the experts’ IOC evaluation relating to each item measuring the cognitive processing in reading. Apart from Test 1, all three tests show a similar number of items, indicating LOT and HOT skills. The following facts explain how they are equivalent. First, all Test 1 and Test 3 texts are taken from the same CEFR-level tests belonging to the LRN validated tests. The second valid reason is that all these four tests are going to be horizontally linked using the process of common item calibration. Therefore, it is acceptable to say

that the tests are equivalent. To summarise, it can be safely concluded that almost all the items clearly measure or somewhat measure the intended socio-cognitive processes of reading.

Table 3.9 Summary of Cognitive Processes of Reading of Each Test According to IOC Indices

	WR	LA	SP	EPM	I	BMM	CTLS	CITR	Total	LOT	HOT
Test 1	2	3	9	15	1	2	8	0	40	29	11
Test 2	3	2	7	10	5	9	3	1	40	22	18
Test 3	1	3	11	8	4	9	4	0	40	23	17
Test 4	0	7	7	9	3	9	5	0	40	23	17
Total	6	15	34	42	13	29	20	1	160	97	63
%										61	39

3.5.2.1.1.3 Common Item Linking Procedure

The present study intends to profile the performance level of students of four different faculties; therefore, based on the samples' interests and their familiarity with the contents, several texts have been selected to ascertain the reading performance. Therefore, around 13 CEFR-related texts associated with different contexts such as sports, technology, travel and transport, food and agriculture, education, literature, business, and holidays, were selected by the researcher. However, testing the students on all these texts and items is unnecessarily tedious, and the lengthy tests will lead to fatigue (Wells & Wollack, 2003). This would negatively affect the validity of the responses to the items. Thus, there was a great need to develop different tests and connect them using a common item technique.

Linking tests using common items will minimise exhaustion and safeguard the validity of the answers since only a small number of items are needed to be addressed by the test takers. In addition, without making the test unnecessarily long, it allows for more items to be field-tested. Furthermore, it is the easiest and simplest way of equating (Linacre, 2020c). Therefore, four different tests were designed to ensure that

the tests are equal in test difficulty level using the common item linking procedure illustrated by Leon (2008).

Opinions differ regarding the optimum number of common items for linking various testlets. The number can, for example, be as low as five items, whilst ten items are considered ideal for the function (Ingebo, 1997). However, the provision of at least 20 % common items from the complete test was proposed by Kolen and Brennan (2013), while North (2000) suggested a size of 30% common items from total items in adjacent experiments. Nevertheless, the decision should be based not only on the number of items or the percentage but what is more important lies in the essence of the quality of the selected items (Wright and Stone, 1979). This can be achieved by selecting the best common items that are representative of the test.

The texts selected for these four tests belong to CEFR levels B1, B2, C1, and multi-level IELCA texts. While arranging the texts according to item difficulty in an ascending or descending order out of these five level texts, the C1 level is suitable for equating, as the items do not represent the easiest or the hardest level of the item difficulty. At the same time, Khalifa and Weir's (2009) cognitive validity of the socio-cognitive validation framework is based on the HOT and LOT skills of meta-cognition. The first four processes belong to the LOT skills whereas the others represent the HOT skills. Further, as the items are based on a particular reading text, without violating the local dependency rule, only a limited number of items can be produced from the same text. All the tests contain only the Selected Response (SR) format, not the Constructed Response (CR) format. Constructing the response on test takers' understanding takes much time, therefore, taking the CR as a common item may require more effort and is not feasible. To ensure that the result of the respondents are reliable, Multiple Choice (MC) items that are under the SR category would be more advisable as common items.

Among the commonly known linking methods, like virtual equating, vertical equating, polytomous equating, separate-estimation equating, random-equivalence equating, alternate-forms equating, and anchored-form equating, horizontal linking is equating tests that are similar in difficulty level, as illustrated by Linacre (2011, p. 2) that in horizontal linking "the two tests are intended to have the same difficulty", and

by Skaggs and Lissitz (1986, p. 496) that “The tests are written to measure the same "ability" at a comparable level of difficulty”. Since all four tests in the current study are similar in test difficulty, this research involves horizontal equating including eleven items as common items from the text belonging to CEF C1 level, labeled ‘Holidays’. This number is considered sufficient for the linking procedure as it falls between 10-20 items (Wright & Stone, 1979). Linking research design is considered to be the most realistic since test takers do not answer all 13 texts, in the sense that they only have to answer only four texts consisting around 40 items. Therefore, tiredness or boredom, that is often associated with answering lengthy tests, is minimised.

As suggested by Wright and Stone (1979, p.101), “With four or more tests we can construct a network of loops”. Figure 3.2 explains the network of 10 tests using nineteen networks.

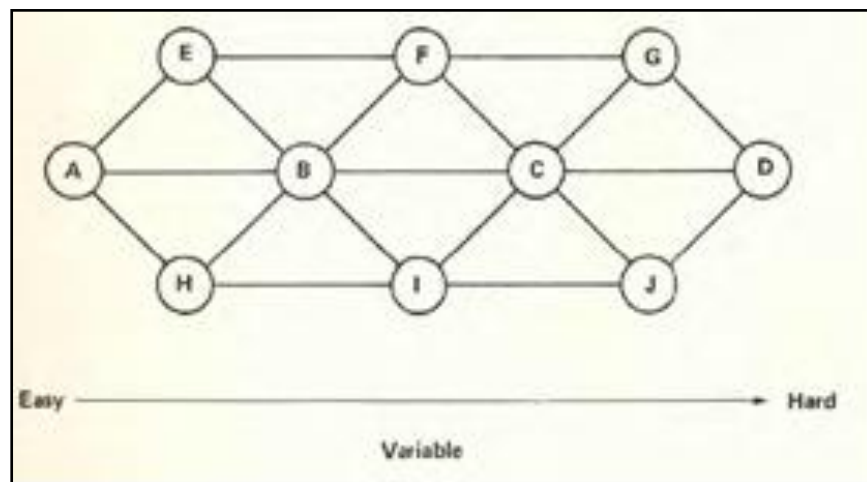


Figure 3.2 Networks of Tests (Adopted from Wright & Stone (1979, p. 101))

Common Item Linking Design

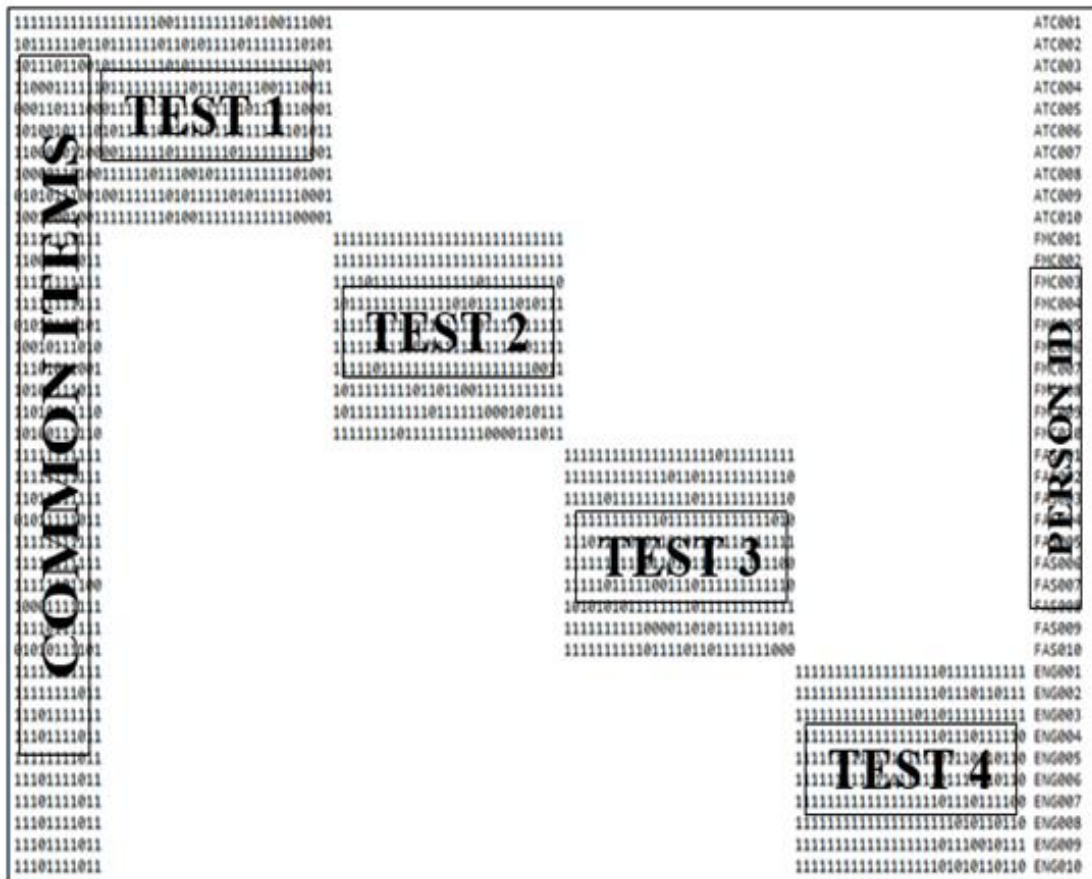


Figure 3.3 Linking Procedure Using Common Item Equating (Concurrent Analysis) in the Current Study

There are two ways the common items can be linked in Rasch MM. One is using the values as anchor (or fixed) values in an analysis of the other Form. This puts the measures of the second Form into the measurement frame of reference of the first Form. The second way is combining the data from the two Forms into one analysis (concurrent equating) in which all the items and persons are measured in the same frame of reference (León, 2008). The present research applies concurrent analysis for linking the common items

According to León (2008), an equation for analysing this concurrent is as follows:

Form B rescaled as Form A = (mean of Form A common items) + (Form B measure - mean of Form B common items) *S.D. of Form A common items / S.D. of Form B common items. (León, 2008, p. 1172).

Table 3.10 shows the way how common items were shared among all four tests. The highlighted 11 items are the common items in all the tests. Each test has 29 unique items with the 11 common items, thus the total of each test is 40 items.

Table 3.10 Common Items in all Four Tests

Tests		Items		
1-FAC	10	9	11	10
2-FMC	10	9		10
3-FAS	10	9		10
4- FE	10	9		10

3.5.2.1.2 Text Inspector Analysis for Readability

The most common methods of assessing readability involve examining the word and sentence length. Gunning Fog Index, Flesch Reading Ease, Flesch-Kincaid Grade level, Fry graph, and Smog are the commonly known methods to assess reading text difficulty (Klare, 1984).

Table 3.11 indicates the readability indices of the selected texts based on the analysis of the Text Inspector software which is an award-winning professional text analysis tool. This tool helps to get reliable information about the lexical composition, level of text difficulty, and its overall level with the CEFR (Bax, 2012). Using Text Inspector, the texts were analysed according to the Flesch Reading Ease and CEFR aspects. The Flesch Reading Ease is measured using a ratio of total words, sentences, and syllables. Texts that are easier to read will have higher measures (up to 120), whereas more complex texts will have lower scores (below 40). It indicates that the

texts selected are ranging from easy to difficult without meeting the extreme ends. ‘Having friends’ is the easiest text with an index of 76.46, whereas ‘London Olympics’ is the most difficult text (40.52).

Table 3.11 Readability Index according to Text Inspector Analysis

Text No	Text Title	LRN CEFR Level	Text Inspector CEFR level	Flesch Reading Ease	Token count	Sentence count	Test FRE total
Test 1 P1	Ice Cream	B1	B1+	71.93	150	8	237.91
Test 1 P2	Jet Lag	C1	B2+	70.66	458	27	
Test 1 P3	Holidays	C1	C1+	46.25	385	18	
Test 1 P4	Technology	IELCA (M/L)	C1+	49.07	796	30	
Test 2 P1	Educational Programmes For Adults	B2	B2+	45.86	140	8	184.89
Test 2 P2	The Tradition of Coffee Drinking	B2	C2	52.26	405	16	
Test 2 P4	London Olympics	IELCA (M/L)	C2	40.52	733	22	
Test 3 P1	Playing Outdoors	B1	B1+	61.89	150	10	200.44
Test 3 P2	Supersonic Flight	C1	C1+	50.13	396	18	
Test 3 P4	Sir William Empson	IELCA (M/L)	C2	42.17	693	31	
Test 4 P1	Having Friends	B2	A2+	76.46	180	12	241.55
Test 4 P2	The magic of the Cinema	B2	B2	62.19	420	15	
Test 4 P4	The Shard	IELCA (M/L)	C1	56.65	659	35	

The CEFR level of each text, according to the LRN validation, is given in the third column of the above table. The fourth column indicates the CEFR analysis of the Text Inspector. While comparing the LRN CEFR level to the Text Inspector CEFR level, (except for one text titled ‘Having Friends’), all eleven texts indicated a similar or a higher level of CEFR based on the Text Inspector analysis. The text ‘Jet Lag’

belongs to B2+ according to the Text Inspector, while it indicated a C1 level according to the LRN; however, the Text Inspector B2+ can be considered similar to the C1 level. Thus, the texts in the proposed tests are put together based on three assumptions that they are comparable in text difficulty level. One is based on the CEFR level of the LRN validation and the Text Inspector analysis. The next is based on the Flesch Reading Ease matrix. Finally, all the tests are joined together with a common text, including eleven items in all tests.

3.5.2.2 Pilot Study

The purpose of the pilot study is to determine the level of difficulty of the items and to assess the correspondence between the estimates of the empirical difficulty and the level of difficulty obtained from the expert judgment. Another reason for piloting is to check whether the items are able to evaluate the ability level of the target population. Further, it was recapped by Fulcher that “This is a reminder that, no matter how experienced one may be in test development, there is always a need for pretesting all items before tests become operational” (1997, p. 116). Further, as Boone (2016) suggested, “Following the thoughtful construction of the measurement instrument, our researcher should collect pilot data, conduct a Rasch analysis of the pilot data, and then refine the instrument” (Boone, 2016, p. 4). With these guidelines, this study decided to collect the data for piloting and conduct a Rasch analysis to refine the instrument to finally be ready for operation.

Using the representative sampling method, a total of 124 students were selected from four faculties, namely the Faculty of Arts and Culture, the Faculty of Management and Commerce, the Faculty of Applied Sciences, and the Faculty of Engineering of SEUSL, for piloting four different tests; 30, 30, 34, and 30 samples, respectively, were selected from each faculty. According to Linacre (2020), a minimum of 30 samples for a test consisting of 40 items is adequate to provide insightful results if the selection of the samples is effective.

All four tests have 40 items in each, and they all are selective response types, including MCQ, multiple matching, and yes/ no, or not given varieties of questions. Texts and items were taken from the LRN materials; however, some items were constructed by the researcher. Two items were deleted after the expert judgment before they were administered for piloting. The test used the scoring method of 1 for correct response and 0 for incorrect items as it is handled in the dichotomous data of the Rasch measurement model. And the total score of each test is 40 marks. Among the number of software available to analyse the data received, *WINSTEPS* version 4.4.7 was used due to its effectiveness in investigating.

3.5.2.2.1 Data Analysis Procedure for Pilot Study

In the beginning, common items including only one reading text and 11 items were piloted before they were added as the common items to linking. The test, consisting only of common items, was initially conducted among a few assistant instructors of DELT, SEUSL, and two IIUM PG students. Then it was administered to the students. The internal validity of the individual tests was determined, and subsequently, the concurrent analysis of the four tests was examined for their validity and reliability.

Since the tests fall under the category of dichotomous data, the following formula informed by Reach (1980), satisfies the linearity, stochasticity, and conjoint additivity requirements for useful measurement, implied in this study.

$$P_{ni}\{x_{ni}=1/B_n, D_i\} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)].$$

Where $P_{ni}\{x_{ni}=1/B_n, D_i\}$ is the probability of person n on Item i scoring a correct ($x=1$) response rather than an incorrect ($x=0$) one, given person ability (B_n) and item difficulty (D_i). This probability is equal to the constant e , or natural log function (2.7183) raised to the difference between a person's ability and an item's difficulty ($B_n - D_i$), and then divided by 1 plus this same value (Bond & Fox, 2015, pp. 497-8).

This model is pertinent in the context of making measurements in the case of correct or wrong test items, considering the difficulty of each test item along with the overall ability level of a test taker, with respect to the single variable; in this scenario, reading construct.

3.5.2.2.1.1 Common Item Linking

There are three methods to calibrate equating, such as common item calibration, common person calibration, and common scale calibration (Kolen & Brennan, 2013). For Wright and Stone (1979), common item linking is more desirable than common person linking as it is the more cost-effective way (Wright, 1978). The common items are used as connecting items to obtain common values from the various tests so that the results of the various assessments could be directly compared (Wright & Linacre, 2001).

As this research uses common items to connect the various sets of test forms, validating these items to ensure that they are accurate is an essential step to be taken. Using the scatterplot is the most common technique to calibrate the items (Ingebo, 1997; Wright & Stone, 1979). However, because of the vast number of test items involved in this analysis, this research used concurrent analysis to examine all the items. Another significant focus of the pilot study was to examine the functioning of common items. As the stability of the calibration depends on the quality of common items used, this is an important factor to consider. The eleven common items that reflect both LOT and HOT cognitive processes in reading (intended to be evaluated), were examined in this analysis. The results of this analysis are shown in Table 3.12.

Table 3.12 Item - Person Reliability of 11 Common Items

	Total score	Total Count	Measure REAL	S.E	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	6.3	11.0	0.51	0.88	1.02	0.1	0.92	0.0
P.SD	2.6	0.0	1.58	0.33	0.26	0.8	0.37	0.6
REAL RMSE	0.94	True SD	1.27	Separation	1.36	Person Reliability	0.65	
Mean	70.9	124.0	0.00	0.24	1.00	0.1	0.92	-0.3
P.SD	19.7	0.0	1.09	0.04	0.18	1.8	0.30	1.3
REAL RMSE	0.24	True SD	1.06	Separation	4.34	Item Reliability	0.95	
Person Logit Range		-2.83 to 2.66						
Item Logit Range		-2.65 to 1.35						

According to the Rasch analyses, the common items almost met the measurement requirements proposed by Linacre (2020), except for a few cases. To meet the measurement requirement, the item and person reliability indices, and unidimensionality statistics should statistically be analysed. Unidimensionality is measured through three indicators such as item polarity, item fit, and PCA residuals (Bond & Fox, 2015; Linacre, 2020). Consequently, person and item reliability should be $>.80$, as the data implicate a dichotomous scale (Bond & Fox, 2015; Linacre, 2020).

It is clear that although item reliability was very high at 0.95, person reliability was a bit low, at 0.65. The reason for the low person reliability may probably be because of the small number of items ($n=11$) included in this analysis. The separation indices for both person and items were recorded at 1.36 and 4.34, respectively, which is also not far from the recommended value of >2 for the person (examinee). Item difficulty spanned about 4 logits (-2.65 to 1.35) while the person ability estimates spanned about 5.49 logits (-2.83 to 2.66).

Unidimensionality of Common Items

Table 3.13 illustrates the principal component analysis of standardised residual variance for common items (see Appendix F.1.b). Though the total variance clarified by measures is 30.3%, this does not affect the application of Rasch. As Linacre (2018) noted, the amount of variance explained by the Rasch dimension is irrelevant. The eigenvalue of unexplained variance in the 1st contrast is less than 2 items strong, indicating 1.8157.

Table 3.13 Principal Component Analysis of Standardised Residual Variance For Common Items

	Eigenvalue	Observed	Expected
Total raw variance in observations	15.7820	100.0%	100.0%
Raw variance explained by measures	4.7820	30.3%	31.3%
Raw variance explained by persons	2.3393	14.8%	15.3%
Raw Variance explained by items	2.4427	15.5%	16.0%
Raw unexplained variance (total)	11.0000	69.7%	100.0%
Unexplained variance in 1st contrast	1.8157	11.5%	16.5%
Unexplained variance in 2nd contrast	1.4724	9.3%	13.4%
Unexplained variance in 3rd contrast	1.4040	8.9%	12.8%

Further, based on infit, outfit, and correlation indices obtained, as shown in Table 3.14 (see Appendix F.1.c), item CI28 has an infit MNSQ 1.36 and outfit MNSQ 1.52, which is not far away from the recommended value of 0.7-1.3 for cognitive tests (Bond & Fox, 2015; Linacre, 2020); and similarly, item CI27, overfitted indicating 0.37 outfit MNSQ. However, the particular items had been investigated further, and as per expert consultancy, the wording of the items had been modified. The point measure correlation coefficients showed very good statistics. All the items were above 0.30, which is a piece of satisfactory evidence that the items were useful indicators for measuring the cognitive processing in reading and student performance. Further, the Wright Map for common items is given in Appendix F.1.d. Therefore, all eleven common items were retained after modification of the two aforesaid items.

Table 3.14 Item Statistics for Common Items

ENTRY	MEASURE	IN.MSQ	IN.ZSTD	OUT.MSQ	OUT.ZSTD	PTMA-E	RMSR	NAME	
9		1.00	1.36	3.49	1.52	2.49	0.35	0.54	CI28
3		1.35	.75	-2.58	0.61	-1.96	0.67	0.54	CI22

3.5.2.2.1.2 Internal Validity Analysis of Individual Tests

To calibrate item difficulty and determine the examinee’s ability to see how the test performed individually, each test was analysed separately. Checking for the reliability and validity of the test is crucial since it helps to take safety measures to sort out the issues in the future. Reliability, separation, item polarity, item map, dimensionality, etc., were examined in this analysis.

Person and Item Reliability

The high reliability of the test indicates that it is highly likely that person ability ordering could be replicated with items of the same difficulty. The findings of the individual tests as shown in Table 3.15 demonstrate good statistics for reliability and separation indices for both person and item (see Appendix F.2.a), except for Test 1 which had 0.74 person reliability and 1.69 person separation indices. This may be because of the smaller sample size (30), and difficulty recognising suitable persons for the study. When determining the proper size of samples, the person and item reliability may increase dramatically. Further, the person separation index (which was > 2.00 for the majority of the tests) is a good indication of acceptable reliability (Bond & Fox, 2015; Linacre, 2018).

Table 3.15 Summary of Person and Item Reliability of Four Tests of Pilot Study

Test Faculty	No of items	No of persons	Person reliability	Item reliability	Person separation	Item separation
1 Arts & Culture	40	30	.74	.86	1.69	2.44
2 Management & Commerce	40	30	.80	.84	2.00	2.26
3 Applied Sciences	40	34	.83	.83	2.22	2.18
4 Engineering	40	30	.83	.82	2.19	2.10

Dimensionality map

The Principal Component Analysis (PCA) of Standardized Residuals supported the unidimensionality of the reading construct as there was no clear secondary factor extracted. “According to Linacre’s recommendations for unidimensionality, the variance explained by measures must be above 40%, with unexplained variance less than 15% in the first contrast, on the strength of at least 3 items” (Isa et al., 2016, p. 4). This is considered to be acceptable according to Rasch simulation. The findings of the PCA of Standardised Residuals for all four tests are given in Table 3.16 (see Appendix.F.2.b).

Table 3.16 The PCA of Standardised Residuals for all Four Tests

Test	Faculty	Raw variance explained	Unexplained variance in I st contrast	Infit	Outfit	Point-measure correlation
1	FAC	30.8%	7.5%	All items fall between 0.70-1.30	All items fall between 0.70-1.30 except for 5 Items >1.30 5 items <.70	No negative correlations 11 items were <.30 but >.13
2	FMC	35.8%	10.1%	All items fall between 0.70-1.30	All items fall between 0.70-1.30 except for 7 Items >1.30 7 items <.70	No negative correlations 19 items were <.30 but >.12
3	FAS	27.2%	8.4%	All items fall between 0.70-1.30	All items fall between 0.70-1.30 except for 4 Items >1.30 7 items <.70	No negative correlations 7 items were <.30 but >.12
4	FE	40.2%	9.6%	5items >1.30 7 items <.70	5 items >.130 18 item <.70	No negative correlations 3 items were at 0.26

Unexplained variance in the 1st contrast for all four tests is acceptable according to Linacre (2020) and Bond and Fox (2015), indicating less than 10%, even though Test 2 showed 10.1%, which is not far from the acceptable measurement requirement. Further, the unidimensionality requirement, with above 30% for raw variance explained, is considered a moderate measurement dimension (Conrad et al., 2011). In light of this literature, this area seems to be nonproblematic at this empirical stage. Therefore a “secondary dimension in the data” does not clearly “appear to explain more variance than is explained by the Rasch item difficulties” (Linacre, 2020c, p. 414).

The positive point-measure correlation coefficients of all four tests provided evidence that items on the test were working together in defining the reading construct. Further, “The positive residuals indicate unexpectedly high responses, while the negative residuals indicate unexpectedly low responses” (Smith, 2003, p. 21). Only a few items in some tests were identified with <.30 value; however, they are not

less than 12. While further investigating this consequence, it was observed that the items were somewhat easy in the item difficulty level.

Fit statistics, in the form of infit and outfit mean-square, were applied to ensure that the items were contributing meaningfully to the measurement of the variable or construct as expected by the model. The items within the recommended ranges 0.7-1.3 are considered meaningful to the measurement; whereas, the values below the ranges are considered as overfitting, and those above the ranges are considered as misfitting (Bond & Fox, 2015; Linacre, 2020c). Almost all items out of 127 unique items have a great infit mean-square, whereas, only twelve items showed different values. This may be because of the high or poor score of the students for the particular questions. Sometimes, the high performance for the difficult item can happen because of guessing (Bond & Fox, 2015; Boone, 2016; Boone et al., 2014; Linacre, 2020c).

Hence, none of the items were removed after the piloting as the person and item reliability in all four individual tests, the unidimensionality showed good statistics, and the items do not pose serious threats to measurement.

Table 3.17 Person Statistics: Misfit Order

Test	Faculty	Infit	Out fit	Ptm Correlation
1	FAC	All persons fall between 0.70-1.30 except for 1 person >1.30 1 person <.70	All persons fall between 0.70-1.30 except for 5 persons >1.30 8 persons <.70	No negative correlations All persons were >.30
2	FMC	All persons fall between 0.70-1.30 except for 8 persons >1.30 7 persons <.70	All persons fall between 0.70-1.30 except for 9 persons >1.30 10 persons <.70	No negative correlations All persons were >.30
	FAS	All persons fall between 0.70-1.30 except for 1 person >1.30	All persons fall between 0.70-1.30 except for 4 persons >1.30 5 persons <.70	No negative correlations 3 persons were <.30 but >.12
4	FE	All persons fall between 0.70-1.30 except for 6 persons >1.30 10 persons <.70	All persons fall between 0.70-1.30 except for 8 persons >1.30 17 persons <.70	No negative correlations All persons were >.30

Similarly, as in item fit statistics, Table 3.17 depicts the person statistics for all four tests individually. Person fit statistic shows how consistent the test taker (person) is in answering the questions. It provides empirical guidance on how well the model can predict the pattern of the person's answers (Bond & Fox, 2015). The person fit or consistency illustrates that the person answers all the items with a difficulty level below their capabilities and does not answer any item with a difficulty level above their capabilities. Also, person misfit or inconsistency can occur due to carelessness, guessing, or dishonesty/cheating (Bond & Fox, 2015; Wolfe & Smith, 2007).

However, these misfitting persons or items were not removed, but they were investigated further as the sample size could be the contributing factor in this case (Keeves & Alagumalai, 1999). The outfit indices suggest that any noise in the data is likely to be due to guessing or carelessness that can be remedied.

The results of the pilot study indicate two things. One is that the items work together to measure the reading construct and they are considered productive for measurement. The second one is that the student population selected in this study is reliable to produce consistent results, except that the limitation of fewer samples can reduce the reliability index, as indicated in the results. Overall, the tests proved to be productive for measurement.

3.5.2.2.1.3 Concurrent Analysis of all Four Tests

After checking for the statistics for common items and individual tests, the last stage is to examine the acceptability of the combined tests using concurrent calibration analysis. This analysis can be applied using the common item linking procedure. In this analysis, all four tests are analysed together in one common scale, whereby the estimation of item parameters in all the tests are done simultaneously.

In large-scale testing programs, it is inevitable to create more than one form of a test to be administered in a different place at different times. It is critical in these instances that all of the forms assess the same skill, ability, or trait, maintaining the same content and statistical specifications for all. Even though these forms are carefully created, inconsistencies between the tests may persist to the point that the

scores from the forms cannot be interchanged without the process of equating. Equating is used to ensure that scores from different versions of a test are comparable (Tsai et al., 2001). As the common items were included in all tests, there are larger responses for these items compared to any other items. One unique feature of the concurrent analysis is that it treats all forms as equally discriminating and influencing in the testing process. In all four tests, common items were used to create a common frame of reference to make a comparison among all items of the different testlets, so that an examinee can answer only one testlet out of the four testlets administered to four different faculties. However, they can be compared with the other examinees who took the different versions of the reading tests in this survey.

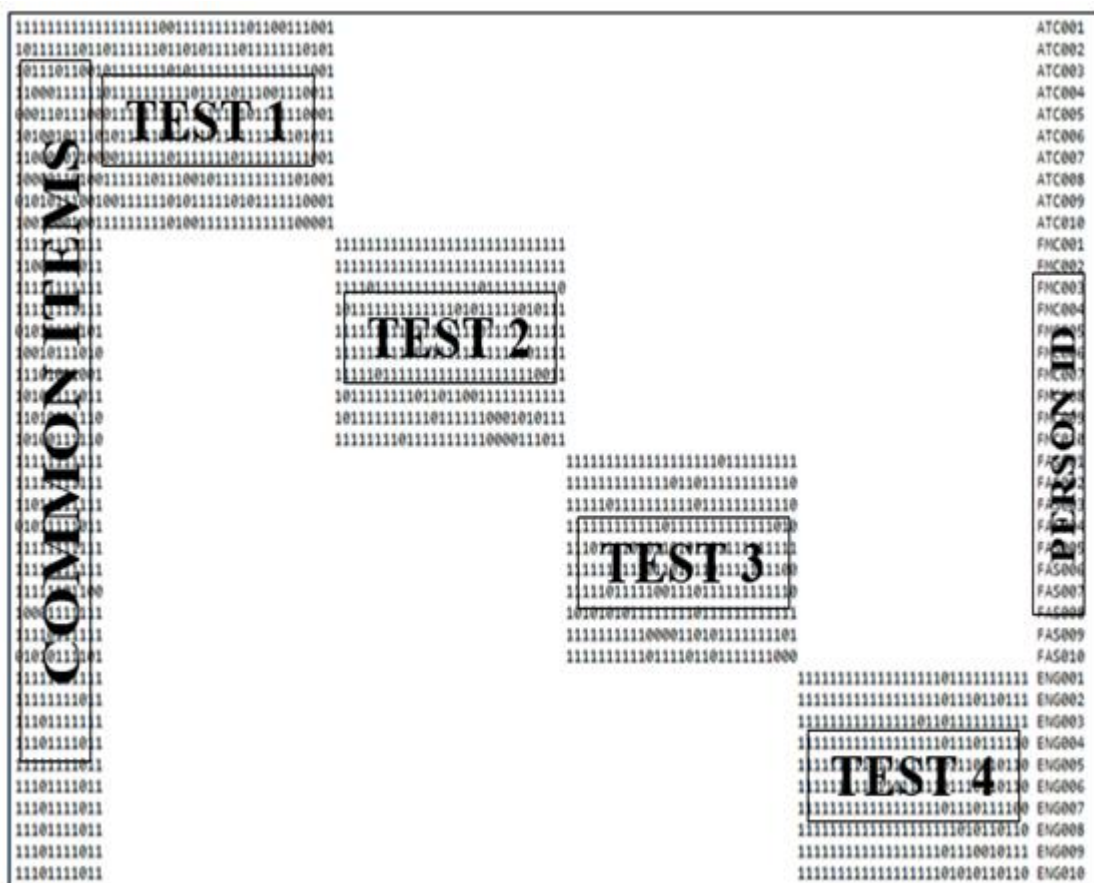


Figure 3.4 Sample of Common Item Equating Data Matrix Configuration

The structure of the equating mechanism used to link the testlets is shown in Figure 3.4. The first eleven items were the same across all exam forms, as can be seen. However, they have been represented from item twenty to item thirty in the real exam paper. In the concurrent analysis, the four testlets were combined into a single test to allow for direct comparison.

The common item design was chosen, based on Linacre (2020a), as it is the most efficient and straightforward approach to equating tests that share items in common. The four testlets were then evaluated using concurrent analysis, which simultaneously involved estimating item parameters for each item. The findings of the concurrent analysis are given below in detail under the following four subsections.

i. Reliability and separation of items and persons

According to the findings of the concurrent analysis, the items produced consistent results in the assessment of student performance. As shown in Table 3.18, the person reliability index is 0.84, whereas the item reliability value is 0.83, which meets the measurement requirements accepted by Rasch. The item separation index of 2.24 indicates “how well a sample of people is able to separate those items used in the test” (Wright & Stone, 1999, p. 151), which means that the items spanned between two and three distinct levels.

Mean error values for both item and person measurement for the concurrent analysis were a little high at 0.55 and 0.45, respectively. These big values could be attributed to the limited number of people and items in the testing, as only 30 or 34 people answered all items in a test, except for common items which were answered by all 124 persons.

Table 3.18 Item and Person Reliability

	Total score	Total Count	Measure	REAL S.E	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	26.7	40.0	1.23	0.45	0.99	.0	0.98	0.0
P.SD	6.5	0.0	1.20	0.16	0.23	1.2	0.53	1.1
REAL RMSE	0.48	True SD	1.09	Separation	2.27	Person Reliability		0.84
Mean	26.1	39.1	-.09	0.55	0.99	.1	0.99	.1
P.SD	16.3	26.2	1.50	0.27	0.18	.9	0.65	1.0
REAL RMSE	0.61	True SD	1.37	Separation	2.24	Item Reliability		0.83

ii. Unidimensionality

Table 3.19. shows that the raw variance explained by measures is 32.2%. The variance unexplained in the 1st contrast is only 2.9%, suggesting that there was no definable secondary dimension in the data. Additionally, both data and modeled expectation measures were almost comparable (32.2% and 32.4 %), which also bolsters the idea of unidimensionality.

Table 3.19 Dimensionality Map of Concurrent Analysis of All Four Tests

	Eigenvalue	Observed	Expected	
Total raw variance in observations	183.0026	100.0%	100.0%	
Raw variance explained by measures	59.0026	32.2%	32.4%	
Raw variance explained by persons	25.2448	13.8%	13.9%	
Raw Variance explained by items	33.7578	18.4%	18.5%	
Raw unexplained variance (total)	124.0000	67.8%	100.0%	67.6%
Unexplained variance in 1st contrast	5.2284	2.9%	4.2%	
Unexplained variance in 2nd contrast	5.0874	2.8%	4.1%	
Unexplained variance in 3rd contrast	4.3359	2.4%	3.5%	

iii. Misfitting Items

The point measure correlation (PTMEA CORR.) and other important data for all 127 original items of the concurrent analysis are shown in Table 3.20. There are no negative point measure correlation coefficients, indicating that the items act in the same direction as the measured construct. It can be noted that a number of the items had $<.30$ PTMEA correlation. Additionally, some items had both misfitting and overfitting infit and outfit MNSQ (See Appendix 3.7 for further details). These items were inspected further to understand the underlying problems in them. Moreover, it is also noted that some items had very large standard errors. One of the reasons for this issue may be the small sample size, as mentioned by Wang & Chen (2005, p. 386): “The critical ranges of the infit and outfit MNSQs should be adjusted according to sample sizes”. However, these items were reviewed further, leading to the removal of some and the modification of others.

Table 3.20 Fit Statistics for Concurrent Analysis

ENTRY	MEASURE	IN.MSQ	IN.ZSTD	OUT.MSQ	OUT.ZSTD	PTMA-E	RMSR	NAME
117	3.54	1.60	1.91	5.17	3.71	0.35	0.47	T4Q19
127	1.87	1.57	3.45	2.01	2.86	0.44	0.56	T4Q40
109	-0.97	1.36	0.86	0.88	0.21	0.39	0.31	T4Q11
124	2.90	1.36	1.83	3.89	4.07	0.39	0.49	T4Q37
97	0.83	1.30	1.94	1.26	0.93	0.39	0.51	T3Q39
98	2.67	1.29	1.25	1.64	1.69	0.43	0.45	T3Q40
73	0.01	1.29	1.29	1.44	0.98	0.33	0.45	T3Q4
20	0.98	1.27	1.81	1.44	2.10	0.35	0.52	T1Q9
12	1.31	1.27	1.53	1.36	1.45	0.33	0.50	T1Q1
9	1.64	1.24	2.64	1.32	2.27	0.46	0.49	CI28
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
118	0.07	0.69	-0.98	0.52	-0.82	0.45	0.29	T4Q31
112	-0.97	0.69	-0.62	0.37	-0.52	0.39	0.22	T4Q14
116	-0.97	0.51	-1.16	0.20	-0.92	0.39	0.19	T4Q18
120	0.07	0.50	-1.80	0.32	-1.46	0.45	0.25	T4Q33

iv. Item Difficulty and Person Ability for Combined tests

The Wright Person-Item Map, shown in Figure 3.5, depicts the distribution of all items and examinees of all four tests along the inquiry scale of the concurrent analysis. The difficulty of the items ranged from -4.00 to +3.50 logits, while examinee ability estimations ranged from -1.5 to +5.0. Around a total of 7 logits for both the items and persons indicates that both these are well-distributed along the inquiry scale, according to the map. 10 items were too easy for most examinees, while 7 items were too difficult for the majority of them. There are no visible gaps identified on the scale. However, the scale lacks persons at the easiest end, which means that many items seemed to be easier for many examinees. Although the map appears as if there were clustering of items at the easiest point and persons at the hardest points along the logit scale, suggesting redundancies, the reality is that the items were from different testlets, measuring different cognitive processes of reading.

TABLE 12.2 concurrent analysis NEM LABEL TO ITEM ZOU323MS.TXT May 21 2021 12:22
 INPUT: 124 PERSON 127 ITEM REPORTED: 124 PERSON 127 ITEM 2 CATS MINSTEPS 4.4.7

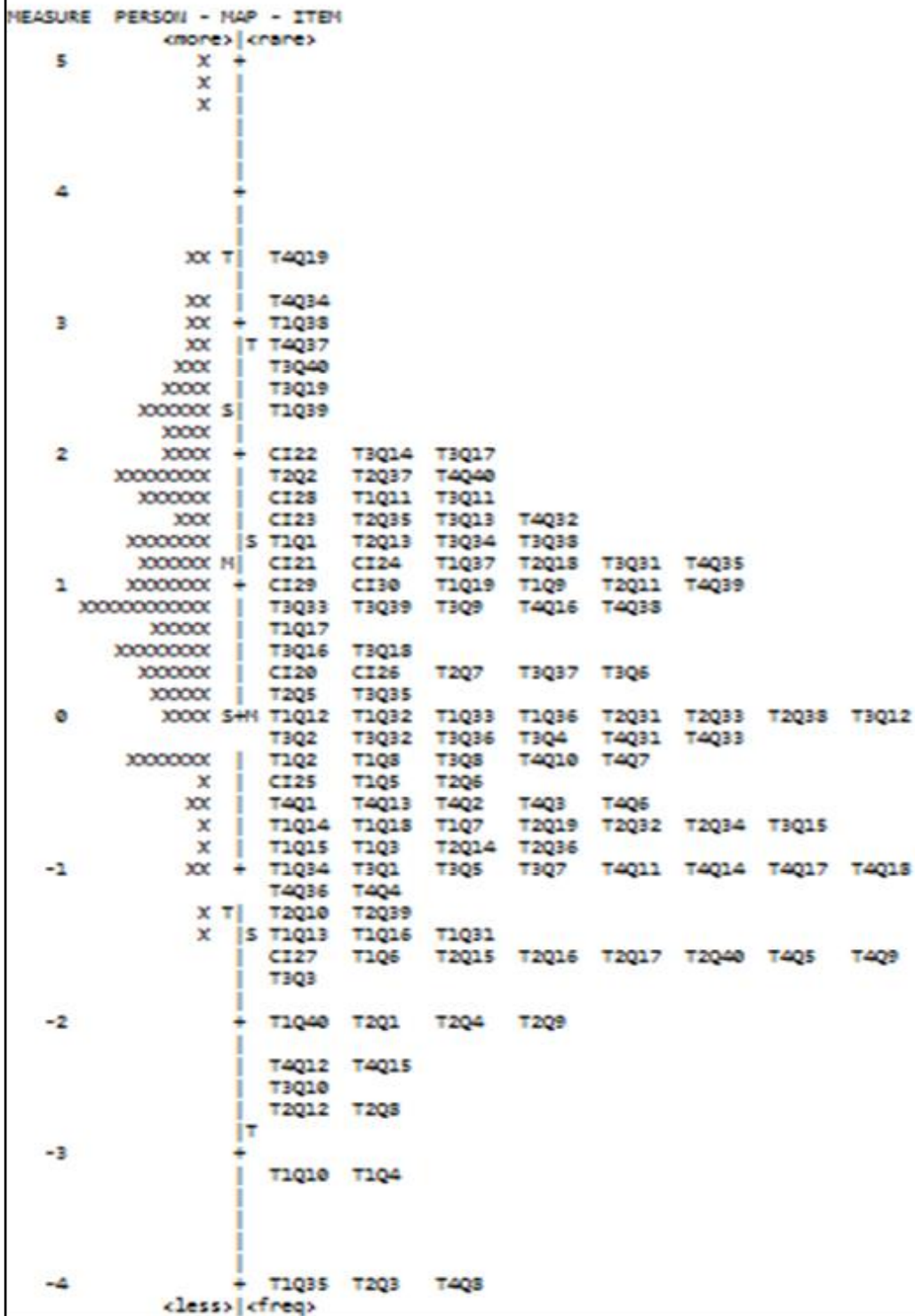


Figure 3.5 Reading Tests: Person- Item Wright Map

3.5.2.3 Modification of the Four Reading Tests Based on Piloting and Experts' Feedback

Based on the findings of the pilot study, and then on experts' feedback, certain changes were made to some of the items to enhance the instruments for the main study. The following decisions were drawn:

1. As suggested by the experts, eleven items were modified. These items were identified as the easiest items misfitting the scale. This was confirmed by students' scoring of these items in the pilot study.
2. No items were removed from the instrument, as it has been suggested by the experts that there are no problems with the item content or format so that the item would not be challenging, and the main data collection among a large number of population can produce better scores for the items.
3. All tests were expected to have a similar format in order to avoid confusing the students even if there were substitutions among the students.

3.5.3 Instrument for Final Study

The instrument for the final data collection was based on the findings of the pilot test and the feedback of the subject matter expert. The findings of the pilot study led to the inclusion or alteration of a number of items in the final version of the instrument. The final version of the instrument featured a total of 127 original items. The items in each testlet in the actual research are summarized in Table 3.19. Each testlet has four texts: one common text, and three others consisting of 40 MCQ items (as Ingebo (1997) advocated that a minimum of 40 items is sufficient for productive instrument development.)

Positive point measure correlation (PTMEA CORR.) exists for all of the selected items, indicating that they are all moving in the same direction. However, the Wright item-person map for the combined tests indicated that 10 items spanned below -2.00 logits. T1Q35, T2Q3, T4Q8, T1Q10, T1Q4, T2Q12, T2Q8, T3Q10, T4Q12, and T4Q15 were modified expecting to produce higher logit scales in the main data. Some of the wording of stems and distractors of these items were changed into harder

vocabulary, and higher grammatical functions were added concerning higher item difficulty levels.

As Hughes (1989) claims that effective choice of reading texts is based on experience, judgment, and a degree of common sense, the next step is based on the judgment and feedback of language experts. Even though the piloting results showed unacceptable statistics for a few items in the earlier section, these items were modified according to their suggestions. Moreover, they proposed to change one item (T3Q10) in Test 3. So, the difficulty level of word choice of the distractors of this item was increased. Table 3.21 illustrates a summary of the final instrument for this study.

Table 3.21 Summary of the Final Instrument

Text No	Text Title	Items	Total Items and score
Test 1 P1	Ice Cream	10	
Test 1 P2	Jet Lag	9	
Test 1 P3	Holidays	11	40
Test 1 P4	Technology	10	
Test 2 P1	Educational Programmes For Adults	10	
Test 2 P2	The Tradition of Coffee Drinking	9	
Test 2 P3	Holidays	11	40
Test 2 P4	London Olympics	10	
Test 3 P1	Playing Outdoors	10	
Test 3 P2	Supersonic Flight	9	
Test 4 P3	Holidays	11	40
Test 3 P4	Sir William Empson	10	
Test 4 P1	Having Friends	10	
Test 4 P2	The magic of the Cinema	9	
Test 4 P3	Holidays	11	40
Test 4 P4	The Shard	10	

Appendix C clearly depicts the IOC indices of all 160 items in detail. The highlighted numbers in each item indicate that they are the objectives (according to the IOC), or cognitive processing in reading (as per the theory). For instance, item 1 of Test1 belongs to the cognitive process; syntactic parsing (SP). Table 3.8 (refer to section 3.5.1.2.2.1.1) depicting the summary of cognitive processes of reading after the IOC analysis, is noted here. Word recognition (WR) is the easiest cognitive process, whereas creating intertextual representation (CITR) is the highest process according to Khalifa and Weir (2009). Thus, besides Test 1, all three tests show a similar number of item difficulty, indicating a similar number of items for LOT and HOT skills. The IOC indices further indicate that Test 4 is the hardest test consisting of 20 items for both LOT and HOT skills. According to Table 3.22, the mean total score of Test 4 is the highest, reporting at 22.4 (out of 40 total count/marks) indicating that it is the easiest test. However, according to Rasch MM, Test 3 is the hardest test scoring the mean measure of 0.00 logit scale, whereas the other three tests share similar logit scales.

Table 3.22 Summary of Mean Score of Individual Tests

Tests	Total score	Measure	Model S.E	Infit MNSQ	Outfit MNSQ
Test 1	17.5	-0.11	.51	.99	.97
Test 2	20.7	-0.10	.56	.97	1.09
Test 3	22.2	0.00	.46	1.00	.94
Test 4	22.4	-0.10	.62	.96	.96

Tables 3.2, 3.3, 3.4, and 3.5 in the Section 3.5.1. need to be considered when discussing how the test was assessed. Altogether there are four texts in each test ranging from CEFR B2 to CEFR C1, and multi-level (as in the IELCA test). Each test comprises 40 items that belong to the SR item response method. It is expected to apply the reading skills that the students have acquired in their classes with different contexts, and text types when a test is designed (Alderson, 2000).

3.6 DATA COLLECTION

Permission from the Post Graduate Office of the Faculty of Education, International Islamic University Malaysia, was granted to collect the data. SEUSL, permitted and facilitated the data collection procedure. The help and collaboration of the administrative staff, academic members, assistant instructors, and students of SEUSL, especially the staff members of the Department of English Language Teaching (DELT) are highly appreciated. Since the physical teaching and learning activities in Sri Lankan universities were restricted because of the Covid-19 pandemic, the pilot data were collected online in January 2021 for Tests 1, 2, 3, and 4. Since the pandemic continues to prevail up-to-date (2021.11.23) in Sri Lanka, from May to July 2021, the main data collection took place online. Two-hour lecture sessions were taken with the permission of the Head of the DELT under the supervision of the lecturers in charge. Zoom meetings were scheduled by the respective lecturers, and they invited the researcher to carry out the briefing session of the reading test and administer it.

In the briefing session, they were instructed as to the purpose and the low-stake nature of the test, the description of test tasks, the relevance of the test to the participants, and the significance of the study. To further ensure that the participants were intrinsically motivated to respond to the tests seriously, they were informed about the automated evaluation of their reading performance following the CEFR benchmarking levels. Since they are adult learners and follow EMI instruction, the importance of self-learning and the evaluation process was impressed upon them. The maximum time limit given to the students was 60 minutes. However, those who completed earlier were requested to leave the meeting of their own accord.

Once the link for the tests was given in the chat box of the Zoom meeting, students were able to access the test papers. All the students were carefully invigilated by the lecturer in charge and the researcher. Since the researcher is a bona fide senior lecturer in English and is well-known to many of these sample students, the majority of them took this exam seriously and performed well. Further, a permission letter (Appendix 3.78) issued by the Post Graduate Office of the Faculty of Education, where this doctoral study is being pursued, was used to

secure approval and permission from the respective faculty deans, course coordinators, and instructors of the classes.

Google forms were employed to design the tests, and the scoring was automatically exported to excel sheets once the students submitted the completed tests. Students from four faculties were chosen for this study, and they sat for four different tests. The duration of the test is 60 minutes. Overall, the majority of the participants needed about 50 minutes to complete the test, although a few of them needed only 30 minutes to respond to all 40 items, and a few of them took more time, and it was reported by the respective instructors that those students' data connection was poor to upload their forms.

3.7 ETHICAL CONSIDERATIONS

The primary data collected for this study were protected by the use of masked student identifiers. For the security and protection of student data and information, the student's identification was substituted by the corresponding number given to each student respective to their faculty. The protection of the human subjects as well as the use of the data was also subject to the standards and permission of the university rules and regulations. The researcher remained cautious and cognizant of any impact this study may have had on the participants, the university, and higher education entities.

3.8 ANALYSIS OF DATA

After the data had been collected, *WINSTEPS* version 4.4.7 (Linacre, 2020c) was utilized to analyse them. Items were named as coding. For example, T1Q1SP is the name given to item 1 of test 1. Here T1 represents Test 1, Q1 represents Question 1, and SP represents the cognitive process of reading; the item measures as per the content judgment. Persons were labelled according to their faculty, representing FAC, FMC, FAS, and FE. The test scores were recorded in Microsoft Excel sheets.

3.8.1 Rasch Measurement Model Analysis for the Final Instrument

RUMM2020, ConQuest, and Winsteps are the commonly used software to analyse dichotomous data of the Rasch Measurement Model; however, several other software are commercially available. Nevertheless, in this study, the researcher opted to use Winsteps 4.4.7 created by Linacre (2020) for the following reasons:

- a. It is very user-friendly.
- b. It is statistically sound and has fewer difficulties with estimation bias (Linacre, 2020a). The evaluation methodology utilized in Quest, on the other hand, is susceptible to estimation bias (DeMars, 2002), although it is versatile in terms of sample size and is not known to demonstrate estimation bias with small samples.
- c. It does not emphasize misfit.
- d. It has a built-in bug-fixer.
- e. In RUMM, missing data need to be frequently accounted for, while it can be imputed or removed in Winsteps.

Winsteps has included several models, named the Georg Rasch dichotomous, Andrich "rating scale", Masters "partial credit", Bradley-Terry "paired comparison", Glas "success model", Linacre "failure model", and most combinations of these models (Linacre, 2020c).

Further, it has three different estimation approaches, JMLE (Joint Maximum Likelihood Estimation), PROX (Normal Approximation Algorithm), and XMLE (Exclusory Maximum Likelihood Estimation). Each estimation has its own benefits and constraints. Therefore, in Winsteps, each method of estimation is subsequently fine-tuned by another, and the constraints of each method are compensated or complemented by another. For instance, XMLE is supposed to compensate for the statistical uncertainty and estimation bias of JMLE.

In the context of small sample sizes, test linking using Winsteps-based Rasch is better than the classical equating pertaining to Babcock and Hodge (2020, p. 1).

WINSTEPS-based Rasch methods that used multiple exam forms' data worked better than Bayesian Markov Chain Monte Carlo methods, as the prior distribution used to estimate the item difficulty parameters biased predicted scores when there were difficulty differences between exam forms.

Therefore, the present study applies Winsteps for analysing the dichotomous data to investigate four aspects of validity evidence: validity of test items, construct validity, consistency with the purpose of measurement, and validity of examinee responses. As well as to present the students' reading performance Wright item-person maps were produced to display the locations of cognitive processing in reading on the logit scale. Consequently, the data analysis and analysis software used in this study is certainly more accurate.

3.8.2 SPSS Analysis

To compare the performances of students of the four faculties on the four reading tests, statistical analysis using SPSS, version 26, was carried out to calculate means, standard deviation, medians, and mean errors. To easily identify the location of the student's performance measures on the CEFR levels among subgroups, Boxplots, as prescribed by Pallant (2020), were also used. (Boxplots are useful for comparing the distribution of scores on different variables.)

Demographic variables investigated in the study were gender (male/ female); faculty they represented (FAC- Faculty of Arts and Culture/ FMC- Management and Commerce/ FAS- Applied Sciences/ FE- Engineering); and year of study (first year/ second year/ third year/ fourth year). According to the research questions, all analyses of students' performance on four English reading tests were presented in tables and figures.

3.9 SUMMARY OF THE CHAPTER

This chapter explains the research methodology, including research design, research procedure, instrument development, construct definitions, the characteristics of the test instruments, sampling procedure, characteristics of respondents, and the processes of the development and validation of the instrument. Accordingly, the results of the pilot study were also discussed followed by data collection and data analysis. The following chapter contains information on sophisticated data analysis processes that were not discussed in this chapter for better clarity and coherence in interpreting the findings of the data analysis.

CHAPTER FOUR

RESULTS OF THE STUDY

4.1 INTRODUCTION

This chapter presents the findings from the data analyses in light of the research questions (RQ) given in Chapter One (section 1.3.2). RQ1 is concerned with the psychometric properties of the CEFR-aligned reading tests. It focuses on five factors: the validity of the test items, the precision and reliability of measurement (or the test's ability to reproduce consistent results in measurement), the construct validity of the test items, the validity of the common items that link the tests, and the validity of the students' responses. The RQ2 highlights four faculty students' reading performance in the CEFR-aligned tests. Finally, cognitive processes of reading are the emphasis of RQ3.

Research Questions:

1. What are the psychometric properties of the CEFR- aligned reading tests?
2. What is the performance of the students in the CEFR- aligned reading tests?
3. What is the performance level of SEUSL undergraduates who follow the EMI system, in the cognitive processes of English reading?
 - a. In which cognitive processes of reading do the SEUSL students indicate higher achievement?
 - b. In which cognitive processes of reading do the SEUSL students indicate lower achievement?

4.2 PRELIMINARY ANALYSIS OF THE MAIN DATA

A preliminary analysis was performed to confirm the item quality, that all items are accurate to be used in the final study. The estimation of the item and person quality was derived using the SPSS and the WINSTEPS software packages.

4.2.1 Screening and Cleaning of Data

The first step before conducting the main analysis is to ensure that no error in the dataset would influence the results, and hence the conclusion made from the findings. Therefore, a preliminary analysis was conducted to screen and clean the dataset.

Checking the missing values is crucial in analysis before processing the main data. As the responses of each test of the study were converted to a separate Excel sheet using Google form, it was convenient to key in the data. Except for the records for demographic information, the data for each item were keyed in using formulas to get the right answer like “1” and the wrong answer as “0”, as they all are used as dichotomous data in the main study. If there is any missing value that has been identified as a wrong answer by the given formula, then there is no missing value found in the data.

Based on the findings of the pilot study, eleven items were modified for the final study, and no items were found to be misfitting in the main data. However, five persons were deleted from the dataset. This was following the fit statistics that found these persons to be misfitting according to the results found in the Boxplot analysis of SPSS and Person fit statistics of *WINSTEPS*. As can be seen from Figure 4.1, the person (FAC174) 174 belonging to Test 1, was removed from the study as per the results found in the SPSS boxplot analysis. Similarly, persons labelled as FAC 210, FAS 135, FAS 167, and FAS 175 were removed as they were reported to be the most misfitting persons from the results of the *WINSTEPS* output table for person fit.

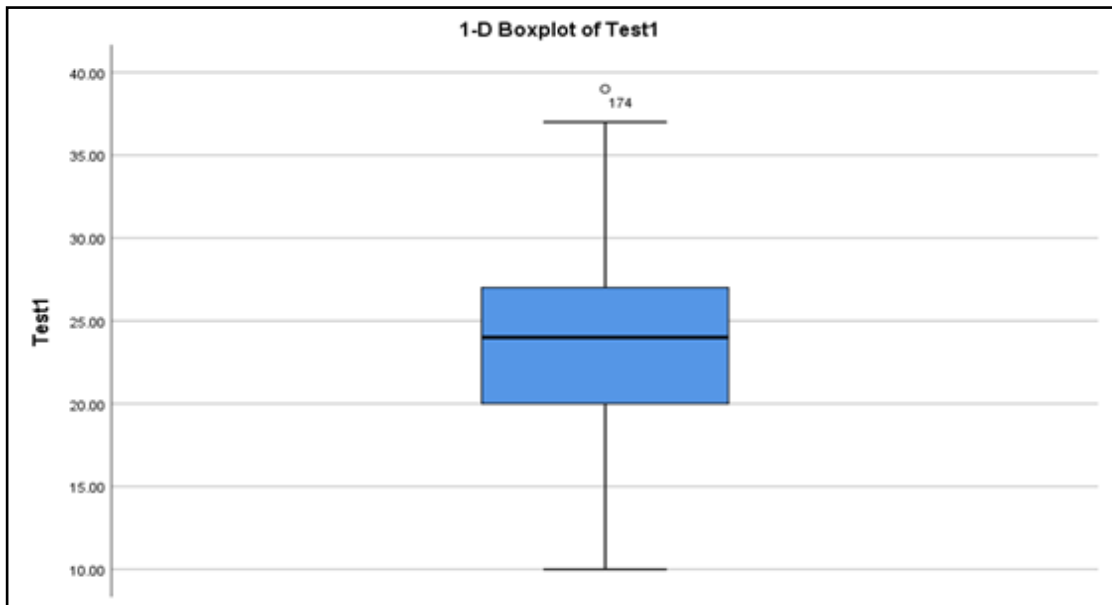


Figure 4.1 Distribution of Scores among Persons in Test 1

4.2.2 Validity of Reading Tests

The quality of a test depends on each item (Freeman, 1962; Sharma, 2000). Thus, first of all, the validity of the test items should be prioritised. Examining the validity of test items involve three indicators: item fit statistics, item polarity, and unidimensionality (through principal component analysis of residuals). The second step is to evaluate the construct validity of the reading tests. This can be done in two ways. The first one is by examining the capability of the test items in delivering a continuum of increasing intensity. The second method is to look at the consistency of empirical scaling and expert judgment. These two kinds of data give the evidence needed to determine whether or not the test items accurately described the measured construct and sub-construct (Bond & Fox, 2015).

The capacity of the tests to produce results that were consistent with the measurement goals was evaluated by three criteria: reliability and separation indices for student ability and item difficulty, the precision and accuracy of person and item measures, and test targeting. To determine the location of students according to their performance in the measured construct, the validity of students' responses is ascertained.

4.2.2.1 Validity of Test Items

In determining the validity of test items, three factors are of concern: item fit, item polarity, and unidimensionality. Each of these results is presented in the subsequent sections.

4.2.2.1.1 Item Fit

Item fit statistics are used to assist in the detection of items that deviate from the Rasch model's expectations, i.e. to ensure that the items are contributing meaningfully to the measurement of the variable or construct (Bond & Fox, 2015). The infit and outfit mean-square statistics are the two of the most commonly used fit statistics. These statistics show “the size of the randomness, i.e., the amount of distortion of the measurement system” (Linacre, 2020a, p. 607).

Infit mean-square statistics is “an information-weighted statistic, which is more sensitive to unexpected behaviour affecting responses to items near the person's measure level” (Linacre, 2020, p. 365). On the other hand, outfit mean-square is “an unweighted statistic, more sensitive to unexpected behaviour by persons on items far from the person's measure level” (Linacre, p. 366), which is estimated based on the conventional sum of squared standardized residuals (Bond & Fox, 2015). 1.0 is the expected value for both infit and outfit mean-square statistics. Values less than 1.0 indicate relatively predictable observations, whereas values greater than 1.0 imply unpredictability or unmodeled noise.

Values for infit or outfit mean square outside the range of 0.7- 1.3 were used to imply that items are misfitting, as this is the accepted range for the cognitive test, including multiple-choice items (Bond & Fox, 2015; Linacre, 2020). Although Muller (2020, pp. 7–8) criticised that “the usual rule of thumb of 0.7- 1.3 is valid for n around 200”, considering the cognitive process of the test items, the above range is applied in the present test. Items that fall within the recommended range are regarded as productive or meaningful to the measurement; values below this range suggest that the items are overfitting, while values beyond this range indicate that the items are misfitting (Bond & Fox, 2015; Wright & Linacre, 1994).

Table 4.1 Item Fit Statistics – Misfit Order

ENTRY NO	MEASURE (logits)	MODL SE	IN.MSQ	OUT.MSQ	PTMA-E	NAME
117	2.36	0.19	1.36	1.76	0.37	T4Q19 B2 CTLS
127	0.23	0.17	1.37	1.60	0.40	T4Q40 C1 I
115	-1.53	0.25	1.07	1.53	0.25	T4Q17 B2 LA
40	-1.10	0.18	1.12	1.49	0.26	T1Q40 C2 CTLS
124	1.74	0.17	1.26	1.36	0.41	T1Q11 C1 EPM
85	0.10	0.14	1.09	1.31	0.35	T3Q16 C1 EPM
75	-0.30	0.15	1.07	1.29	0.32	T1Q2 B1 SP
93	0.36	0.13	1.14	1.26	0.37	T3Q34 C2 BMM
59	0.27	0.14	1.12	1.26	0.34	T2Q19 B2 EPM
45	-0.36	0.15	1.06	1.26	0.31	T2Q5 B2 WR
98	1.61	0.14	1.13	1.25	0.40	T3Q40 C2 BMM
66	1.54	0.15	1.14	1.25	0.36	T2Q37 C2 I
121	1.80	0.17	1.22	1.23	0.40	T4Q34 C1 BMM
97	1.18	0.13	1.16	1.21	0.39	T3Q39 C2 BMM
9	1.06	0.07	1.15	1.20	0.38	CI28 C1 CTLS
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
112	-0.54	0.19	0.84	0.76	0.34	T4Q14 B2 EPM
118	0.40	0.17	0.83	0.76	0.41	T4Q31 C1 BMM
60	0.54	0.14	0.79	0.74	0.35	T2Q40 C2 SP
120	0.68	0.17	0.78	0.74	0.42	T4Q39 C1 EPM
MEAN	173.7	284.1	0.00	0.16	0.99	0.98
P.SD	109.0	193.1	1.06	0.05	0.10	0.19

Infit and outfit mean-squares for individual items are shown in Table 4.1. The first few items and some items on the last part of the item fit statistics (see Appendix G.1.c for the complete statistics of item misfit order) are given in this table.

Table 4.2 Summary Table of Frequency of Item Fit within 0.7- 1.3 infit and outfit MNSQ Range

Mean- Square Value	Infit		Outfit	
	Frequency	Percentage	Frequency	Percentage
Below 0.7	0	0.0	2	1.6
0.7 – 1.3	125	98.4	119	93.7
Above 1.3	2	1.6	6	4.7
Total	127	100	127	100
Mean (logits)		.99		.98
S. D		.10		.19

Table 4.2 summarises the data from the item fit statistics (given in table 10.1 of WINSTEPS 4.4.7 analysis, as shown in Table 4.1), corresponding to the evaluation of the accepted infit and outfit MNSQ ranges, along with their item frequency and percentage. Except for two items all the other items were within the specified range (0.7 - 1.3), showing 98.4% in the infit mean-square. However, the outfit mean-square index reveals that two items were less than 0.7 (or 1.6% of the total), and six items had outfit mean-square values above 1.3, making 4.7% of the total items. Out of 127 items, only a small proportion of items were above or below the accepted range. The mean logits of the infit mean-square (.99 logits) and outfit mean-square (0.98 logits) were very close to the expected value of 1.00. In the meantime, the standard deviations of the infit mean square (.10 logits) and outfit mean square (.19 logits) was only slightly apart from the expected value (0.0), indicating little variation from the prediction of the Rasch MM. The standard deviation of the outfit mean square (.19 logits) was, however, a little higher than that of the infit mean square.

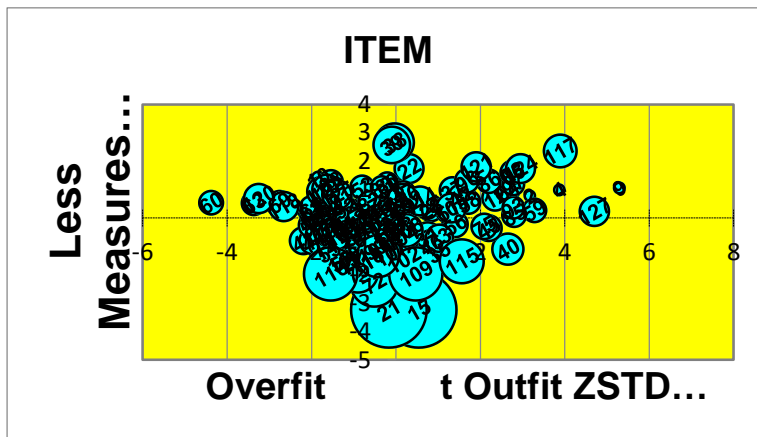


Figure 4.2 Bubble Chartz

Figure 4.2 explains the overfit and underfit items using the outfit ZSTD in the bubble chart. As per the chart, the distribution of the majority of items is close to 0.0 - 1.0 logit.

4.2.2.1.2 Item Polarity

Item polarity is the second indicator used to check the validity of the test items. It is measured by the point-measure correlation coefficient, which indicates how closely test items function together to describe the intended construct. Linacre (2020a) states that items or examinees with negative or zero values are working in the wrong direction. As a result, relatively high positive results are desired in the examination of item polarity.

The point measure correlation (PTMEA CORR.) for the 127 items is completely shown in the eighth column in Appendix G.1.c. In Table 4.3. item polarity statistics for the items in the two ends of the statistics are given.

Table 4.3 Item Polarity Statistics: Measure Order (Reading Test)

ENTRY NO	MEASURE (logits)	MODL SE	IN.MSQ	OUT.MSQ	PTMA-E	NAME
38	2.66	0.22	0.99	0.98	0.27	T1Q38 C2 EPM
39	2.57	0.21	1.03	0.97	0.27	T1Q39 C2 EPM
117	2.36	0.19	1.36	1.76	0.37	T4Q19 B2 CTLS
121	1.80	0.17	1.22	1.23	0.40	T4Q34 C1 BMM
124	1.74	0.17	1.26	1.36	0.41	T1Q11 C1 EPM
22	1.74	0.17	0.98	1.03	0.32	T4Q37 C1 BMM
98	1.61	0.14	1.13	1.25	0.40	T3Q40 C2 BMM
66	1.54	0.15	1.14	1.25	0.36	T2Q37 C2 I
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
79	-1.92	0.23	0.93	0.72	0.20	T3Q10 B1 SP
109	-1.97	0.3	1.01	1.14	0.21	T4Q11 B2 EPM
116	-1.97	0.3	0.89	0.48	0.21	T4Q18 B2 BMM
72	-2.29	0.27	0.94	0.79	0.17	T3Q3 B1 WR
15	-3.24	0.42	1.04	1.20	0.12	T1Q4 B1 SP
21	-3.24	0.42	0.99	0.84	0.12	T1Q10 B1 SP
MEAN	173.7	284.1	0.00	0.16	0.99	0.98
P.SD	109.0	193.1	1.06	0.05	0.10	0.19

The point measure correlation coefficients for all items were positive. No item has a negative point measure correlation coefficient according to the data. This indicates that all items are defined in the measured construct in the same way. Unexpected responses (as suggested by the Outfit MNSQ) had contributed to the low correlation values of less than 0.30. However, 34 items out of 127 were less than 0.30 (between 0.12 to 0.29). The poor correlation coefficients indicate that these items were ineffective in distinguishing between people of high and low ability. Despite this, the test items were all pointing in the same way when it came to measuring the construct.

4.2.2.1.3 Unidimensionality of the Items

The third indicator to validate the items is unidimensionality. According to the Rasch MM, items should function together to evaluate a single construct. This is because Rasch does not attempt to fit the model to the data obtained; instead, it focuses on whether the data fit the model constructively (Wright & Stone, 1979). The threat of secondary or sub-dimensions in determining unidimensionality is a threat to the main construct. Hence, data fit is determined by verifying whether the items on a specific test have successfully established one or a single construct. To check for unidimensionality, the principal component analysis (PCA) of residuals is used.

Table 4.4 PCA of Standardized Residuals of all Items

	Eigenvalue	Observed	Expected
Total raw variance in observations	166.84	100.0%	100.0%
Raw variance explained by measures	39.84	23.9%	24.0%
Raw variance explained by persons	18.37	11.0%	11.1%
Raw Variance explained by items	21.47	12.9%	12.9%
Raw unexplained variance (total)	127.00	76.1%	100.0% 76.0%
Unexplained variance in 1st contrast	3.77	2.3%	3.0%
Unexplained variance in 2nd contrast	2.68	1.6%	2.1%
Unexplained variance in 3rd contrast	2.25	1.3%	1.8%

Table 4.4 illustrates the unidimensionality of the items measured, by showing the results for PCA analysis. The gap between the variance explained by measures (23.9%) and the modelled expectations (24.0%) is very minor (0.1 %). Because all of the components in the first and second contrasts were less than 10%, therefore, there is no secondary dimension. And the greatest factor recovered from the residuals for the 1st contrast is 3.77 Eigenvalue, which is comparable to around 4 items in strength. Compared to the total of 127 items, the average 4 items in the first contrast are small in amount. A minimum of five items should be there to heavily load on a dimension to be considered as a separate factor or a construct (Linacre, 2020a), and since the percentages in the 1st, 2nd, and 3rd contrasts, were less than 10%, namely at 2.3%,

1.6%, and 1.3% respectively, it can be confirmed that there is no secondary dimension.

The representation of the variance components on the log scales had been shown in Figure 4.3 to confirm the unidimensionality.

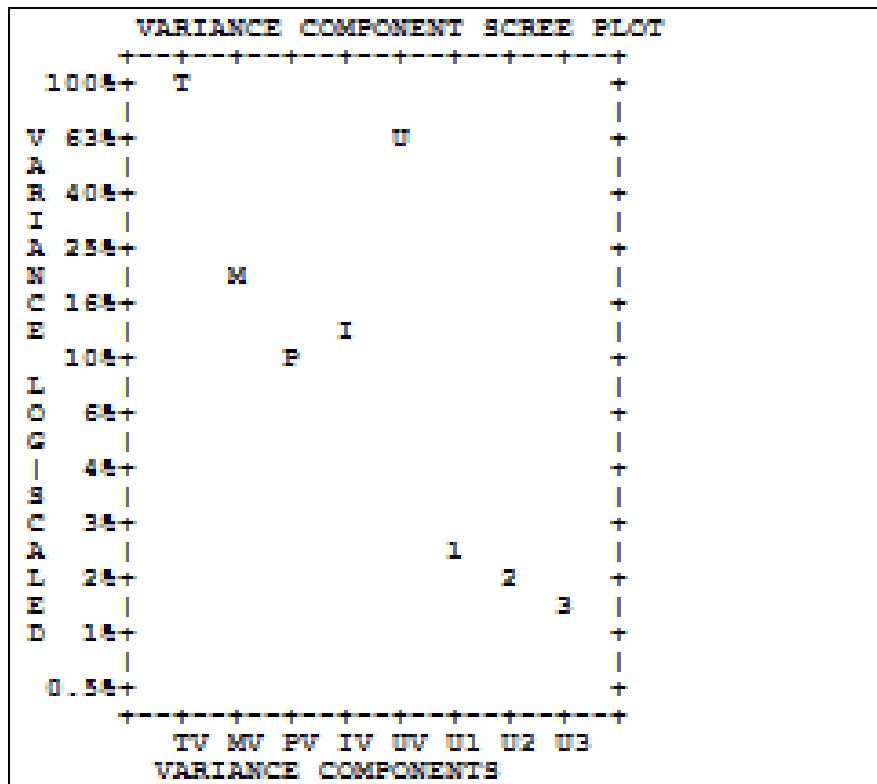


Figure 4.3 Standardized Residual Variance Scree Plot

4.2.2.2 Construct Validity

To find out whether the items are measuring one main construct, i.e., if the reading ability can be checked through using two main ways, the first thing is to look at the capacity of the test items to deliver a continuum of increasing intensity. Next is to examine the consistency of empirical scaling and expert judgment.

4.2.2.2.1 Continuum of Increasing Intensity

Wright and Stone (1979) noted that a variable or construct is well-defined when the items are well-separated. The Wright item-person map is a useful map to locate the items along the logit scale. If there is no significant gap between the item location, it provides evidence of an increasing intensity continuum. Except for slight gaps at the upper and lower ends of the scale between the locations of Items T4Q19 and T4Q34 at the top end, and Items T3Q3 and T1Q10 at the lower end, as shown in Figure 4.4, there were no notable visible gaps between item distributions. Items T1Q4 and T1Q10 were identified as the easiest items located at -3.24 logits, whereas item T1Q38 was the most difficult item on the map, at 2.66 logits. The span between the highest and the lowest difficult items is $(-3.24 - 2.66) 5.90$ logits.

In the concurrent analysis, too, there cannot be seen any serious redundancy, as can be visible in Figure 4.4. This may be because there are many items (127), placed on the map belonging to four separate tests, collaborated by eleven common items. At the same time, most of the repeated items in the same location of the map, are different in terms of the measured cognitive processes. For instance, at the bottom line of the map (Figure 4.4), T4Q11 and T4Q18 are located at the same position; however, T4Q11 evaluates the cognitive process of EPM, whereas the next item examines the process of BMM. Therefore, such redundancy does not pose a danger to construct validity.

It is quite noticeable that there are a few items (around 7 items belonging to three separate tests) located at the lower end of the map; however, there cannot be seen any students at the lowest part of the map. This means that these items are too easy for the students in the tests. Although these items indicate that all the students achieved mastery to certain levels in these tests, these items are an underestimation of the examinees' performance. Therefore, it is advisable to have items that can better discriminate the students according to different ability levels.

Checking the item person map for the separate four tests indicated that there was not much redundancy of items. For instance, Test 1 indicated that there is no significant redundancy among items. The stacks of items in Test 1 shown in Figure 4.5, illustrate that there is no redundancy in Test 1.

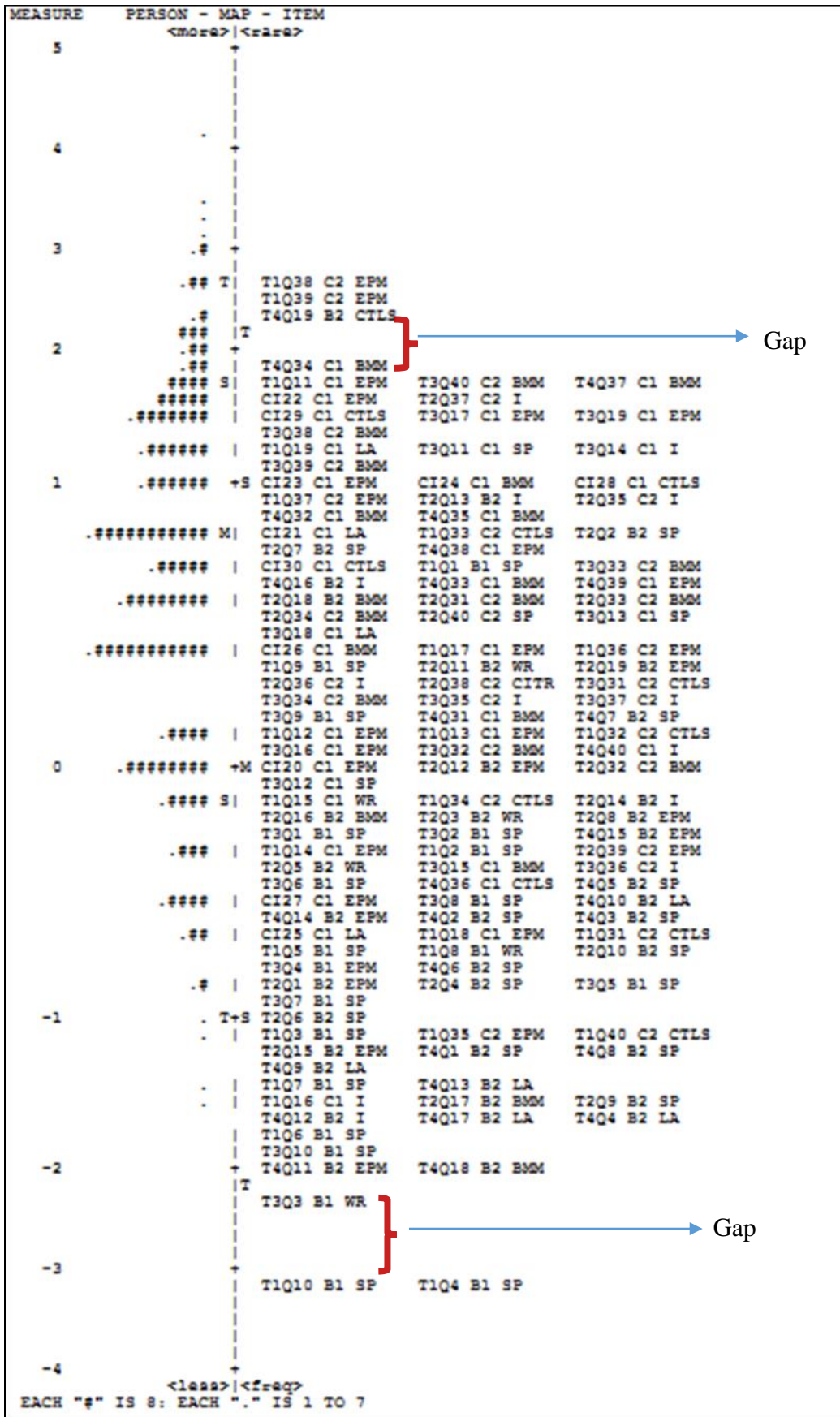


Figure 4.4 Item-ability - Wright Map for all four tests

Table 4.5 Summary of Cognitive Processing of Reading in Each Test Based on Expert Judgment

	WR	LA	SP	EPM	I	BMM	CTLS	CITR	Total	LOT	HOT
Test 1	2	3	9	15	1	2	8	0	40	29	11
Test 2	3	2	7	10	5	9	3	1	40	22	18
Test 3	1	3	11	8	4	9	4	0	40	23	17
Test 4	0	7	7	9	3	9	5	0	40	23	17
Total	6	15	34	42	13	29	20	1	160	97	63
%										61	39
Total¹	6	9	34	30	13	23	11	1	127	79	48
%										62	38

According to Khalifa and Weir's (2009) socio-cognitive validation framework, the cognitive processes are ordered hierarchically from the lowest (here WR- word recognition) to the highest (CITR- creating inter-textual representation) (Bax & Chan, 2016; Bax, 2013). Good tests must have a reliable proportion of easy, average, and the most difficult items (Brown & Abeywickrama, 2010). The ratio of items as per the expert judgment proves the tests to be adequate. A total of 160 items (after item calibration of common items in each test) share 61% of low order thinking processes, whereas 39% of items test high order processes. These percentages are almost similar, even if the items are not linked using the common items; as can be seen in Table 4.5, 127 items share 62% and 38%, respectively, of LOT and HOT processes.

Figure 4.6 depicts the distribution of difficulty of the items based on the eight cognitive processing in reading. It reveals that out of the eight processing, items for determining WR (word recognition) and SP (Syntactic Parsing) were the easiest processes, having a mean of -0.55 logits with an SD of 0.90 for WR and 1.03 for SP, correspondingly. Items for BMM (building mental model) were the most difficult, with a mean of 0.50 logits and SD at 0.91 logits. Other categories had the following item distribution and difficulty estimates: LA (lexical access) (mean = -.48 logits, SD = 1.02 logits); EPM (establishing prepositional meaning) (mean = 0.24 logits, SD =

⁽¹⁾ Total for individual items excluding common items.)

1.07 logits); I (inferencing) (mean = 0.23 logits, SD = 0.93 logits); CTLS (creating text level structure (mean =0.42 logits, SD =0.98 logits); and CITR (creating inter textual representation) with 0.33 logits.

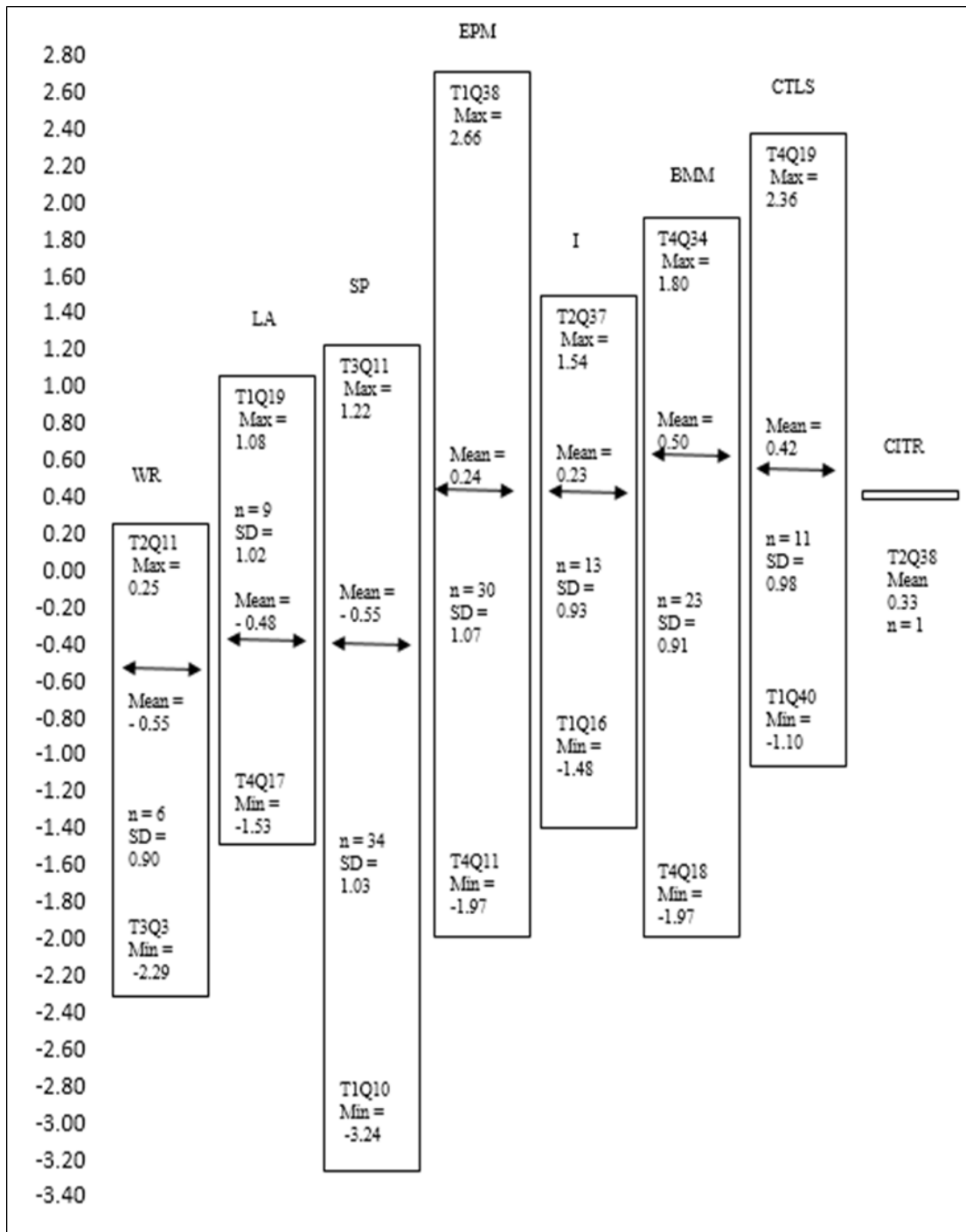


Figure 4.6 Empirical Scaling of Test Items Based on Cognitive Processes of Reading

Another finding was that, apart from the cognitive process of CITER, which had been identified with only one item in all four tests (T2Q38 with 0.33 logits), all other cognitive processes had significant distribution. All the tests in the present study target the CEFR B2 level to C2 level. However, the maximum difficulty level of the tests was expected to be in the C1 level (refer to section 4.3.1 for further clarification). As CITER is tested at the higher level of the CEFR-aligned tests (Khalifa & Weir, 2009), based on the content validation of the expert judgment, this cognitive process was not strictly intended to be measured in the present tests, except for one item out of 160.

The item distribution also indicates the divergence between the content validation of expert judgement, the hierarchical order of Khalifa and Weir's cognitive processes, and Rasch evaluations. However, there ought to be no concurrence between the order of cognitive processes based on expert judgment and the item difficulty level, as per the results of the Rasch Measurement Model, because the RMM assumes that the items scoring lower logits suggest that they are easy to get through, whereas the items having the higher logits suggest that they are difficult to answer. For instance, an item in SP scoring -3.24 logits, may be an easy item to most students, while, at the same time, another item in SP scoring 1.22 logits, may be tough for all students. These findings were similar to some of the previous research (Badrasawi, 2012; Hudson, 2007; Jusoh, 2018; Rosenshine, 2017).

4.2.3 The Precision and Reliability of Measurement

Consistency of the measurement can be read through the reliability and separation indices of both the items and the person.

4.2.3.1 Reliability and Separation

Although reliability does not give details on the quality of the data, it refers to the "reproducibility of relative measure location" (Linacre, 2020a, p. 671). So "high reliability" (of people or items) implies that people (or items) attributed with high measures are more likely to have higher measures than people (or items) assessed with

low measures. Large sample size and/or minimal measurement error are required for high reliability.

So, in order to have high person (test) reliability, a big ability range and or a large number of items in the instrument are needed. Similarly, to have high item reliability a test with a wide item difficulty range and or a large sample size is needed. The person sample size is usually too small to build a reproducible item difficulty hierarchy, resulting in low item reliability.

Table 4.6 Reliability of 127 Measured Items

	Total score	Total Count	Measure	Model S.E	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	173.7	284.1	0.00	0.16	0.99	-0.06	0.98	-0.08
SEM	9.7	17.2	0.09	0.00	0.01	0.16	0.02	0.15
P.SD	109.0	193.1	1.06	0.05	0.10	1.76	0.19	1.71
S.SD	109.4	193.9	1.06	0.05	0.10	1.77	0.20	1.72
MAX	693.0	902.0	2.66	0.42	1.37	5.68	1.76	5.29
MIN	27.0	179.0	-3.24	0.07	0.78	-5.36	0.48	-4.38
REAL RMSE	0.17	True SD	1.05	Separation	6.02	Item Reliability		0.97
MODEL RMSE	0.17	True SD	1.05	Separation	6.11	Item Reliability		0.97
S.E. of Item Mean = 0.09								

Table 4.6 explains the reliability statistics of the 127 measured items. The reliability of item difficulty measurement was high, reporting at 0.97, as seen in Table 4.6 (see Appendix.G.1.a). This indicates that the item difficulty ordering is highly reproducible with additional students of a similar nature and that the items are well-separated in terms of difficulty. The item separation index was 6.02, indicating that the items could be split into at least seven degrees of difficulty.

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD
MEAN	24.5	40.0	.77	.38	1.00	.02	.99	.03
SEM	.2	.0	.03	.00	.00	.03	.01	.03
P.SD	6.3	.0	.91	.08	.15	1.02	.29	1.01
S.SD	6.3	.0	.91	.08	.15	1.02	.29	1.01
MAX.	39.0	40.0	4.23	1.03	1.56	3.54	3.00	3.89
MIN.	7.0	40.0	-1.52	.34	.62	-3.36	.18	-3.01
REAL RMSE	.40	TRUE SD	.82	SEPARATION	2.05	PERSON RELIABILITY	.81	
MODEL RMSE	.39	TRUE SD	.83	SEPARATION	2.11	PERSON RELIABILITY	.82	
S.E. OF PERSON MEAN = .03								
PERSON RAW SCORE-TO-MEASURE CORRELATION = .98								
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .37 SEM = 5.02								

Figure 4.7 Winstep Output Table for Reliability of 902 Measured Persons

Figure 4.7 illustrates the reliability statistics of the 902 measured persons. The reliability of the student ability measure was good, at 0.81 as illustrated in Figure 4.7. This implies that the ordering of students can most likely be duplicated with other items of similar difficulty. The student separation index was 2.05, showing that the reading test was able to separate students into at least three skill categories, although 2.05 is closer to 2, in classifying the person groups, which have to be allocated into three full groups. This is a good indication of discriminating students according to their ability levels, because Linacre (2020, p: 671) stated that “Low person separation (< 2, person reliability < 0.8) with a relevant person sample implies that the instrument may not be sensitive enough to distinguish between high and low performers”.

4.2.3.2 Precision of Measures

The ability of the test to measure the intended construct was also examined for consistency of the measurement. The mean standard error for the students was 0.38. Although this is a little high, which was above the target range of 0.18 to 0.27 ($2/\sqrt{127}$) $<SEM< 3/\sqrt{127}$ ($0.18<0.38< 0.27$), the means student standard error of individual tests was within the acceptable range (see Section 4.2.5). As a result, the reading test measured exactly enough to achieve its goal. In addition, item measurements had a good level of precision ($SEM = 0.16$) for the sample size ($n=$

902). Despite the high standard error for student measurements (0.38), the value was acceptable, due to the utilisation of a diverse group of students with varying skills. Therefore, the overall statistics imply that the tests are precise enough.

4.2.3.3 Test Targeting

Test targeting can reveal how commensurate the test is to the test-takers, which can be determined by how well and accurately the test is appropriate for the students. Item difficulty mean was 0.00 logit (SD= 1.06) on average, whereas student ability mean was 0.77 logits (SD= 0.91). Although Curtis and Boman (2007) suggest that the person mean within 0.50 is a good sign for measurement, as the person mean for the present tests is 0.77, it did not reach the mean of 1.0, because the measurement can be compromised when the person mean is more than 1.0 logits from the data source (Curtis & Boman, 2004). Therefore, the slight difference in the mean cannot have a high impact. However, there were around nine items below the lowest ability students' location (-1.52 logits is the location of the least able student) at the bottom line of the item person map. Although these items may have an impact on poor test targeting, the small amount of mean standard error of items (0.16) is an acceptable statistic.

4.2.4 Validity of Common Item Linking

The validity of the common items is evaluated in a way similar to the validation of all items for the concurrent analysis achieved. The person and item reliability and separation indices are the prominent factors to identify the consistency of the measurement of the common item.

Table 4.7 indicates the reliability indices of items and persons for the 11 common items. Although the item reliability and separation indices of the common items were having good values of 0.99 and 9.24, respectively, the reliability and separation values for the students were low, at 0.55 and 1.10, as shown in Table 4.7. Nevertheless, this may be due to the low number of items (only 11 items).

Table 4.7 Reliability Indices of Person and Item for the Common Item Calibration

	Total score	Total Count	Measure	Model S.E	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	5.8	11.0	.17	.71	1.00	.0	1.00	.1
P.SD	2.2	.0	1.17	.20	.20	.8	.34	.8
REAL RMSE	78	True SD	0.86	Separation	1.10	Person Reliability		0.55
Mean	478.7	902.0	.00	.08	1.00	.0	1.00	.1
P.SD	123.6	.0	0.73	.00	.06	1.9	.08	1.8
REAL RMSE	0.08	True SD	0.73	Separation	9.24	Item Reliability		0.99

Table 4.8 describes the item fit statistics of the common items. The item polarity of the common items indicates a good value as indicated by Table 4.8. The point measure correlation (PTMEA CORR.) for eleven items was excellent, possessing positive point measure correlation coefficients for all (all of them were above 0.30.) This indicates that the items measured the construct in the same way as directed.

Table 4.8 Item Fit Indices for Common Items

ENTRY	MEASURE	COUNT	SCORE	IN.MSQ	OUT.MSQ	PTMA	NAME
1	-0.55	902	578	0.99	0.98	0.42	CI20 EPM
2	0.20	902	446	1.03	1.06	0.41	CI21 LA
3	0.90	902	325	0.96	1.03	0.46	CI22 EPM
4	0.43	902	405	1.05	1.05	0.41	CI23 EPM
5	0.45	902	401	0.96	0.97	0.47	CI24 BMM
6	-1.30	902	693	0.99	0.95	0.39	CI25 LA
7	-0.33	902	540	0.93	0.89	0.49	CI26 BMM
8	-1.22	902	683	0.93	0.89	0.44	CI27 EPM
9	0.51	902	391	1.14	1.20	0.33	CI28 CTLS
10	0.88	902	327	1.01	1.00	0.44	CI29 CTLS
11	0.03	902	477	0.98	0.96	0.46	CI30 CTLS
Mean	.00	902	478.7	1.00	1.00		
SD	.73	.0	123.6	.06	.08		

Item fit analysis for these items proved to be good as well, as the infit and outfit MNSQ of all items were reported to be within the expected range of 0.70 and 1.30. No items had a misfit or overfit value, which entails that the items functioned in the same direction along the measurement scale.

The unidimensionality of the common items is determined by using PCA residuals. The raw variance explained by measures was 24.3%; however, the residuals for the unexplained variance in the 1st contrast is below the Eigenvalue 2 reporting at 1.51, which is less than 2 items in strength, which is a good indication of measurement (Linacre, 2020). Further, the gap between the observed variance explained by measures (24.3%) and the modelled expectations (24.2%) is very minor (0.1 %).

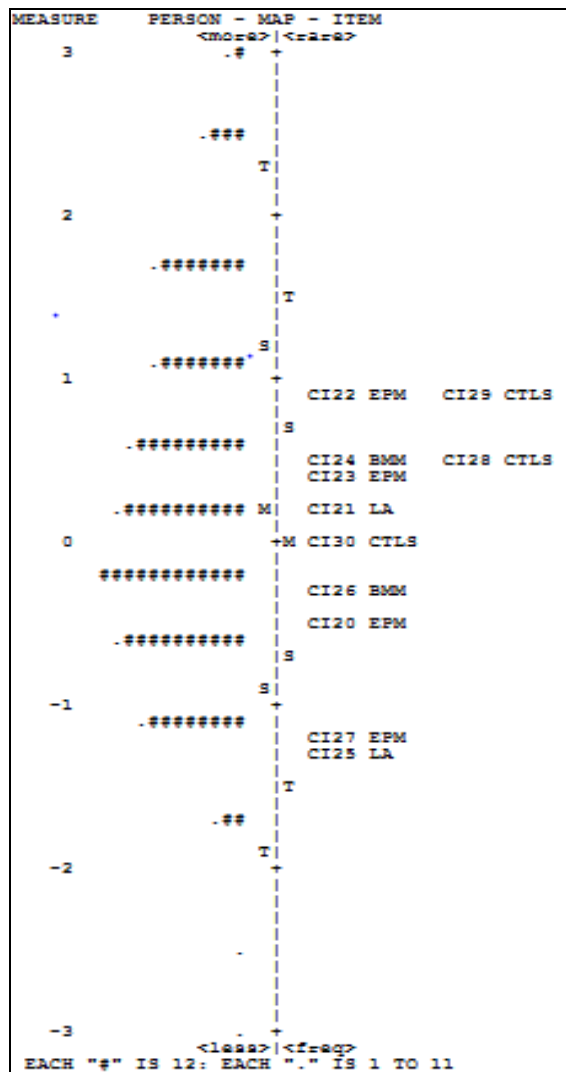


Figure 4.8 Wright Item- Person Map for Common Item Linking

Figure 4.8 illustrates the Wright item-person map for the eleven common items. The distribution of the eleven common items according to the students' ability is displayed in Figure 4.8. These common items belong to the CEFR C1 item difficulty level. There is no significant gap between the item location, and the overlap of items CI22 and CI29; as can be viewed from Figure 4.8, it is not between the same cognitive processing. They are two different processes, namely, EPM and CTLS, correspondingly. These items span between the highest logits of 0.90 to the lowest logits of -1.30.

4.2.5 Validity of Individual Tests

Since there were four different sets of tests developed, four distinct analyses of the tests were undertaken to determine the internal validity of each. To ensure the adequacy of the items of each test individually, this step was taken. This was accomplished by looking into the precision of measurement, fit statistics checking for misfitting items, particularly item polarity, as well as unidimensionality.

Table 4.9 explains the summary of the reliability indices of all four tests. The total number of students for each test applied, in the final analysis, is given in the second column of Table 4.9. All tests showed high statistical values for item reliability at 0.98 or 0.97, which is a good indication for the measurement. Except for Test1 and Test2, other tests had acceptable person reliability indices, even Test 2 possessed a value of 0.79 for person reliability, which is very close to 0.80, the acceptable statistics. Similarly, the separation indices for items of all four tests indicated a minimum of six to seven categories (6.45 means in Test1 means 7 types of item difficulty levels), whereas person separation for Test 1 and Test 2 is a little lower, showing only two types of ability level students.

Table 4.9 Summary of Reliability Indices of all Four Tests

Test	No of persons	Person reliability	Item reliability	Person separation	Item separation	Person mean measure	Item mean measure	Person SD	Item SD
Test 1	208	0.73	0.98	1.64	6.45	0.45	0.00	0.39	0.18
Test 2	247	0.79	0.97	1.94	5.26	0.56	0.00	0.38	0.15
Test 3	268	0.83	0.97	2.19	5.84	0.62	0.00	0.39	0.15
Test 4	179	0.82	0.97	2.11	5.28	0.91	0.00	0.41	0.20

The means of item difficulty measure of all tests were 0.00, whereas the means for student ability were between 0.45 to 0.91. Curtis and Boman (2007) suggested that a small difference of below 0.50 between the item and person means measures indicates that the test targeted students quite accurately; however, Davis and Boone (2021) mentioned that the difference of less than 1.00 logit indicates good targeting of items and persons. Therefore, all the tests have been well-designed, targeting the population appropriately. The means standard error for the students in all four tests were comparably high with that of means standard error of items. All means for the students' standard error (person SD) were higher than the value for $2/\sqrt{40}$, which is above 0.32 and lower than $3/\sqrt{40}$, which means below 0.47. This was substantially within the target range of 0.32 to 0.47 ($2/\sqrt{40} < SEM < 3/\sqrt{40}$) ($0.32 < SEM < 0.47$). Item SDs for all tests were equal to or less than 0.20 indicating a high level of precisions. With the use of a wide group of students with varying skills, all the tests showed an acceptable precision of measurement.

Table 4.10 Summary of Fit Statistics and PCA Residuals of all Four Tests

Tests	No of persons	Infit MNSQ	Outfit MNSQ	PTM Corr. (person)	λ Eigenvalue in 1st Contrast	PCA Residuals RVEM(%)
Test 1	208	0.99	1.02	>0.16	2.58(4.8%)	26.0
Test 2	247	1.00	0.99	>0.18	4.41(8.7%)	20.7
Test 3	268	1.00	0.99	>0.10	2.76(5.3%)	22.7
Test 4	179	0.96	0.96	>0.14	2.95(5.4%)	27.3

Table 4.10 explains the summary of the mean fit statistics and PCA residuals of all four tests. The means of infit MNSQ and outfit MNSQ of all tests were within the recommended range of 0.70-1.3, as can be seen in Table 4.10. There were no tests identified with negative point measure correlation coefficient (PTM Corr.), and all of the items scored at least above 0.10 correlation coefficient. Although the PCA residuals of raw variance explained by measures, should be above 40% of the total items, it is not a concern of unidimensionality, according to Linacre (2018); the eigenvalue in the first contrast should be less than 10% strength (Linacre, 2020) to indicate that there is no secondary dimension. According to the results of these tests, it can be concluded that the four tests are unidimensional. Hence, each test proved to be reliable and valid, considering the validity evidence from item polarity, item fit, and PCA residuals, as well as the reliability evidence from the precision of measurement.

4.2.6 Validity of Students' Responses

The summary of fit statistics for the student responses can be seen in Table 4.11 (see Appendix G.1.d.). Out of a total of 902 students, 863 had a reasonable infit MNSQ between 0.7 and 1.3, which was 95.7%; however, only 683 students (75.7%) had a reasonable outfit MNSQ between 0.7 and 1.3. Nevertheless, the difference in the outfit MNSQ will not affect the measurement significantly.

Table 4.11 Summary of Person Fit Statistics

Mean- Square Value	Infit		Outfit	
	Frequency	Percentage	Frequency	Percentage
Below 0.7	6	0.7	118	13.1
0.7 – 1.3	863	95.7	683	75.7
Above 1.3	33	3.7	101	11.2
Total	902	100	902	100
Mean	1		0.99	
S. D	0.15		0.29	

Furthermore, the mean of the infit MNSQ value was 1.00 logit, which is as perfectly expected by the model's expected value of 1.00. The mean of the outfit MNSQ value 0.99 was also very close to the model's prediction. For both infit and outfit, the standard error was not far off. Infit MNSQ had a value of 0.15 logits, while outfit MNSQ had a value of 0.29 logits. All these statistics matched the predictions of the Rasch measurement model.

DATA	OBSERVED	EXPECTED	RESIDUAL	ST. RES.	MEASDIFF	ITEM	PERSON	ITEM	PERSON
0	0	.99	-.99	-10.92	4.78	8	716	CI27 EFM	FAS261
0	0	.99	-.99	-8.11	4.19	115	844	T4Q17 LA	FE121
0	0	.98	-.98	-7.96	4.15	72	688	T3Q3 WR	FAS233
0	0	.98	-.98	-7.85	4.12	85	597	T3Q16 EMM	FAS142
0	0	.98	-.98	-7.82	4.11	109	871	T4Q11 EFM	FE148
0	0	.98	-.98	-7.19	3.95	15	199	T1Q4 SP	FAC199
0	0	.98	-.98	-7.19	3.95	15	192	T1Q4 SP	FAC192
0	0	.98	-.98	-7.19	3.95	21	164	T1Q10 SP	FAC164
0	0	.98	-.98	-6.96	3.88	89	470	T3Q31 CTLS	FAS015
0	0	.98	-.98	-6.90	3.86	93	695	T3Q35 I	FAS240
0	0	.98	-.98	-6.81	3.84	40	202	T1Q40 CTLS	FAC202
0	0	.98	-.98	-6.73	3.81	21	143	T1Q10 SP	FAC143
0	0	.98	-.98	-6.30	3.68	15	57	T1Q4 SP	FAC057
0	0	.98	-.98	-6.26	3.67	115	871	T4Q17 LA	FE148
0	0	.97	-.97	-5.91	3.55	15	177	T1Q4 SP	FAC177
0	0	.97	-.97	-5.91	3.55	15	51	T1Q4 SP	FAC051
0	0	.97	-.97	-5.54	3.42	21	130	T1Q10 SP	FAC130
0	0	.97	-.97	-5.54	3.42	15	121	T1Q4 SP	FAC121
0	0	.97	-.97	-5.49	3.41	31	188	T1Q31 CTLS	FAC188
0	0	.97	-.97	-5.46	3.40	102	866	T4Q4 LA	FE143
0	0	.97	-.97	-5.33	3.35	75	599	T3Q6 SP	FAS144
0	0	.96	-.96	-5.24	3.31	79	549	T3Q10 SP	FAS094
0	0	.96	-.96	-5.24	3.31	79	498	T3Q10 SP	FAS043
0	0	.96	-.96	-5.15	3.28	111	737	T4Q13 LA	FE014
0	0	.96	-.96	-5.15	3.28	106	754	T4Q8 SP	FE031
0	0	.96	-.96	-5.15	3.28	68	309	T2Q39 EFM	FMC101
0	0	.96	-.96	-5.15	3.28	45	285	T2Q5 WR	FMC077
0	0	.96	-.96	-5.08	3.25	57	290	T2Q17 EMM	FMC082
0	0	.96	-.96	-5.00	3.22	71	599	T3Q2 SP	FAS144
0	0	.96	-.96	-4.99	3.22	109	882	T4Q11 EFM	FE159
0	0	.96	-.96	-4.91	3.18	18	102	T1Q7 SP	FAC102
0	0	.96	-.96	-4.88	3.17	107	785	T4Q9 LA	FE062
0	0	.96	-.96	-4.87	3.17	113	850	T4Q15 EFM	FE127
0	0	.96	-.96	-4.82	3.15	56	379	T2Q16 EMM	FMC171
1	1	.04	.96	4.71	-3.10	117	727	T4Q19 CTLS	FE004
0	0	.96	-.96	-4.69	3.09	115	849	T4Q17 LA	FE126
0	0	.96	-.96	-4.69	3.09	115	758	T4Q17 LA	FE035
0	0	.96	-.96	-4.69	3.09	115	736	T4Q17 LA	FE013
0	0	.96	-.96	-4.68	3.09	59	289	T2Q19 EFM	FMC081
0	0	.96	-.96	-4.64	3.07	109	768	T4Q11 EFM	FE045
0	0	.96	-.96	-4.64	3.07	109	730	T4Q11 EFM	FE007
0	0	.95	-.95	-4.60	3.05	72	608	T3Q3 WR	FAS153
0	0	.95	-.95	-4.60	3.05	72	511	T3Q3 WR	FAS056
0	0	.95	-.95	-4.55	3.03	67	289	T2Q38 CTR	FMC081
0	0	.95	-.95	-4.55	3.03	102	743	T4Q4 LA	FE020
0	0	.95	-.95	-4.52	3.02	75	605	T3Q6 SP	FAS150
0	0	.95	-.95	-4.52	3.02	75	464	T3Q6 SP	FAS009
0	0	.95	-.95	-4.52	3.02	99	775	T4Q1 SP	FE052
1	1	.05	.95	4.46	-2.99	38	149	T1Q38 EFM	FAC149
1	1	.05	.95	4.40	-2.97	117	725	T4Q19 CTLS	FE002

Figure 4.9 Most Unexpected Responses of the Students

Figure 4.9 portrays the WINSTEPS output table indicating the most unexpected responses of the students. Many students, as shown in Figure 4.9, were influenced by unexpected responses. Although the items were easy enough for them to answer correctly, as expected by the Rasch measurement model, they have responded wrongly to the items, which may also cause the high outfit mean square, as many high achievers answered wrongly when they were expected to get them right. For example, Item CI27 EPM was wrongly answered by the person identified as FAS261, even though the model expected the person to answer it correctly (0.99). Similarly, Item T4Q19 CTLS was expected to be answered wrongly by the model (at 0.04); however, this was answered correctly by the student named FE004 (see Figures 4.9 and 4.10). Comparably, there were many observed data scoring “0”, while the model expectations were between 0.99 and 0.95; only a few items had the observed data score “1”, while the model expectations were close to 0.00 or 0.05, as depicted in Figure 4.9. Overall, there were unexpected responses, some high ability students missed easy questions or vice-versa. Although this may be due to carelessness or lucky guessing, as Linacre (2021) stated, it could be due to other reasons, too. This could possibly be due to the selection of fewer high ability students.

```

MOST MISFITTING RESPONSE STRINGS
| ITEM
| 11 1 1 11 1 1 1 1 1
|217107115420110304947703 50 06475718856809 1 92133
OUTMNSQ |15269975797281756096634160088855613159795320382798 PERSON
high-----
3.00 A| . . . . . 0 . . . . . 716 FAS261
2.98 B| .0 0 . . . . . . . . . . 871 FE148
2.44 C| 0 . . . . . . . . . . 688 FAS233
2.30 D| .0 . . . . . 0 . . . . . .1 199 FAC199
2.29 E| .0 . . . . . 0 0 . . . . . 192 FAC192
2.26 F| . . 0 . . . . . . . . . . 0 . . . . . 844 FE121
2.19 G| .0 . . . . . 0 . . . . . . . . . . 164 FAC164
2.05 H| . . . . . . . . . . . . . . . . 1.1 1 727 FE004
2.00 I| .0 . . . . . 0 . . . . . . . . . . .1 177 FAC177
1.97 J| .0 . . . . . 0 . . . . . . . . . . . 143 FAC143
1.85 K| .0 . . . . . . . . . . . . . . . . 1. 57 FAC057
1.82 L| .0 . . . . . 0 . . . . . . . . . . . 130 FAC130
1.80 M| 0 0 . . . . . . . . . . . . . . . . 678 FAS223
1.80 N| . . . . . . . . . . . . . . . . . . . . .1 1 725 FE002
1.74 O| .0 . . . . . 0 . . . . . . . . . . . . . . . 51 FAC051
1.73 P| . . . . . 0 . . . . . . . . . . . . . . . 1 11 186 FAC186
1.72 Q| 0 . . . . . . . . . . . . . . . . . . . . .1 . 804 FE081
1.71 R| .0 0 . . . . . . . . . . . . . . . . . . . . . 121 FAC121
1.71 S| .0 . . . . . 0 . . . . . . . . . . . . . . . . 768 FE045
1.70 T| . 0 . . . . . . . . . . . . . . . . . . . . .111 532 FAS077
1.66 U| . . . . . 0 . . . . . . . . . . . 0 . . . . . 866 FE143
1.66 V| .0 . . . . . 0 . . . . . . . . . . . . . . . . 882 FE159
1.63 W| . . . . . . . . . . . . . . . . . . . . .1 227 FMC019
1.61 Y| . . . . . . . . . . . . . . . . . . . . .1. 426 FMC218
1.61 Z| . . . . . . . . . . . 0 0 . . . . . . . . . . 599 FAS144
low-----
|21711711542111131494771365181647571885681921392133
|152109717970810500966301 00 08556111597903 0 82198
| 69 5 2 17 6 4 0 8 3 5 7

```

Figure 4.10 Most Misfitting Students' Response Strings

Figure 4.10 describes the strings of the most misfitting students' responses. The figure indicates that the student's response patterns have contributed to high person-misfit statistics. Low achievers denoted by the ID numbered FE004, FE002, FAC186, FAS077, etc., might resort to lucky guessing or higher achievers like FAS261, FE148, FAC192, etc., might perhaps be somewhat lackadaisical in answering the questions, which could cause the high misfit numbers, as Linacre (2021) suggested.

4.2.7 Summary of Acceptability of Reading Tests

Based on the analysis for the concurrent analysis, common item linking, individual tests, and students' response validation, the findings of the study revealed that the tests were highly valid for describing the students' reading performance.

4.3 STUDENTS' READING PERFORMANCE ALIGNED WITH CEFR LEVEL

Since the study utilized the CEFR-aligned reading passages, it is one of the concerns of the study to evaluate the reading performance level of the selected students on the CEFR level. Thus, the findings related to the reading performance levels of all four faculty students are presented in this section.

4.3.1 CEFR Levels of the Tests

Reading passages for the tests were selected from the Learning Resource Network materials, which are aligned with the CEFR levels. Table 4.12 illustrates the CEFR levels of the LRN passages and the readability analysis of the passages, as per the evaluation of the Text Inspector software analysis.

Table 4.12 Readability Indices of the Selected Passages

Passage No	LRN CEFR Level	LRN CEFR	Text Inspector CEFR level	Flesch Reading Ease	FRE total	Mean FRE total
Test 1 P1	B1	3	B1+	71.93	237.91	59
Test 1 P2	C1	5	B2+	70.66		
Test 1 P3	C1	5	C1+	46.25		
Test 1 P4	IELCA(M/L)	6	C1+	49.07		
Test 2 P1	B2	4	B2+	45.86	184.89	46
Test 2 P2	B2	4	C2	52.26		
Test 2 P3	C1	5	C1+	46.25		
Test 2 P4	IELCA(M/L)	6	C2	40.52		
Test 3 P1	B1	3	B1+	61.89	200.44	50
Test 3 P2	C1	5	C1+	50.13		
Test 3 P3	C1	5	C1+	46.25		
Test 3 P4	IELCA(M/L)	6	C2	42.17		
Test 4 P1	B2	4	A2+	76.46	241.55	60
Test 4 P2	B2	4	B2	62.19		
Test 4 P3	C1	5	C1+	46.25		
Test 4 P4	IELCA(M/L)	6	C1	56.65		
Average	C1	5 (19/4=4.75)				54

The overall CEFR level of the LRN passages is calculated by measuring the value for the CEFR levels. From the CEFR A1 to C2, there are six levels and the value for each CEFR level is achieved accordingly as given under the column “LRN CEFR”. As IELCA (International English Language Competency Assessment) passages belong to multi-level test difficulty targeting the B2 level to C2 level, they possess the maximum value (6). So, the total value of the CEFR level LRN passages in each test was 19 (for example, the total of Test 1 is 3+5+5+6=19). Therefore, the mean of four passages is 5 ($19/4 = 4.75$), which denotes the CEFR C1 level.

The Flesch-Kincaid Reading Ease statistics for each passage is given in the fifth column of Table 4.12. The means of Flesch-Kincaid reading ease are 59, 46, 50, and 60, respectively, for Test 1 to Test 4. The cumulative mean of Flesch-Kincaid reading ease for all tests is 54.

Linguapress, an English language learning website (*A comparison of different readability scales*, n.d.), has mapped Flesch-Kincaid reading scores onto the CEFR levels. According to this website, the ranges of Flesch-Kincaid reading ease between 50 and 60 are conflated to be CEFR C1 level. A similar approach has been presented by Natova (2019), who claims that the range of 50-60 Flesch-Kincaid reading ease belongs to the CEFR C1 level. Based on these works, the present study justifies that each test probably belongs to the CEFR C1 level, as the mean of the Flesch-Kincaid reading ease of all four tests was reported to be 54.

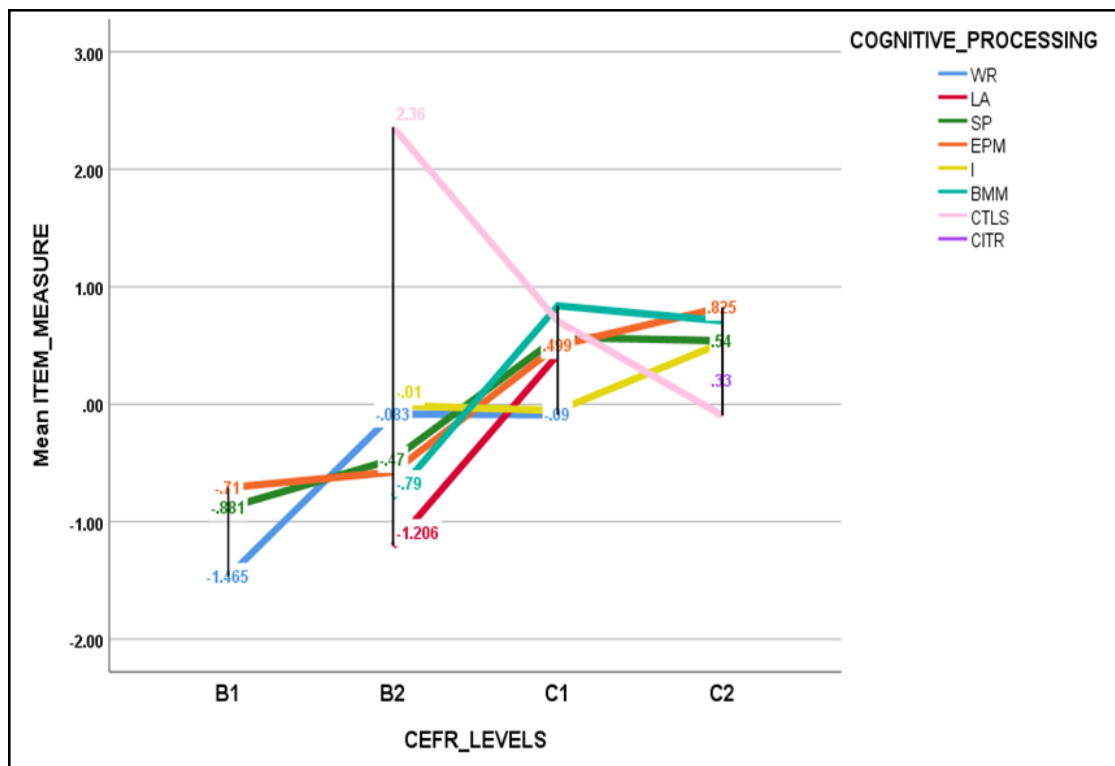


Figure 4.11 Mean Item Measure of Cognitive Processing along with CEFR Levels

Figure 4.11 presents a multiple line graph of individual items as per their representation of the CEFR level and cognitive processing. It is noted that the CEFR levels from B1 to C2 levels have an ascending order (hierarchy) of item difficulty levels. Most of the items in the line graph show a reasonable incline for the cognitive processes according to their CEFR level except for one, which is known as the process of CTLS. In the CEFR B2 level, it has the highest span of 2.36 logits, whereas in the C2 level, it does have the lowest location of -1.10 logits. The data value labels of each line were typed in the same colour as their cognitive processing. The purple colour data value label represents only one item belonging to the CITR reporting at 0.33 logits. Except for the B2 level, which has a span between -1.206 and 2.36 logits, all the other three categories of the CEFR levels have the span of similar ratios; for instance, the distribution of B1 spans between -1.465 and 0.71 logits, and the distribution of C2 level spans between -0.26 and 0.825 logits.

4.3.2 Grading Scheme of Tests

The LRN ESOL international tests for the A1 level to the C2 level have four components, such as listening, reading, writing, and speaking. However, their grading system is different for each test level. For example, the LRN ESOL international A1 level test has 20 reading questions, which carry 20 marks, whereas the C1 level test has 50 reading items with 25 total scores for reading tests (www.lrn-global.org, n.d.-a). Similar to this, all the other skills share equal marks. However, their grading systems represent fail, pass, merit, and distinction categories. Therefore, considering a feasible effective scoring system, the grading system of the IELCA academic reading test (issued by the LRN, covering the CEFR B2 to C2 levels), is applied in this test. Since the test passages in the study are related to reading for information, orientation, and argumentation, the scoring procedure for the IELCA academic reading test was employed. Further, this test has 40 reading questions possessing one mark for each. Hence, applying this grading system to reading tests consisting of 40 items is comparably efficient. Table 4.13 portrays the grading system of the IELCA academic reading test.

Table 4.13 Grading system of IELCA Academic Reading Test

IELCA Academic Reading		
Score & CEFR Level		Raw Score out of 40
10	A2	0-10
20	B1	11-22
30	B2	23-34
40	C1	35-39
50	C2	40

(Adopted from Qualification Specification - IELCA Retrieved from http://www.lrnglobal.org/web_docs/qualifications/esolint/ielca/ielca_spec.pdf)

The target of the texts and test items do not focus on the A1 or A2 levels, as the tests are for university students, denoting their academic achievements and reading performance. Therefore, as can be seen from Table 4.13, the grading of the tests starts from the A2 level. Based on this grading, a student achieving 40 marks out of 40 items will be placed on the CEFR C2 level. However, it is noted that the highest difficulty level of the test is the CEFR C1 level.

4.3.3 Students' Performance Level

This section presents the findings that relate to the second research question. The RQ2 was answered by two methods. The first method applied the Classical Test Theory (CTT), while the second approach employed the Item Response Theory (IRT).

RQ2: What is the performance of the students in the CEFR- aligned reading tests?

The first method applying the CTT is discussed first to answer the RQ2. This theory of testing is based on the concept that a person's observed score on a test is the sum of a true score and an error score (Hambleton & Jones, 1993). As Suen (2012) suggested that this is a leading measurement theory, it is sensible to apply this method to evaluate the performance levels of students using this approach.

Table 4.14 Summary of Test Scores in four Tests according to CEFR levels

CEFR Level	Test1	%	Test2	%	Test3	%	Test4	%	All Tests	%
A2	1	0.5	3	1.2	3	1.1	0	0.0	7	0.8
B1	88	42.3	94	38.1	102	38.1	56	31.3	340	37.7
B2	114	54.8	135	54.7	144	53.7	110	61.4	503	55.8
C1	5	2.4	15	6.1	19	7.1	13	7.3	52	5.8
C2	0	0	0	0	0	0	0	0	0	0
Total	208	100	247	100	268	100	179	100	902	100

Table 4.14 explains the summary of the test scores based on the CEFR levels in four tests. The test scores were measured according to the IELCA academic reading criterion. Table 4.14 signifies that the majority of the students were categorized between the CEFR B1 and B2 levels. When measuring the total percentages of these two categories as a whole in all four tests, 93.5% of them, which means 843 out of a total of 902 students fall under these categories.

Compared to all four tests, Test 4, which was administered to the Faculty of Engineering students, achieved the highest performance, indicating 0% for the lowest level (A2), a lower percentage (31.3%) for the other lower level (B1) of the scoring, and the highest percentages, 7.3 % and 61.4% (a total of 68.7%), for the highest levels of C1 and B2, respectively. Similarly, Test 2 and Test 3 scored similar percentages in almost all the levels except for the B2 and the C1 levels. In these two levels, Test 3 performance indicated a higher level as it gained 7.1% for the C1 level, whereas Test 2 gained 6.1%. The scores reported for Test 1 illustrated that this test is counted for the minimum performance, indicating higher percentages for the lower levels (42.5% for the B1 level), and lower percentages for the higher levels (2.4% for the C1 level). Figure 4.12 and Table 4.15 illustrate the same in graphical forms.

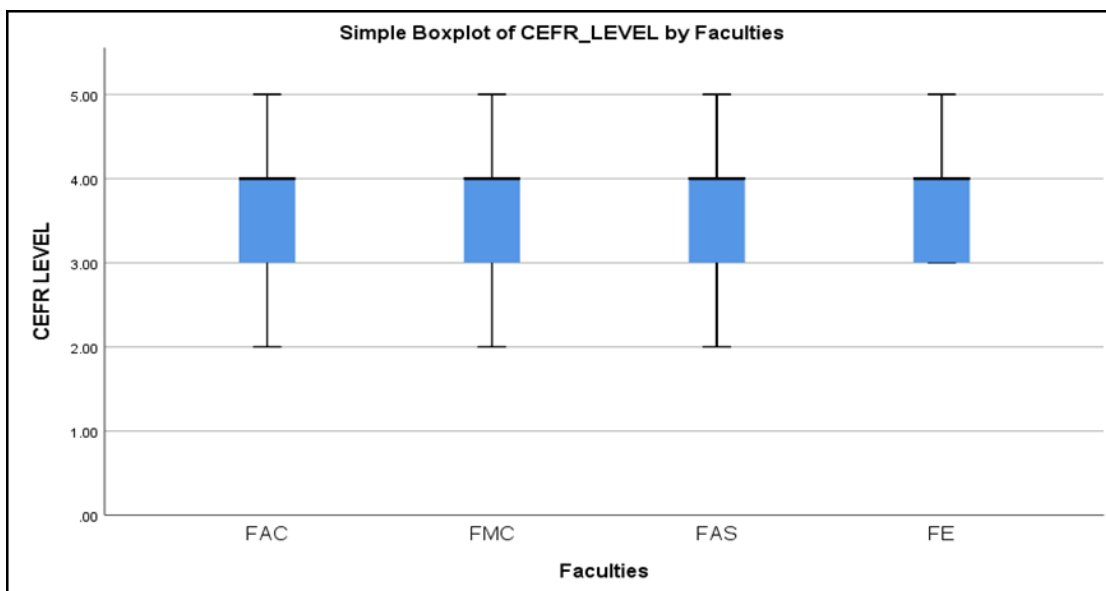


Figure 4.12 Boxplot for Inter-Faculty Reading Performance in CEFR-aligned Test

Figure 4.12 portrays the performance of the four faculties based on the CEFR levels in the four reading tests, utilizing the Boxplot analysis of the SPSS.

Table 4.15 Descriptive Statistics of Students' Performance

Faculties	N	Minimum	Maximum	Mean	Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error
FAC	208	2.00	5.00	3.59	.038
FMC	247	2.00	5.00	3.66	.039
FAS	268	2.00	5.00	3.67	.038
FE	179	3.00	5.00	3.76	.043

Table 4.15 shows the performance of the four faculties based on the CEFR levels in the four reading tests, utilizing descriptive statistics. The maximum score achieved by students in these tests is 5.00, which is represented by the CEFR C1 level. At least some students from each faculty achieved this level. The minimum statistics scored by them is 2.00, except for the students of the Faculty of Engineering (FE), who gained 3.00 as the minimum score. So as shown in Figure 4.12, their distribution

is between the CEFR levels 3 and 5. The mean statistics for the FE is the highest, indicating 3.76. However, the lowest standard error (0.38) and standard deviation (0.55) were reported for the FAC.

4.3.4 Students' Performance Level according to Faculty Background

The second method to answer the RQ2 is using the Item Response Theory (IRT). The classical test theory assumes that each person has a true score that would be earned if there were no measurement errors. Nevertheless, measuring instruments are flawed, each person's observed score may differ from their true ability (Magno, 2009). Therefore, the approach of the IRT, which measures the true ability of the students according to the item difficulty, is applied in this analysis.

Table 4.16 Reading Performance of the Four Faculties

Faculty	N	Minimum	Maximum	Mean	Std. Error	SD
FAC	208	-1.36	3.09	0.53	0.05	0.78
FMC	247	-1.22	4.09	0.73	0.05	0.84
FAS	268	-1.52	4.23	0.88	0.06	0.98
FE	179	-1.02	4.21	0.95	0.07	0.99

The distributions of reading performance measures on the scale of inquiry, as well as in all cognitive processing, were used to assess four faculty students' reading performance, namely, FAC, FMC, FAS, and FE. Table 4.16 demonstrates the finding of the reading tests according to IRT analysis using Rasch MM. The means logit measures of the four faculties have been arranged in an ascending order, illustrating the FAC (0.53) with the lowest and the FE (0.95) with the highest performance. It shows that the FE students performed the best compared to all four faculty students, scoring a mean of 0.95 logits.

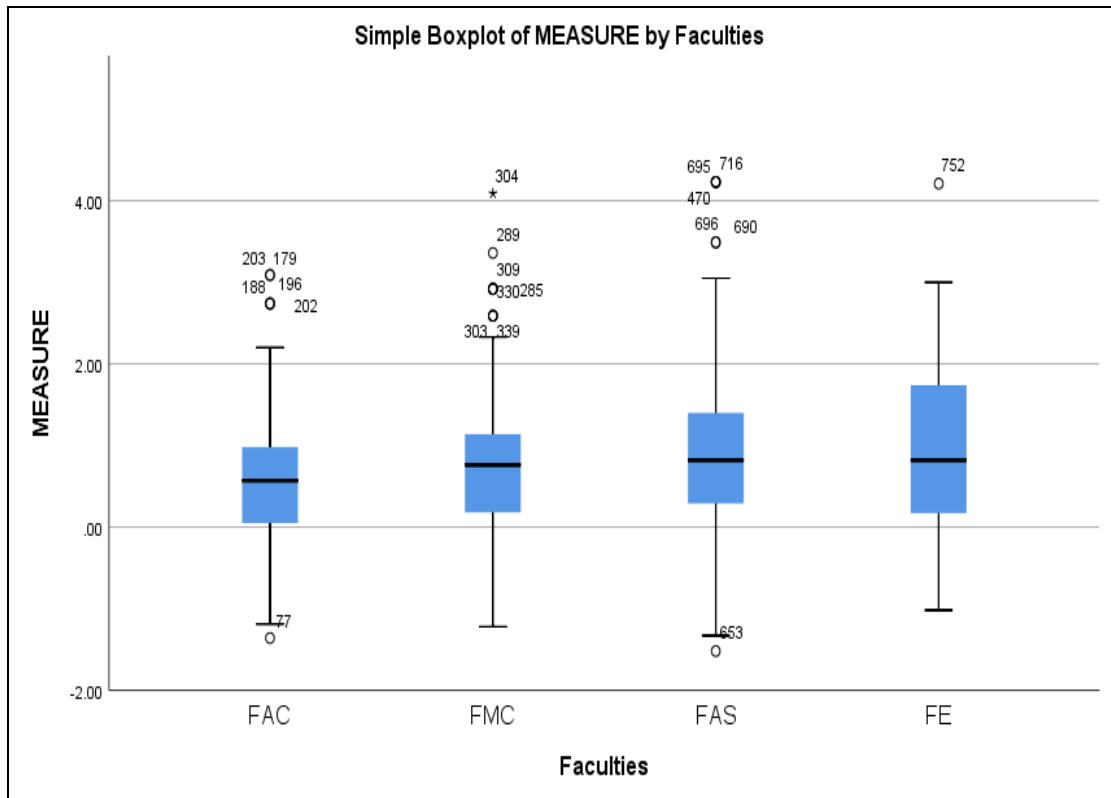


Figure 4.13 Distribution of Reading Performance of FAC, FMC, FAS, and FE Students on Logit Scale

Figure 4.13 shows the distributions of reading performance of students of the four faculties, FAC, FMC, FAS, and FE, on the logit scale. It is clear that the FE students fared the best, their distribution spanned between -1.02 to 4.21, and around half of the students scored above the mean logits (0.95). The span of the FAS students was 5.75 (-1.52 to 4.23), which is the highest span compared to the other faculties. (The spans of 4.45, 5.31, and 5.23 were reported by the FAC, the FMC, and the FE, respectively). Although the students of the FMC (mean 0.73 logits) scored better than those of the FAC (mean 0.53 logits), their performance was a little lower than that of the FAS students.

The maximum score of 4.23 logits was achieved by four students belonging to the FAS. Figure 4.14, illustrating the Wright person map portrays the same, (here “A” refers to the Faculty of Arts and Culture, “M” refers to the Faculty of Management and Commerce, “S” refers to the Faculty of Applied Sciences, whereas “E” refers to the Faculty of Engineering). Although six persons seem to have achieved the same

Both the CTT and the IRT analyses of the students' performance in reading tests indicate similar findings. As a whole, Test 4, administered to the FE students, was reported to be the test with the highest mean statistics in both investigations. The ranking of the mean statistics of both analyses was the same as the FAC, scoring the lowest, while the FE gained the highest; however, there were minor differences that can be seen in the ranking of the standard error and standard deviation of all tests between the two analyses.

4.4 COGNITIVE PROCESSING IN READING

This study aims to establish the performance level of the SEUSL undergraduates in English reading skills, as well as to determine the cognitive processes they excelled in, and those whereby they attained lower levels. As a result, the analysis and discussion in this section are based on the reading performance levels of all students, as a whole. The dichotomous analysis was performed to calculate the item difficulty measure of individual items of each test. The outcomes of the analyses based on the following research questions are presented in this section.

RQ3: What is the performance level of the SEUSL undergraduates who follow the EMI system, in the cognitive processes of English reading,?

- a. In which cognitive processes of reading do the SEUSL students indicate higher achievement?
- b. In which cognitive processes of reading do the SEUSL students indicate lower achievement?

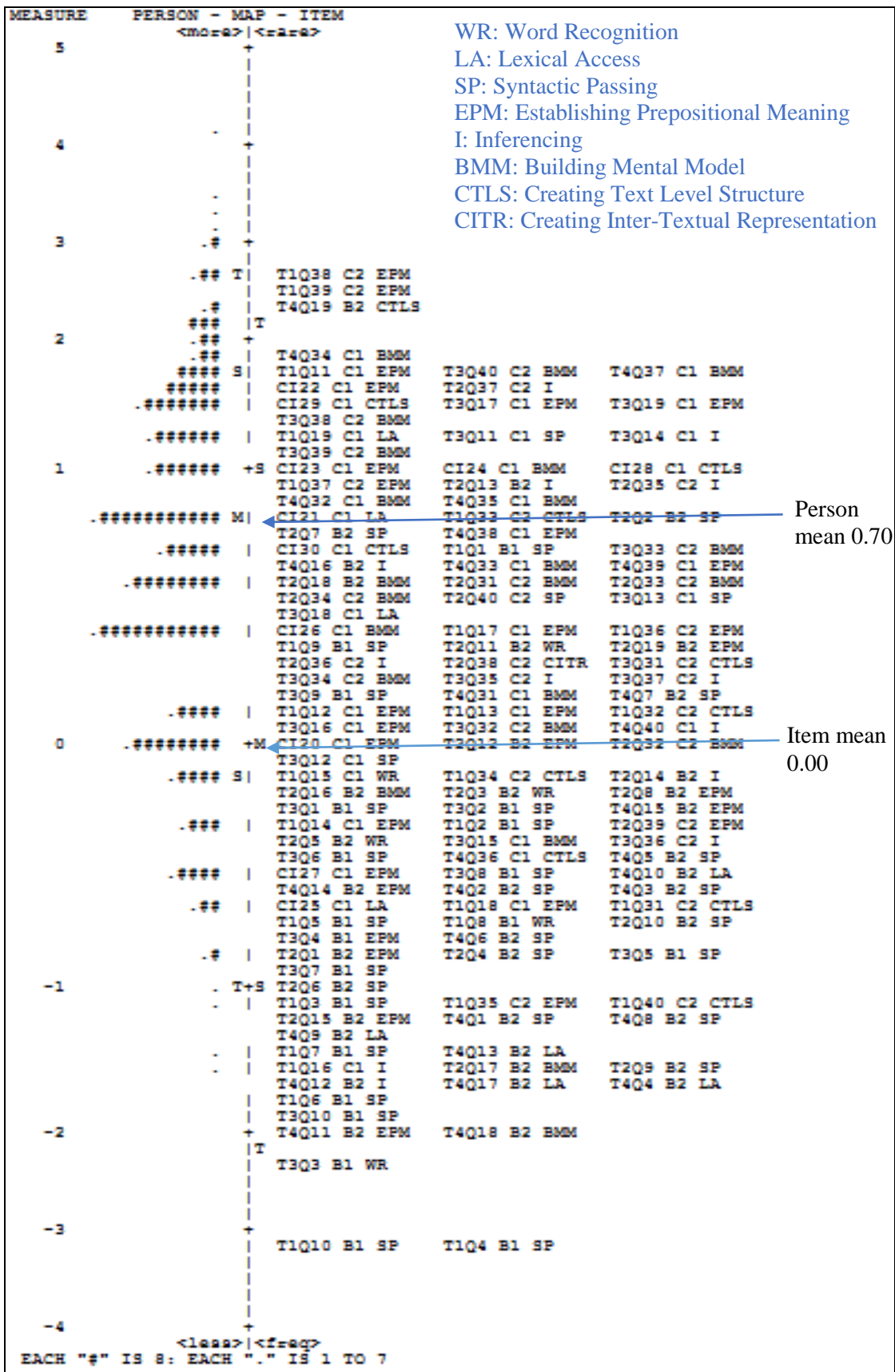


Figure 4.15 Wright Item Person Map: Students' Performance on Reading Tests

The distribution of all cognitive processing categories of reading involved in each item included in the reading exam is depicted in Figures 4.15 and 4.16, respectively. The levels of difficulty in each processing were different. According to the mean results of the processes, the most difficult process was building a mental model (BMM) with the highest mean of 0.50 logit, and SD of 0.91, however, only a single item measuring the process of the CITR was ranked to the eighth place in the hierarchy order. The easiest processes were identified as word recognition (WR) (mean = -0.55, SD = 0.90), as well as syntactic parsing (SP), with the same mean (-0.55) but different SD (1.03).

4.4.1 Cognitive Processes Achieved by Many Students

Table 4.17 provides the descriptive statistics for the cognitive processing of reading. This includes the number of items measuring the processing, the minimum measure of the processing, maximum logit, means, ranking, mean errors, and standard deviations for all cognitive processing categories. 127 items were designed to measure the eight cognitive processing of reading variances.

Table 4.17 Descriptive Statistics for Cognitive Processing

Cognitive Processes	N	Minimum	Maximum	Mean	Rank	Std. Error	SD
WR	6	-2.29	0.25	-0.55	1	0.37	0.90
LA	9	-1.53	1.08	-0.48	3	0.34	1.02
SP	34	-3.24	1.22	-0.55	1	0.18	1.03
EPM	30	-1.97	2.66	0.24	5	0.19	1.07
I	13	-1.48	1.54	0.23	4	0.26	0.93
BMM	23	-1.97	1.80	0.50	8	0.19	0.91
CTLS	11	-1.10	2.36	0.42	7	0.30	0.98
CITR	1	0.33	0.33	0.33	6	-	-

According to Khalifa and Wier (2009), there is a hierarchical order among the cognitive processing indicating WR is the easiest and CTR is the most difficult processing. A test targeting the C1 level of the CEFR does not focus much on the CTR or the CTLS. In addition, according to Brown's (2005) recommendations, a good test probably must have many items focusing on the average ability of students; hence, the present tests possessed many items on the average skills, rather than including more items on the two extremes (too easy or too hard) of the cognitive processing. Thus, there were 15 items shared by the first two easiest cognitive processing, namely, WR and LA, whereas, 12 items were shared by the last two most difficult processing like the CTLS and the CTR. Around 100 items tested the middle-level cognitive processes like SP, EPM, I, and BMM. Students performed well for almost all processes as can be seen from Figure 4.15.

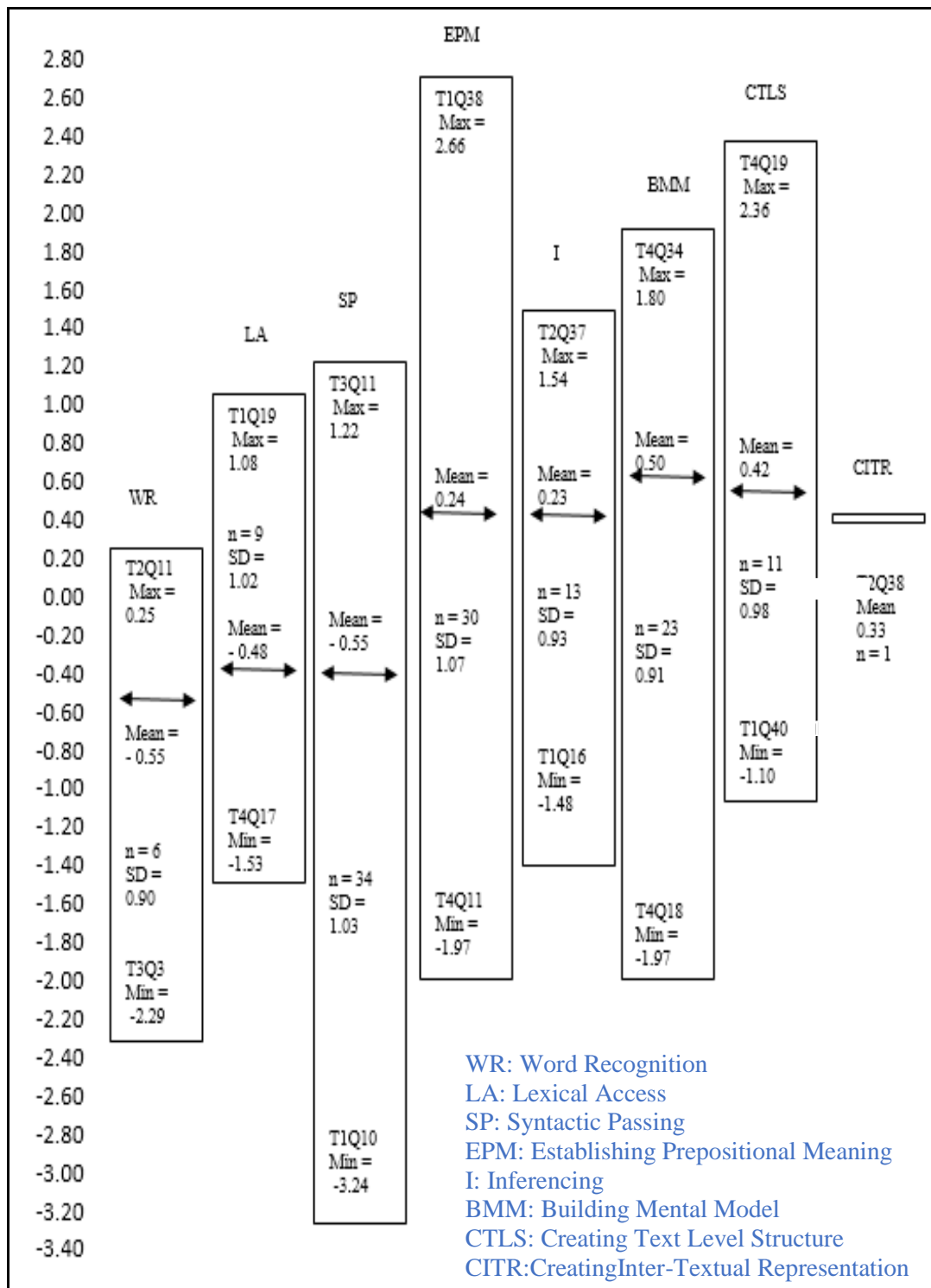


Figure 4.16 Distribution of Items Based on Cognitive Processing of Reading

Figure 4.16 explains the distribution of 127 items based on the cognitive processes of reading. Some items of particular cognitive processing were placed at the top of the scale of inquiry, while others were placed in the middle or at the bottom,

indicating whether students were able to use the particular cognitive processing as illustrated in Figure 4.16. One of the easiest cognitive processing according to the results of the present study is SP, which is ranked first, indicating that it is the easiest cognitive processing of reading. As it can be seen from Table 4.17, 34 items measuring SP processing scored -0.55 mean logit. The items having minus logit values indicate that they are easy items achieved by many students according to the Rasch Measurement model's prediction. This becomes true considering the items like T1Q10 and T1Q24 (-3.24 logit), T3Q10 (-1.92), T1Q6 (-1.66), etc. However, an item measuring this processing, for example, Item T3Q11, possessed a value of 1.22 logits indicating that this particular individual item representing SP is challenging, according to another situation.

This is the same with processing like WR too, which is also ranked number 1, the easiest. Compared to the items assessing SP, only 6 items measured the cognitive processing of WR. The distribution of WR is between -2.29 and 0.25 logits, between the minimum and maximum ends. Item T2Q11 was located at 0.25 logits, whereas item T3Q3 was at -2.29 logits, which measured the same cognitive processing.

The next easiest processing is LA, ranked third place, having a mean of -0.48 logits. The distribution of this processing spanned between -1.53 to 1.08 logits. Out of nine total items assessing LA, six items were easy, as they were having negative logit measures. Item T4Q17 was located at the bottom of the Wright map at -1.53 logits, whereas item T1Q19 was positioned at 1.08 logits. The present study considered the first three cognitive processes, which scored above the item mean of 0.00 logit values as the processes, which were easy to achieve by many students.

The Wright item person map (Figure 4.15) indicates that there were students above the most difficult level of items, and there were items below the lowest level of student ability. Individual items were difficult or easy for students depending on many characteristics, including the number of items measuring the cognitive processing, whether the information was explicitly or implicitly given key choice, distracter characteristics, and student familiarity with items.

4.4.2 Cognitive Processes Underachieved by Many Students

The processes, which had the mean logits above the item mean of 0.00 logit value, were labelled as cognitive processing in which students showed a lower achievement.

Table 4.18 Ascending order of Item logit measures of Cognitive Processing

Summary of Item Measure under each Cognitive Processing							
WR	LA	SP	EPM	I	BMM	CTLS	CITR
-2.29	-1.53	-3.24	-1.97	-1.48	-1.97	-1.1	0.33
-0.64	-1.47	-3.24	-1.2	-1.47	-1.52	-0.67	
-0.36	-1.35	-1.92	-1.19	-0.34	-0.32	-0.26	
-0.14	-1.24	-1.66	-0.8	-0.25	-0.23	-0.18	
-0.09	-0.62	-1.52	-0.71	0.23	0.08	0.25	
0.25	-0.44	-1.4	-0.7	0.25	0.24	0.35	
	0.45	-1.14	-0.55	0.35	0.29	0.62	
	0.78	-1.13	-0.54	0.36	0.36	0.77	
	1.08	-1.09	-0.36	0.6	0.4	1.06	
		-0.94	-0.28	0.98	0.52	1.41	
		-0.87	-0.18	1.08	0.54	2.36	
		-0.87	-0.17	1.09	0.56		
		-0.8	-0.04	1.54	0.56		
		-0.7	0.08		0.68		
		-0.64	0.09		0.7		
		-0.62	0.1		0.93		
		-0.58	0.2		1.01		
		-0.52	0.27		1.01		
		-0.44	0.29		1.18		
		-0.3	0.34		1.31		
		-0.3	0.68		1.61		
		-0.26	0.82		1.74		
		-0.17	0.99		1.8		
		-0.17	0.99				

	WR	LA	SP	EPM	I	BMM	CTLS	CITR
			-0.07	1.27				
			0.35	1.29				
			0.35	1.42				
			0.4	1.74				
			0.54	2.57				
			0.56	2.66				
			0.71					
			0.76					
			0.87					
			1.22					
Total items	6	9	34	30	13	23	11	1
Mean	-0.55	-0.48	-0.55	0.24	0.23	0.5	0.42	0.33
Ranking	1	3	1	5	4	7	6	8

Table 4.18 explains the ascending order of item logit measures of the eight cognitive processes. The item difficulty is ranked based on the mean results of each construct. Only one item (T2Q38) out of 127 total items measured the CITR according to the expert judgment, which had a measure of 0.33 logits. This process was challenging to put in the hierarchical order because it does not have many items; therefore, the mean statistics applied in all other cognitive processes cannot be employed in analysing this process, because the mean represents a central tendency, the average value for a set of values. Since the CITR had only one item, it does not make any sense to talk about the mean value. Therefore, this process is excluded from the systematic ranking procedure applied in all the other processes. According to Khalifa and Weir (2009), CITR is the highest cognitive process involved in reading, ranking in eighth place. So, the researcher considered this process as the highest level of cognitive processes based on two reasoning; one is Khalifa and Weir's concept, and the second reason is that this process was not mainly measured in the CEFR-aligned ESOL tests, which target the CEFR C1 level or below. The maximum difficulty level of all four tests is the C1 level. Therefore, the present study decided to rank the CITR as the highest level process, ordering it the eighth place.

BMM was identified as the most challenging process, as it has the highest mean (0.50) logit in the present study (which is at the seventh ranking in the difficulty level). Out of 23 items measuring the BMM, except for 4 items, all the other items (19 items) were reported to have a positive logit value as can be seen from Table 4.18 (23:19 ratio for positive values). Item T4Q34 is the hardest item measuring BMM of 1.80 logits.

The next challenging process is CTLS, which is at the sixth ranking in the difficulty level out of the eight processes. Eleven items tested CTLS out of which four items were reported to have minus logit values and seven scored positive values (11:7 ratio for positive values). Similarly, 13 items were designed to measure the cognitive processing of I (inferencing), which also had four items possessing minus logit values (13:9 ratio for positive values). This cognitive processing was reported to be in the sixth ranking. Comparing the ratio of all these processes (23:19, 11:7, and 13:9), BMM had the highest ratio rate.

EPM was reported to have a mean value of 0.24 logits. It was identified as the fifth most challenging cognitive processing in the hierarchy from the easiest mean logit measure to the hardest mean logit measure (See Table 4.18). The distribution of items fell between the measures of -1.97 logits (Item T4Q11) and 2.66 logits (Item T1Q38). Although BMM was reported to be the most challenging cognitive processing for the students, items like T1Q38 (2.66 logits) and T1Q39 (2.57 logits), checking the EPM process, were reported to be the most challenging, scoring the highest logit measures. The logit measure of the most difficult item in BMM was 1.80 (T4Q34), which is lower compared to that of the logit measures of the most challenging items in EPM (T1Q38 and T1Q39). Finally, inferencing (I) had a mean of 0.23 logits for 13 items. Its span was 3.02 logits. The aforesaid five cognitive processes having an item mean above 0.00 logit were identified as the processes underachieved by many students.

Table 4.19 Fit Statistics of 127 Individual Items

ENTRY NO	TOTAL SCORE	TOTAL COUNT	MEASURE (logits)	MODL SE	IN. MSQ	OUT. MSQ	PTMA-E	NAME
1	578	902	0.08	0.07	1.01	1.00	0.35	CI20 EPM
2	446	902	0.78	0.07	1.1	1.11	0.38	CI21 LA
3	325	902	1.42	0.08	1.03	1.08	0.38	CI22 EPM
4	405	902	0.99	0.07	1.11	1.14	0.38	CI23 EPM
5	401	902	1.01	0.07	0.99	1.00	0.38	CI24 BMM
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
123	131	179	-0.26	0.18	1.02	0.96	0.37	T4Q36 CTLS
124	61	179	1.74	0.17	1.26	1.36	0.41	T4Q37 BMM
125	94	179	0.82	0.17	0.92	0.89	0.42	T4Q38 EPM
126	99	179	0.68	0.17	1.01	1.01	0.42	T4Q39 EPM
127	115	179	0.23	0.17	1.37	1.60	0.40	T4Q40 I
MEAN	173.7	284.1	0.00	0.16	0.99	0.98		
P.SD	109.0	193.1	1.06	0.05	.10	0.19		

Table 4.19 describes the fit statistics of the first and the last five items of the item fit statistics according to their entry order (see Appendix G.1.c, explaining the misfit order of all 127 items, to have a clear understanding of the item distribution.) It was discovered that the influence of individual items is crucial in defining the item mean measure of cognitive processing. Although the ranking of the cognitive processes was defined based on their item means, in some cases the same process is found to be easier, whereas it was found to be difficult in other situations, as was brought out by the findings of Badrasawi (2012) and Jusoh (2018). On the other hand, individual differences in reading fluency may thus be influenced by higher-level memory processes and text-level comprehension processes (Stanovich, 1982).

Figure 4.17 illustrates the distribution of cognitive processes, and their item means, which facilitates comparing and contrasting the difficulty levels of cognitive processes, their item means, and the logit value of the easiest and the hardest items in each process. Among the measures of item difficulty distribution between -3.24 and 2.66 (5.9 logits span), the mean statistics of five processes were located above the item mean of 0.0 logits and the mean of three processes was below it.

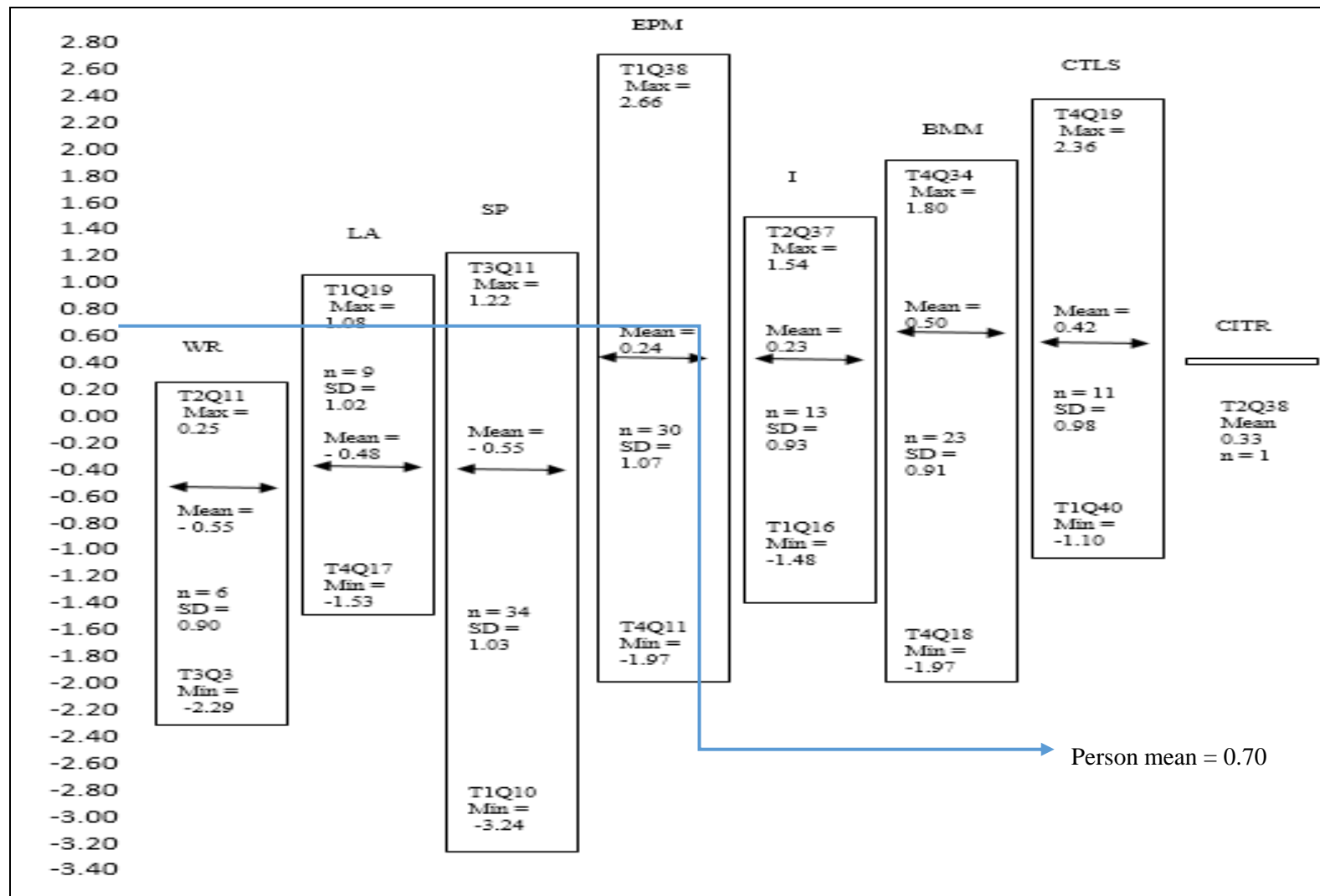


Figure 4.17 Means of the Item difficulty level of Cognitive Processing and Person

4.5 SUMMARY OF THE KEY FINDINGS

This section explains the summary of the key findings of the study involving the CEFR-aligned four reading tests across four faculties of the SEUSL, Sri Lanka. Each test includes three specific passages, and a common passage, which is common to all four tests. Finally, a test has forty items, checking the cognitive processes depicted by Khalifa and Weir (2009). The WR, LA, SP, EPM, I, BMM, CTLS, and CITR are the eight processes checked in this study. Since these tests are identified to check the CEFR C1 level, the CITR was not considered much, based on the content validation, except for a single item.

The objectives of the study are to develop the CEFR-aligned reading tests and to validate them, to check the reading performance of the students, and to profile and ascertain their cognitive processing in reading. Therefore, before analysing the main data preliminary analyses were conducted. First, the passages and the items were analysed both qualitatively and quantitatively to receive the content validation of the instruments. In this procedure, the agreement on the cognitive processes of reading was checked. Secondly, a pilot testing was carried out to check the validity of these items among 124 students representing four faculties. The majority of the items were deemed acceptable to be used in the final data collection, based on the requirements of the Rasch measurement model.

The major analysis was carried out with WINSTEPS, utilising the dichotomous analysis of the Rasch measurement model 4.4.7, and SPSS. The important findings from the analysis are summarised in Table 4.20 based on the research questions.

Table 4.20 Summary of the Key Findings

Research Questions	Findings																																																	
What are the psychometric properties of the CEFR-aligned reading tests?	Psychometric properties, like reliability and validity of the tests, were identified in three steps. First by identifying the psychometric properties of the common items, followed by the individual items of each test. Finally, the concurrent analysis of all tests was checked to evaluate the psychometric properties of the entire analysis. The results indicated that the tests are significantly valid and reliable.																																																	
What is the performance of the students in the CEFR-aligned reading tests?	<p>Four tests were carried out among 902 SEUSL students representing four faculties. A summary of the performance is given:</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center;">CEFR Level</th> <th style="text-align: center;">Test1</th> <th style="text-align: center;">Test2</th> <th style="text-align: center;">Test3</th> <th style="text-align: center;">Test4</th> <th style="text-align: center;">All Tests</th> <th style="text-align: center;">%</th> </tr> </thead> <tbody> <tr> <td style="text-align: center;">A2</td> <td style="text-align: center;">1</td> <td style="text-align: center;">3</td> <td style="text-align: center;">3</td> <td style="text-align: center;">0</td> <td style="text-align: center;">7</td> <td style="text-align: center;">0.8</td> </tr> <tr> <td style="text-align: center;">B1</td> <td style="text-align: center;">88</td> <td style="text-align: center;">94</td> <td style="text-align: center;">102</td> <td style="text-align: center;">56</td> <td style="text-align: center;">340</td> <td style="text-align: center;">37.7</td> </tr> <tr> <td style="text-align: center;">B2</td> <td style="text-align: center;">114</td> <td style="text-align: center;">135</td> <td style="text-align: center;">144</td> <td style="text-align: center;">110</td> <td style="text-align: center;">503</td> <td style="text-align: center;">55.8</td> </tr> <tr> <td style="text-align: center;">C1</td> <td style="text-align: center;">5</td> <td style="text-align: center;">15</td> <td style="text-align: center;">19</td> <td style="text-align: center;">13</td> <td style="text-align: center;">52</td> <td style="text-align: center;">5.8</td> </tr> <tr> <td style="text-align: center;">C2</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0</td> </tr> <tr> <td style="text-align: center;">Total</td> <td style="text-align: center;">208</td> <td style="text-align: center;">247</td> <td style="text-align: center;">268</td> <td style="text-align: center;">179</td> <td style="text-align: center;">902</td> <td style="text-align: center;">100</td> </tr> </tbody> </table>	CEFR Level	Test1	Test2	Test3	Test4	All Tests	%	A2	1	3	3	0	7	0.8	B1	88	94	102	56	340	37.7	B2	114	135	144	110	503	55.8	C1	5	15	19	13	52	5.8	C2	0	0	0	0	0	0	Total	208	247	268	179	902	100
CEFR Level	Test1	Test2	Test3	Test4	All Tests	%																																												
A2	1	3	3	0	7	0.8																																												
B1	88	94	102	56	340	37.7																																												
B2	114	135	144	110	503	55.8																																												
C1	5	15	19	13	52	5.8																																												
C2	0	0	0	0	0	0																																												
Total	208	247	268	179	902	100																																												
What is the performance level of SEUSL undergraduates who follow the EMI system, in the cognitive processes of English reading,?	<p>Eight cognitive processes of reading were evaluated in this study. A summary of the findings of these processes is given:</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">WR</th> <th style="text-align: center;">LA</th> <th style="text-align: center;">SP</th> <th style="text-align: center;">EPM</th> <th style="text-align: center;">I</th> <th style="text-align: center;">BMM</th> <th style="text-align: center;">CTLS</th> <th style="text-align: center;">CITR</th> </tr> </thead> <tbody> <tr> <td style="text-align: center;">Total items</td> <td style="text-align: center;">6</td> <td style="text-align: center;">9</td> <td style="text-align: center;">34</td> <td style="text-align: center;">30</td> <td style="text-align: center;">13</td> <td style="text-align: center;">23</td> <td style="text-align: center;">11</td> <td style="text-align: center;">1</td> </tr> <tr> <td style="text-align: center;">Mean</td> <td style="text-align: center;">-0.55</td> <td style="text-align: center;">0.48</td> <td style="text-align: center;">0.55</td> <td style="text-align: center;">0.24</td> <td style="text-align: center;">0.23</td> <td style="text-align: center;">0.5</td> <td style="text-align: center;">0.42</td> <td style="text-align: center;">0.33</td> </tr> <tr> <td style="text-align: center;">Ranking</td> <td style="text-align: center;">1</td> <td style="text-align: center;">3</td> <td style="text-align: center;">1</td> <td style="text-align: center;">5</td> <td style="text-align: center;">4</td> <td style="text-align: center;">7</td> <td style="text-align: center;">6</td> <td style="text-align: center;">8</td> </tr> </tbody> </table>		WR	LA	SP	EPM	I	BMM	CTLS	CITR	Total items	6	9	34	30	13	23	11	1	Mean	-0.55	0.48	0.55	0.24	0.23	0.5	0.42	0.33	Ranking	1	3	1	5	4	7	6	8													
	WR	LA	SP	EPM	I	BMM	CTLS	CITR																																										
Total items	6	9	34	30	13	23	11	1																																										
Mean	-0.55	0.48	0.55	0.24	0.23	0.5	0.42	0.33																																										
Ranking	1	3	1	5	4	7	6	8																																										
In which cognitive processes of reading do the SEUSL students indicate higher and lower achievements?	Among the measures of item difficulty distribution, five processing were located above the item mean 0.0 logits, which were hard to achieve by the students. The remaining three processes were achieved by many students.																																																	

4.6 SUMMARY OF THE CHAPTER

The findings of data analysis using WINSTEPS and SPSS were reported in this chapter. The psychometric properties of the CEFR-aligned reading tests were tested in the first analysis of the study, which focused on the validity of the test items, the precision and reliability of measurement (or the test's ability to reproduce consistent results in measurement), the construct validity of the test items, the validity of the common items that link the tests, and the validity of the students' responses. The second analysis checked the students' reading performance in the CEFR-aligned tests across four faculties. Finally, the cognitive processes of reading were discussed including a comprehensive examination of all items used in the study among students of all faculties. The data revealed that WR and SP were the easiest cognitive processes in the hierarchy, whereas BMM was identified to be the most challenging process. However, this situation can be different for individual items under each process.

CHAPTER FIVE

DISCUSSION, RECOMMENDATIONS, AND CONCLUSION

5.1 INTRODUCTION

This chapter provides an overview of the study, a summary and discussion of the major findings, and the implications of the findings on the theory, methodology, and practice. Further, it also explains the limitations of the study and the pointers for future research with recommendations, and a conclusion.

5.2 OVERVIEW OF THE STUDY

The goals of this study are three-fold. The first is to collect evidence to demonstrate the validity of the CEFR-aligned reading tests based on the cognitive processing in reading prescribed by Khalifa and Weir (2009), and the second goal is to measure the university students' reading performance according to these validated tests. The last is to investigate the students' achievement level in the cognitive processes of reading. To achieve the objectives of the study, four testlets were designed to measure eight cognitive processes of reading, namely: Word Recognition (WR), Lexical Access (LA), Syntactic Parsing (SP), Establishing Prepositional Meaning (EPM), Inferencing (I), Building a Mental Model (BMM), Creating Text Level Structure (CTLS), and Creating Inter-Textual Representation (CITR). Each testlet had been equated with others using a set of 11 common items as anchoring to link all four tests. Finally, a single test had a total of 40 selected-response (SR) objective items. The concurrent Rasch analysis was used to horizontally equate the tests of similar difficulty levels by analysing the dichotomous data.

The first phase of the study primarily focused on instrument construction using test adaptation procedures and content validation procedures, as an *a priori* test validation approach. The second stage concentrated on piloting the designed

instruments, refining them, and analysing the final data statistically, as an *a posteriori* test validation.

5.3 SUMMARY AND DISCUSSION OF THE FINDINGS

In this section, the findings of the processes of adapting test items from existing tests, the validity of the test items (item spread and target using the Rasch measurement analysis) and examinee responses, the process of ensuring construct definition, test equating procedures, and the cognitive processing and the academic reading are discussed. These are the key features, which give significant value to this research.

5.3.1 The Processes of Test Adaptation

This section provides a summary of the test adaptation procedures and a description of the factors influencing test adaptation. Test adaptation is a scientific procedure, which possesses many psychometric properties of the original test development (Hambleton, 1996; Merenda, 2006). The psychometric properties depend on the notions of validity, reliability, factor structure, and norms. To achieve the psychometric properties, the researcher must first have a solid understanding of test theory, psychometrics, multivariate statistics, factor analysis, and research methods. Thus, the adaptation procedures of the present study followed the majority of the guidelines presented by Hambleton (1996), in adapting educational and psychological tests. Overall, this study included the practices of job analysis: involving the selection of items and training in item writing, the use of an expert committee to validate the items, piloting the test, and analysing results to refine the items for the final operation.

One of the concerns that test adaptation involves is getting copyright permission (Hall et al., 2018). It is essential to get authorization from the testing agencies to use their materials for research purposes, even though it is considered a '*fair use*'. After several follow-up communications with LRN, a CEFR-aligned testing agency, permission was granted to utilize the test materials (refer to Appendix H). As a practice of job analysis, after a meticulous selection of texts along with their items

by the researcher, these selected materials were analysed by two experts in the field of language testing. This selection and the experts' evaluation were based on the text type, content, text length, item format, CEFR level, cultural and linguistic features, etc. Finally, 13 texts along with their items were adapted from the LRN, representing texts of different difficulty levels.

Another important factor in adapting a test is fit, which is to examine if the proposed items capture the spirit of the original items, as seen through the source test (Matthews-López, 2003). Therefore, a test specification was developed for the testlets used in this study. The test specifications consist of text type, content (title), the CEFR level, item format, explicit/implicit status of answering for the items, underlying constructs, and the nature of adaptation, whether adopted (original), newly constructed or adapted, were used to guide the right direction of test development. This procedure facilitated the test fit study.

After a careful selection of a common passage along with its eleven items, four test papers were designed, out of the selected 13 texts. Each test included a common passage and three more specific texts, along with their items. The test adaptation procedures were based on identifying the readability indices, received from the analysis of the *Text Inspector* software. The final assembling procedure of texts under each test was carried out after calculating the Flesch-Kincaid reading ease and the CEFR level, and qualitative analysis of the experts. All these long methodical procedures were carefully conducted for the successful adaptation of the tests. The cross-sectional analysis also helped answer the issue of fit.

The relevancy of the construct is another important factor in adapting tests. It is crucial to make sure the items measure the same construct as the source items in an appropriate, relevant, and interpretable way. Thus, a few items were created by the researcher to match some of the cognitive processes of reading, which were not covered by the original items. The number, or the ratio of adopted, adapted, and newly developed items is also a matter of importance in test adaptation. However, due to the purpose of adaptation, and the function of the test being adapted, the extent to which a test is encompassed may differ in scale (Hambleton & Bollwark, 1991). Therefore, the

present study included seven self-constructed items before they were validated at the *a priori* stage.

To ensure the proper test adaptation procedures at every step, a systematic judgement of evidence, both linguistic and psychological, was achieved, and the techniques of adaptation were carefully considered. A thorough expert judgment procedure using IOC analysis, which investigates the agreement between the items and their measured constructs, is explained in detail in Chapter Three (section 3.5.2.1.1.2.2 Quantitative Approach). Although this procedure was time-consuming and challenging, the present study applied this method to provide more value in test adaptation procedures, as an *a priori* test validation, since the IOC analysis is considered one of the more comprehensive methods for content validation (L. Crocker & Algina, 1986). Therefore, a clear discussion and explanation of the quantitative approach to content validation, which is crucial to test developers, item writers, and language testers, are in this study.

Another important process of test adaptation involved piloting the tests on representative samples of the target population to achieve *a posteriori* validation of the intended tests. Section 3.5.2.2 of Chapter Three explains the procedures involved in pilot tests in detail. However, the number of participants representing the population should reasonably resemble when administering the real tests (L. Crocker & Algina, 1986). To resolve this issue, this study employed a sample and item-free measurement model known as Rasch MM. As Boone (2016) suggested, this study collected the data for piloting and conducted a Rasch analysis to refine the instrument to finally be ready for operationalisation after the *a posteriori* validation. The processes of test adaptation involved were scientific in nature, and they were the basics of test adaptation.

5.3.2 Validity and Adequacy of the Reading Tests

Four requirements must be met for useful measurement: (1) item validity, (2) a clear definition of the construct, (3) capacity to yield consistent results with the purpose of measurement, and (4) the validity of the examinee responses (Wright & Stone, 1979).

Therefore, by applying the Rasch measurement model, the psychometric properties of the tests: the validity and the reliability, as well as the evidence supporting the claim that the tests are suitable for their intended use, were verified with regard to the abovementioned four facts (Messick, 1980). Consequently, item polarity, fit statistics, unidimensionality, reliability and separation indices, students' response validation of individual tests, as well as concurrent analysis, were assessed (Bond & Fox, 2015). Sections 4.2.2 to 4.2.7 in Chapter Four, indicate the findings for the items and person analysis comprehensively. These findings depicted good statistics for items and person reliability and validity. Therefore, it can be assured that the reading tests are adequate to measure the reading proficiency of the students.

In the university system, there is a need to create different equivalent tests to check the performance level of students in reading skills, they need to be conducted on different faculty students at different times. In the university, it is impractical to conduct exams for all faculty students simultaneously at a given time, due to a lack of physical and human resources, different academic calendars among inter-faculty administrations, etc. Therefore, to reduce the security bias when administering the same difficulty level tests, with different faculty students at different times, there is a need to create different forms of tests, which are equivalent in difficulty levels. Thus, four tests had been designed to be administered among four different cohorts, following comprehensive pilot testing and vetting procedures, using the Rasch measurement model approach.

5.3.3 Validity of Examinee Responses

The results of the student response analysis revealed that a substantial percentage of the student's responses fell within the permissible infit MNSQ range. The outfit mean square values, on the other hand, revealed that many students did not respond as the model predicted (See section 4.2.6 for further detail). This is not considered a serious threat to validity, as high outfit MNSQ can be because of a few random responses by low performers (Linacre, 2020). This can happen in two ways: (1) carelessness and (2) guessing (Linacre). Further, this suggestion was emphasised by Bond and Fox (2015), and Curtis and Boman (2007). One of the important findings revealed that the mean

item difficulty of the tests was less than 0.0, while the mean person ability of the tests is high (0.77). Therefore, the level of the examinees' ability is high compared to the level of the item difficulty. Consequently, it can be assumed that the reason for the high outfit MNSQ may be carelessness, rather than lucky guessing. However, the mean person difficulty of the tests is 0.77, which is lower than 1.0, which indicates that the tests target the examinees effectively (Curtis & Boman, 2004). Overall, the responses of the measured examinees were consistent with the Rasch model's predictions.

Useful measurement does not depend only on the use of valid items, but also on valid responses. Achieving the validity of the examinees' responses is one of the four requirements of the measurement (Wright & Stone, 1999). Thus, the findings of the tests assured this requirement.

5.3.4 Construct Definition

Construct definition was demonstrated by two methods in this study. One is through an *a priori* validity evidence, which provides a kind of job analysis evaluating test design decisions, and the evidence supporting these decisions during the test adaptation procedure. The other is through an *a posteriori* validation procedure, which focuses on scoring validity, criterion-related validity, and consequential validity, after the operationalisation of the adapted tests. These two pieces of evidence illustrate how far the reading tests have defined the intended constructs, according to the unified validity concept of Messick (1989).

The processes of the *a priori* validation include the selection of the tests, evaluation of the selection by the experts, and content validation of the finalised tests by a pool of experts. The present study conducted a thorough evaluation of the content validation procedure using the IOC analysis, which is more comprehensive. The evidence of this *a priori* validation illustrates a clear understanding of the process of item writing, because the experts being the item writers, were able to easily agree with many of the items that measured certain underlined constructs (which are defined as objectives in IOC). The most agreed upon cognitive processes of reading, in the IOC

analysis, were easy to conceptualise by the experts. However, the items that they disagreed with, showed that the constructs, which were measured by the items, were difficult for them to conceptualize.

Although the researcher was aware that there were certain reservations about the application of the experts' judgment (Alderson & Kremmel, 2013), it is considered to be a reliable method by many scholars in the assessment field. The *a priori* evidence assured that the experts were able to easily conceptualize most of the LOT processes, such as WR, LA, and SP, while there was confusion regarding the EPM, in certain cases. At the same time, it was a little challenging to conceptualize the HOT processes. Among these processes, however, the I and the CITR were somewhat more convenient to conceptualize. (In addition, since the tests aimed at the CEFR C1 level, the CITR was not measured in many items, except for one, in all four tests.) The cognitive processes, the BMM and the CTLS, were rather challenging for the experts to theorize, as is visible from the results of the study indicated in Table 3.9. Therefore, training is prescribed to item writers to boost their ability to conceptualize the challenging cognitive processes like the EPM, BMM, and CTLS. This is necessary because in item writing, the item writers must be able to conceptualize the constructs that they evaluate, since they have to operationalize their understanding in testing.

The processes of the *a posteriori* validation were carried out using the Rasch Measurement Model, and the findings of the tests assured that the psychometric properties, namely, validity and reliability of the tests, were achieved. Further, the evidence of a continuum of increasing intensity in checking the substantial gap between the item locations, indicated that there were no noticeable gaps between item distributions, except for slight gaps at the upper and lower ends of the scale, between the locations of Items T4Q19 and T4Q34 at the top end, and Items T3Q3 and T1Q10 at the lower end, respectively (See Figure 4.3). Since there was no noticeable gap that hinders taking a defensible decision on the measuring scale, the tests were identified to be effective (Schulz, 1995).

Further, although Khalifa and Weir's (2009) socio-cognitive processes in reading are arranged according to a hierarchy (Bax, 2013; Bax & Chan, 2016; Brunfaut & McCray, 2015; Khalifa & Weir, 2009; O'Sullivan & Weir, 2011), the results of the present study did not seem to indicate the same hierarchy, except for the processes of WR and CITR (see Figure 5.1 below). Although the items for determining WR (Word Recognition) and SP (Syntactic Parsing), were the easiest to determine out of the eight cognitive processing, according to their item mean value, in some cases these processes showed higher item difficulty. This is similar to the higher cognitive processes, too. These results indicate little congruence between the empirical scaling and expert judgment, because some items (examining cognitive processes) that were assumed to be easy by the experts, turned out to be difficult for students during the empirical testing, and vice versa.

The findings also suggested that cognitive processes could be used to measure the items across different levels. Table 4.18, explained the ascending order of item logit measures of the eight cognitive processes. Many of the cognitive processes, especially, SP, EPM, I, BMM, and CTLS can be used to measure across low and high ability levels. Since item difficulty is influenced by many facets, it can be concluded that differences in individual items can influence the cognitive processing in reading, as Stanovich (1982) stated.

Overall, the tests met the RMM requirement to fulfil the construct definition. However, the ambiguous findings on the hierarchy of cognitive processing in reading, as illustrated by the evidence of the *a priori* and the *a posteriori* validations, were supported by the spread of item distribution within the cognitive processes.

5.3.5 Test Equating Procedures and Validity of Common Item Linking

One of the values of this research is the demonstration of how equal difficulty level tests are linked, using the common item linking procedure using the Rasch model analysis. In this study, 160 items distributed over four testlets were assembled to represent eight cognitive processes of reading. This study was based on a horizontal equating procedure since all four tests had the same difficulty levels. The common

item linking demonstrated that all four tests were equivalent in test difficulty level to make comparable decisions between different groups, as the test equation defined by Crocker and Algina (1986), that equating is a mathematical procedure for determining the results of several assessment instruments.

A single lengthy test would violate the validity of the test (Wells & Wollack, 2003). Therefore, all the items in the item bank were allocated into four separate tests, using the common items linking procedure of the IRT within the Rasch MM (Bond & Fox, 2015; Hambleton et al., 1991; Kolen & Brennan, 2013; Moulton, 2015; Wright & Stone, 1999; Yu & Osborn-Popp, 2005), because different tests can be equated to have the same frame of reference and measure student performance, in the same way, using proper equating procedures (Baker, 1984; Bond & Fox, 2015; Kolen & Brennan, 2013).

The challenges that psychometricians, the test constructors, or the item writers face, when utilising the linking design using concurrent analysis, is to equate test forms to determine whether any discrepancies in overall outcomes between the different populations are attributable to differences in students, tests, or both. Therefore, a careful understanding of students' group differences from test differences, would minimise such issues. Considerable variations between groups may demand a closer examination of the sampling methods or an investigation into any underlying variables affecting the entire cohort. However, tests are given to distinct cohorts of students in anchor-test design, also known as common item non-equivalent groups. In contrast to equivalent groups, the distribution of abilities in the groups can be varied. Therefore, anchoring tests using concurrent analysis can be handy, when examining different cohorts of students.

In equating different forms of tests, there is a dilemma about how to objectively evaluate different cohorts of students. This can be clearly done in the process of vertical linking, which is to measure the different difficulty levels of tests. In contrast, horizontal linking is a bit challenging. For example, among the university students from four selected cohorts (FAC, FMC, FAS, and FE), there may be students with varying ability levels, as there is a common notion that a student in the engineering field has better critical thinking skills (Almerino et al., 2020; Benigay et

al., 2018). It is important to evaluate these students' reading performance, but if the test is too difficult, the low ability students would be affected, and vice versa. When comparing the different groups of students, different separate exams (in an unsystematic approach), would result in distorted grades. For instance, should a student in one faculty who scored 95% on Test 1, be treated as same as a student in another faculty who scored 95% on Tests 2, 3, or 4? In the common item linking procedure of the Rasch MM, however, the relative difficulty of items on these tests can be measured on a single scale, without having any bias in comparison, even if the examinees have different ability levels.

One advantage of employing the concurrent analysis methodology is that it eliminates the need for equivalent groups of test-takers to provide a foundation for linking and equating the various test types. Test 1 may, for example, be administered to FAC students (different groups of students) with different sample sizes and/or ability levels, and Test 2 for FMC students (a different group). However, a comparison can be made between these two tests without FMC students sit for Test 1. The second benefit of this technique is that different tests can be interchangeably administered among different groups of students in different time frames, without showing any security bias.

Significant test differences may necessitate a closer examination of various item-level influences (e.g., scoring differences, issues in test form construction, differences in test administration), that could complicate the equating process and test score comparability. Thus, a judicious evaluation of anchoring items based on the following factors would influence the suitability of the items for use: content, item format, text type, item parameter drift, item difficulty level, differences in nearby items that could hint at the key to the anchor item, compromise the security of a test item resulting in performance changes for an anchor item, and test-takers seeing an anchor item on another test (or tests) (Kolen & Brennan, 2013; Ryan & Brockmann, 2009; Wright, 1993; Yu & Osborn-Popp, 2005).

However, certain careful processes have to be completed before finalising the concurrent design for equating. A mini-test, which consists of common items should be administered before the administration of the entire test to check the validity of the

common items. Next, the same geographical location of the common items in Test 1, for example, should appear in Tests 2, 3, and 4 in a similar spot (item number). Then, the common items should be the same in all tests, which are calibrated with no rewording, different response options, different directions, or any other change that might affect student performance from one test to the next. Finally, the same item format should be applied to all forms of the tests. The steps implemented in this study as recommended by Kolen and Brennan (2013), and Ingebo (1997), provide more strength and empirical evidence to the test developers.

Although test equating is possible with classical test theory, the IRT method is more suitable (Hambleton et al., 1991). Therefore, the Rasch MM applying IRT was utilized to link the tests among students of four different faculties, as it is “a practical and defensible method of test equating to make fair comparisons of scores from one test to another” (Wright, 1993, p. 298). Further, the Rasch provides the easiest linking procedures, like concurrent calibration or anchor test, which can be calibrated on the same interval scale using common items to equate the tests of different forms. The results achieved from the concurrent analysis or anchoring are amenable to comparing the difficulty levels of different tests (Hambleton et al., 1991).

Another reason to choose the Rasch MM to equate tests is that it can solve equating problems related to test difficulty, item difficulty distribution, sample ability, missing data, test length, standard error, linear scale, and quality control. Each item and each student completing the exams is given critical information, such as standard errors, fit statistics, and measurements separately in the Rasch MM, as they are combined, calibrated, and analysed in a single linear scale (Wright, 1993). Therefore, many issues related to equating tests are sorted out by the application of the Rasch MM, which is recognised as a versatile measurement model.

WINSTEPS, the software for conducting the Rasch analysis, can be used to do the test equating and linking in which common item equating, common person equating, virtual equating of test forms, random equivalence equating, paired item equating - parallel items, and other equating approaches can function well (Linacre, 2020). Using a horizontal equating procedure, the tests were linked to measuring the performance of the four groups of students having similar abilities (Baker, 1984;

Linacre, 2011). Using common item equating, the present study linked the four tests under a single scale by applying concurrent analysis as linking tests using common items is much more convenient rather than using common person equating (Baker & Al-Karni, 1991; León, 2008; Linacre, 2020a; Yu & Osborn-Popp, 2005).

Test specifications are another concern when equating different tests. To work properly, test specifications must be reasonably steady from one test to another. According to Kolen and Brennan (2013), only if the test specifications are well-defined and stable, equating becomes successful. The interchangeability of test scores can be brought into doubt if test forms change dramatically from one test to another. Although minor changes can be generally accommodated, the major differences in test specifications make equating more challenging. Further, they highlighted another issue related to the use of multiple item formats (selected response and constructed response), which can complicate the equating process. Therefore, the use of the MCQ was practised in the present study, as it is common in many equating tests (Ryan & Brockmann, 2009).

The number of items representing the anchoring items is also a matter of concern in test equating. The rule of thumb according to Kolen and Brennan (2013), is that the common items must be at least 20 % of the length of a total test consisting of 40 or more items, which is similar to Hambleton's (1991) suggestion. Consequently, in the present test, a set of 11 common items represent 27.5% of the total test items.

Underlined constructs that are measured by the anchoring items are in question. Out of eight cognitive processes of reading, four constructs were measured by the eleven common items, as it was determined by the experts' validation process. The number of constructs that the anchoring items should examine should be informed by empirical testing, without determining it based on mere assumption. However, since the anchoring items were based on the CEFR C1 level in the present study, the easier LOT processes like WR and SP, as well as the harder HOT process like CITR, were not considered the constructs of the anchoring items.

Finally, checking the validity of these common items is crucial in the *a posteriori* validation process, to confirm the adequacy of the anchoring items. The item polarity of the common items implies a positive value. The point measure correlation (PTMEA CORR.) for 11 items was excellent, with all of the items having positive point measure correlation coefficients and all of them were greater than 0.30 (See Table 4.8). The infit and outfit MNSQ of all items were reported to be within the expected range of 0.70 and 1.30. The raw variance explained by measures was 24.3 % in the PCA, but the residuals for the unexplained variance in the first contrast were less than 2 items in strength, which is a strong indication of measurement. The item reliability and separation indices of the common items were having good values of 0.99 and 9.24, respectively. However, the reliability and separation values for the students were low at 0.55 and 1.10 (See Table 4.7). The Wright item-person map for the eleven common items illustrated that there was no significant gap between the item location. Overall, the common items were able to meet the requirement of the Rasch MM.

Further, test equating facilitates the research study to compare the performance of different groups of students under the same scale, as Petersen et al. affirmed that it is possible to “measure examinees’ growth, to chart trends in the variable measured, and to compare or merge data, even when the separate pieces of data derive from different forms of a test with somewhat different item characteristics” (1989, p. 242).

5.3.6 Student’s Reading Performance Aligned with CEFR Level

This section provides a summary of the student’s reading performance aligned with the CEFR level along with the discussion. In addition, the student’s reading performance about their cognitive processes of the reading given in the following section (5.3.7), has a connection to answering the students’ reading performance completely.

The maximum difficulty level of the reading tests is the CEFR C1 level as determined by the study in Chapter Four. When the students’ reading performance was measured according to their degree programmes (with English as their medium of

instruction), the students from the FE outperformed those from FAS, FMC, and FAC. The majority of the students were categorized between the CEFR B1 and B2 levels indicating that 843 (93.5%) students out of a total of 902, fell under these categories. Test 4, which was given to the Faculty of Engineering students, had the best results, with 0% for the lowest level (A2), and the highest percentages (68.7%) for the highest levels (C1 and B2), compared to all four tests (Refer Tables 4.14 and 4.15). Except for the B2 and C1 levels, Test 2 and Test 3 scored similarly in almost all of the levels. The results of Test 1 showed that this test was used to determine the minimum level of performance. To name the tests according to faculties, Test 1, Test 2, Test 3, and Test 4 represented the FAC, FMC, FAS, and FE respectively.

To operationalize the theory, Khalifa and Weir (2009), employed the CEFR A2 level to C2 level Cambridge ESOL Main Suite Reading papers to check the socio-cognitive validation framework. Similarly to this, the present study tried to use the CEFR scaling to measure the performance of the students in the *a posteriori* validation procedure to provide insights into the level of the reading performance of the university students, needed for academic success.

A few previous research studies have been identified to support the claim that a minimum requirement to achieve academic success in the EMI scenario of ESL or EFL students is the B2 (Carlsen, 2018) or B1 levels (Laborda et al., 2017; Wu, 2011). In addition to this, the findings of the research administered among Taiwanese EFL students revealed that the majority of these students were at the A1 and A2 levels (Waluyo, 2019). However, the present study targeted to achieve at least a B1 level, because in the Sri Lankan context, a UTEL band score of 5, which means an upper level of B1, was prescribed as the exit level English qualification for the undergraduates who pass out from the university (Wikramanayake et al., 2012).

As per this requirement, the findings of the present study found that the majority of the test-takers, including the students of the Faculty of Arts and Culture, achieved the B1 or B2 levels. This indicates that the students have adequate ability to continue their studies in the EMI situation, and it is fair to predict that the majority of these students are at least in the average reading ability level to achieve academic success.

It is popularly believed that the students who achieve the best performance in the G.C.E. (O/L), prefer Physical Science, Biological Science, or Technology Streams for their A/L. The rest of them, generally, select Arts or Commerce Streams, although there are exceptions. Although this idea was contradictory to the findings of Rushika (2019), who mentioned that the selection of the A/L streams depends greatly on family factors, many students and parents select the STEM (Science, Technology, Engineering, and Mathematics) streams once they achieve their best performance in the O/L. In many cases, the low performers enrol in the HEMS (Humanities, Education, Management, and Social Sciences). Besides, the requirements for enrollment in the STEM fields in the A/Ls are higher compared to that of the HEMS.

Thus, one of the reasons for the best performance of the Faculty of Engineering is perhaps due to their superior critical thinking acumen, academic performance, intelligence, and skills (Almerino et al., 2020; Benigay et al., 2018), compared to the rest of the respondents of the present study (according to popular belief). The students from the Faculties of Engineering and Applied Sciences belong to the top two categories: Physical Science and Biological Science Streams in the A/Ls, whereas the Faculties of Management & Commerce and Arts & Culture students mostly followed Arts or Commerce streams in their A/L studies. As predicted by the aforementioned belief, the students from the Faculty of Engineering scored the best performance in this CEFR-aligned English reading test. In consecutive order are the Faculty of Applied Sciences, Faculty of Management and Commerce, and Faculty of Arts and Culture.

The common drawbacks confronted by Sri Lankan students for their low proficiency in the English language were highlighted in a cluster of literature (Aloysius, 2015; Attanayake, 2017; Azeera et al., 2016; Kareema, 2016; Navaz, 2016; Rameez, 2019; Rathnayake, 2013; Umashankar, 2017; Wijesekera, 2012). Students' psychological dimensions and their socio-cultural backgrounds (enrolment of students from rural backgrounds) (Rameez, 2019; Rathnayake, 2013), insufficient physical and human resources (Aloysius, 2015; Navaz, 2016; Rameez, 2019; Rathnayake, 2013), or poor teaching and learning environment and a lack of ELT professionals (Aloysius, 2015), lack of student-centred teaching methods (Azeera et al., 2016), inexperienced teachers and improper teacher training (Rameez, 2019), poor motivation for ESL

learners and teachers (Kareema, 2016; Rathnayake, 2013), lack of reading habit (Azeera et al., 2016; Rameez, 2019), unmatched curriculum (Rathnayake, 2013), and poor assessment methods (Umashankar, 2017), are a few reasons for the students' low proficiency in the English language in Sri Lanka.

Furthermore, government language policy (Coperahewa, 2009; Walisundara & Hettiarachchi, 2016), and aversion towards the English language among vernaculars (Fonseka, 2003; Gunasekera, 2005; Rathnayake, 2013; Walisundara & Hettiarachchi, 2016), are also some of the other factors indirectly influencing the English language instruction, which in turn results in the students' poor language performance. Proper measures, educational management, and policy-level changes are recommended to sort out these issues for better performance.

5.3.7 Cognitive Processing and Academic Reading

Khalifa and Weir's (2009) cognitive validity includes three important factors: metacognitive activities like goal setting and monitoring, the central processing core (includes the eight cognitive processing), and the knowledge base (Brunfaut & McCray, 2015). Cognitive validation appears to be growing in importance as a consideration in test design (Field, 2012). Thus, the findings of the study provide insightful information on the central processing core of cognitive validation. The distribution of all reading cognitive processes involved in each item on the reading exams is shown in the Wright map (See Figures 4.13 and 4.14). Among the measures of item difficulty distribution between -3.24 and 2.66 (5.9 logits span), five processes were located above the item mean of 0.00 logits and three were below this. Each process had a distinct amount of complexity. Figure 5.1 illustrates the overview of cognitive processing in reading according to Khalifa and Weir (2009) and the findings of the present study.

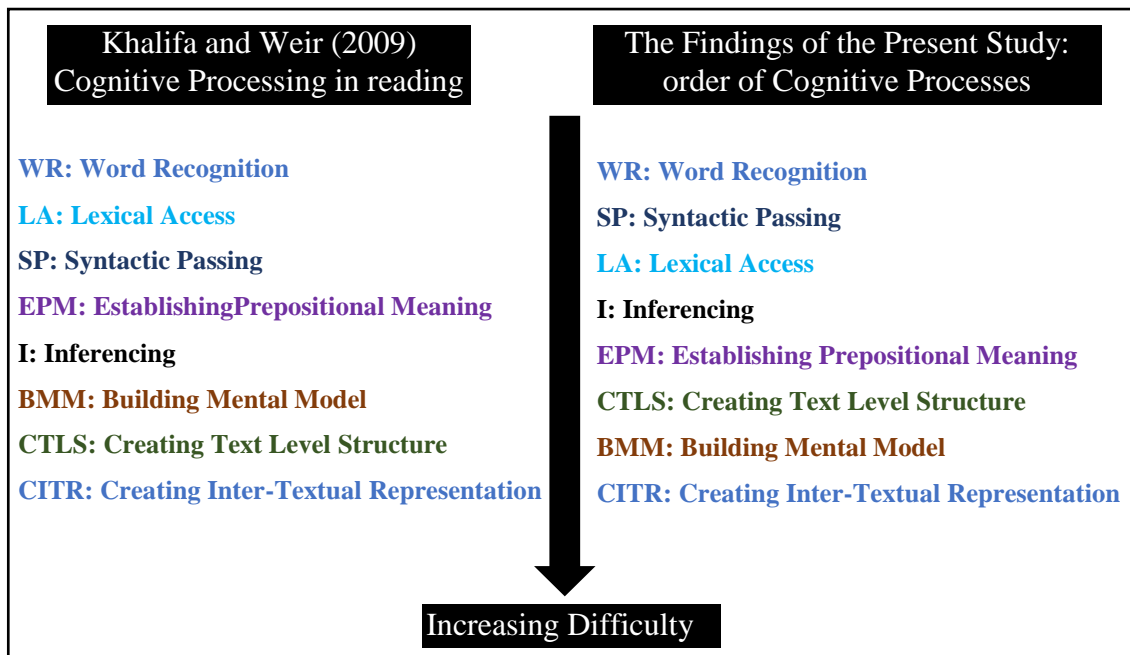


Figure 5.1 An Overview of the Cognitive Processes of Reading according to Khalifa and Weir (2009) and the Present Study

The theory claims that the unit of analysis grows larger in each processing stage, and each level indicates a 'higher' degree of processing, increasing the cognitive strain on the reader. As a result, each level is thought to be tougher than the one before it (Bax, 2013; Khalifa & Weir, 2009; Owen, 2016; O'Sullivan & Weir, 2011). As a result, WR is the easiest and CITR is the hardest according to the theory. However, the results of this study showed contradictory facts for the *a posteriori* approach. Although WR proved to be the easiest in the current study, SP, which should be ranked third, was also identified as the easiest process for scoring the same item mean value as WR. The WR and CITR were in the same order, while there was a mismatch between the other processes. However, the order of LOT and HOT processes remained the same, both in the theory and the empirical evidence.

Alfassi (2004), Green and Hawkey (2011), Moore et al. (2012), Weir et al. (2012), Chalmers and Walkinshaw (2014), Bax (2015), and Owen (2016), unanimously agree that academic reading assessments should include higher-order rather than lower-order reading processes. Higher-level cognitive processes aid in the creation of coherent, integrated, and elaborate mental representations of content

(McNamara et al., 2015); therefore, to provide settings that encourage readers to participate interestingly, more items focusing on this area should be of great use. Accordingly, in the proportions of items representing LOT and HOT processes in the present study, 79 and 48, respectively, proved to be good enough to measure cognitive processing.

A university education requires academic reading rather than general reading since the purpose of reading at the university level is reading for information and orientation. Academic reading is a kind of performance, which requires the students to act on it in some way. They are expected to discuss, evaluate, criticise, and be tested on what they have read. They need to pay great attention to the text, remember it precisely, and compare it to other texts in terms of style, structure and content. Therefore, academic reading requires high order mental processes at universities for both L1 and L2 students, and expeditious (quick) skills and methods are equally as important as they are for academic study, and in some circumstances, they are much more difficult (Weir et al., 2012). Elective search reading followed by an intensive careful reading of relevant text sections locally and globally (in terms of information processing within a text) is also expected at the university level. Putting information together across phrases to find the information, and needing to respond to the tasks, is another technique that academic reading requires among university students. Compared to WR, LA, SP, and EPM processes (LOT processes), the techniques mentioned previously depend more on high-order mental processes. The findings of the study as described in detail in sections 4.4.1 and 4.4.2, indicated that the respondents of the present study showed low achievement in the high order cognitive processes.

An interesting finding is that the items within the same process did not have the same difficulty level. For example, some of the items of the CTLS, which is one of the HOT processes, indicated low difficulty levels and vice versa. Therefore, it may be mentioned that certain cognitive processes can be used across different difficulty levels, as Stanovich (1982) stated. The facets including text variables and reader variables can have an impact on the difficulty of an item. The possible reasons for the difference in item difficulty may be due to linguistic characteristics linked with the input material, as well as variables associated with reading purpose, which can have

an impact on cognitive processing. Reader variables include formal schemata (knowledge of the language, text type, linguistic knowledge, content) or topic schemata, world knowledge, and cultural knowledge. Text variables include text length, topic, type, text difficulty, text content, text readability analysis, linguistic, non-linguistic characteristics, item format, test specification, reader purpose, etc. These two variables are crucial in defining the difficulty of an item (Alderson, 2000) because the knowledge base (monitoring) column of Khalifa and Weir's (2009) reading model explicitly connects test takers' lexical, syntactic knowledge, world knowledge, and textual aspects to cognitive functioning (Owen, 2016).

As a result, item difficulty is dependent on the complexity of the cognitive processing that is triggered by the item design, which is activated by the features of discourse highlighted in the text by the test developers. The capacity to accurately answer items is dependent on the ability of test-takers to interpret complicated multi-clause phrases, which is the core of cognitive processing (Alderson, 2000; Bachman, 1990; Koda, 2005; O'Sullivan, 2011).

Another plausible reason for the inconsistency of item difficulty of the present results and Khalifa and Weir's hierarchical order of cognitive processing is that there were big differences in the number of items that each processing tested. The WR had examined only 6 items, whereas the SP was measured in 34 items as per expert judgment at the *a priori* validation. Similarly, all eight cognitive processes were tested in a different number of items, because to have an equal number of items for each processing is impossible, as a good test must have varying levels of item difficulty. The use of an unequal number of items among cognitive processing can affect the test scores, and it will cause the analysis to have a high mean to BMM, which had been tested on 23 items. Employing an equal number of items sharing the cognitive processing would have supported the hierarchical order of Khalifa and Weir's cognitive processing.

Thus, the results of the present study are consistent with the findings of Owen (2016) who researched the academic reading tests of IELTS and TOEFL under the theoretical concern of Khalifa and Weir (2009), and concluded his findings that "The majority of identified cognitive processes were lower-level processes for both tests"

(Owen, 2016, p. 367). In both tests (which focus on the CEFR B1-C2 levels and above), high-level processing is primarily focused on building a mental model “rather than inferencing or creating a text-level representation” (Owen).

Jang (2017) emphasised that the best way to understand a learner's cognitive capacity for task execution is to look at the cognitive capacity as a dynamic system. For her, the evidence-based reasoning about the learner’s cognitive capacities is flawed if learners are assessed without considering the dynamic interaction of their multiple traits. Moreover, Weir (2005), the founder of the socio-cognitive validation framework, mentioned that merely the performance outcomes cannot adequately explain cognitive processes because, in assessment scenarios, varying task types and reading purposes place different demands on cognitive processes.

In conclusion, while the reader interacts with the text, closer observation is needed to understand the real processes occurring in each cognitive process. Thus, although the findings of the present study provided mismatching information about the hierarchy of the cognitive processes prescribed by Khalifa and Weir, closer observations of the process and characteristics of individual test-takers, and the dynamic interaction of multiple trait texts and tasks must be accompanied to measure the concrete processing involved in item design.

5.4 IMPLICATIONS

From a theoretical, methodological, and practical standpoint, the findings of this study have widespread implications.

5.4.1 Theoretical Implications

The research has contributed to a better understanding of the theoretical aspects of reading assessment, and test development and validation, particularly in terms of construct validation. Construct validation, exploring the extent to which the test performance is consistent with the theoretical expectation (Bachman, 1990; Messick,

1989) is a great concern in language testing development. The underlined reading construct applied in the study focused on Khalifa and Weir's (2009) cognitive processing validation.

As Bax (2013) highlighted for item writing of reading tests, the application of cognitive processing prescribed by Khalifa and Weir (2009) is a great concern.

Test item writers can therefore usefully draw on Khalifa and Weir (2009), for example, to plan the kinds of items they design, so as to test different levels of cognitive processing, with a view to achieving greater cognitive validity in their reading tests. (Bax, 2013, p. 461)

Researching the cognitive validity of Khalifa and Weir's reading model is significantly important: (1) as this model is built on a componential approach, and it is easy for coding purposes, (2) the model explicitly accounts for both local and global cognitive processing, (3) it has been successfully used in a diversity of contexts, and (4) it has been empirically validated in both L1 and L2 settings.

Information regarding the construct incorporated in tests is not readily available to many test developers, item writers, teachers, individual test-takers, or educational institutions. Data about how the architecture was realised in tests may be considered private or commercially sensitive. Limited information on essential abilities may be provided on publicly accessible websites, but it may be insufficient for stakeholders to make decisions regarding the usefulness of exams for certain objectives. Test specification documents especially a table of specifications including data on cognitive processes important to items, and test completion have limited information about the realisation of a construct.

Generally, the table of specifications in the previous studies has always been two-dimensional, with merely the content (items) and the process represented (level of cognition). The test developer determines item difficulty by looking at the construct and the cognitive level necessary to answer the item. This rudimentary way of judging whether an item is easy or difficult by looking at the level of cognitive complexity required of pupils to correctly answer the item should be used with caution. This study implies that while designing a reading test, test creators should think about the

cognitive processing along with the types of reading (whether, expeditious or careful reading), levels of knowledge (local or global level), item format, and explicit or implicit status of information as well. The expansion of this table of specifications has theoretical implications on test development in deciding item difficulty.

The study applied the Rasch MM, which is considered a latent trait model (Bond & Fox, 2015; Wright & Stone, 1979) to provide valid evidence. In the latent trait model, unidimensionality is a great concern. The use of this model in language testing is always criticised by Bachman (1990), mentioning that language competence is multi-component. However, it was effectively justified by Henning et al. (1985), McNamara and Knoch (2012), Mokshein (2019), and Aryadoust et al. (2020), to be applied in the construct validation of language tests. To have more comprehensive clarifications, the constructs of language components should always be verified with the theories of second language acquisition.

5.4.2 Methodical Implications

First and foremost, the study used a less-commonly used item objective congruence (IOC) rating process to validate the test materials to reach the content validation (CV) as part of the *a priori* validation. Expert judgment is measured in this approach by calculating the IOC index, which measures the agreement among experts on what objectives the test items appear to measure (the objectives refer to the cognitive demands of students). As many of the items tested multiple objectives, the IOC simplified formula presented by Crocker and Aligna (1986) for multidimensional items, was applied in the study. Although it was time-consuming, this strategy offered an alternative to the traditional qualitative method of assessing content validation, as it is more justifiable and objective, as well.

The next is a common practice at SEUSL, where different norm-referenced assessments are being used to measure language attainment at the semester-end examinations among faculties. As this assessment procedure does not allow the stakeholders to compare the students' performance levels with the levels of the other faculty students, a common framework or a benchmark is needed to measure the

performance levels of the students among different faculties. The present study, utilizing a criterion-referenced procedure applying the CEFR to measure the performance levels of different faculty students under one framework, enables the comparison of students' performance under one umbrella. Further, the application of the Rasch MM, which can be applied to both norm-referenced and criterion-referenced tests, facilitates the measurement of the ability of the students and the difficulty levels of the test items on a single scale. As the students' performance should be monitored over time, the use of Rasch MM is appropriate to monitor (Bond & Fox, 2015; McNamara, 1996).

The third methodical benefit of the research is related to the use of the output of the Rasch model, which is an item-ability map. All items on the map are placed on a scale of difficulty, and all students are located according to their ability levels. The map shows if items representing each cognitive processing meet the Rasch model's expectation and reflect the theoretical framework. It gives readers a sense of the item difficulty of a specific test item. Test developers can determine which items are easy and which are challenging for most examinees based on the results. If an item or a person is at the upper end of the continuum, it is considered a difficult item, or a high ability student, and vice versa.

The next benefit is its fit statistics, which can provide clear empirical information about the misfit and overfit items. Therefore, the items misjudged as difficult or easy by the experts, as per the hierarchy of cognitive processing, can be easily identified through the use of these fit statistics.

In addition, the Rasch MM has performed concurrent analysis on a large number of items. It is critical to have a big number of items to represent each cognitive processing category so that the results reflect the real situation more realistically. This is predicated on the idea that the more items in the testing, the more consistent the results will be. Consequently, 160 items were utilized in all four tests; however, all these tests were calibrated using 11 common items, which is an effective method according to Linacre (2020), using the concurrent analysis procedure. In most cases, examinations with a massive number of items need a significant amount of time and effort from the test takers. Nonetheless, because of concurrent analysis, without

making the students tiresome, the data were analysed concurrently among parallel tests. The concurrent analysis using common item equating procedures enables the comparison of the performance of the different cohorts of students. Further, the application of horizontal equating procedures to link the tests with the same difficulty level gives more empirical evidence for the test equating procedures.

Although many previous studies (Badrasawi et al., 2019; Boone, 2016; Davis & Boone, 2021; Jusoh, 2018), used both dichotomous and rating scale data of the Rasch MM, to have insightful information, the present study handled only multiple-choice question formats of selected response method, applying the dichotomous analysis. Although Sykes and Yen (2000), criticised the use of single analysis, as the same procedure was utilized by Zubairi and Kassim (2006), considering the accountability of data available, the present study, too, applied the dichotomous analysis. The simplicity of applying this analysis is also another great benefit of the methodology.

5.4.3 Practical Implications

Some practical consequences for various groups of individuals, including language teachers, test developers, item writers, test-takers, and policymakers, are proposed based on the findings of the study.

There is very little publicly available information on test construct description and test difficulty level that is freely accessible to stakeholders (Natova, 2019; Owen, 2016). As a result, material designers, test developers, item writers, and teachers have limited information regarding test development and validation. Item writers will get a clear idea of which reading constructs they should focus on in creating the items, on which basis the texts are selected according to the test levels, how to use the readability analysis, selection of item format and text type for a variety of reading purposes, etc. Further, content validation of the experts in the field will help the test writers shape the items and their designs.

The socio-cognitive validation framework has been applied in many common tests (Cambridge ESOL, Baltic States' test of English in Higher education), it has been applied in a variety of contexts. For example, it has been the theoretical basis for CEFR linking projects by the examination boards in the UK, Mexico, Taiwan, Turkey, and Japan, and the CRELLA of the University of Bedfordshire uses it to supply the theoretical and practical basis for professional development courses and training. However, it has been less known among ELT experts in South Asia or South East Asia. Through this research, some experts gained a short training on Khalifa and Weir's socio-cognitive validation framework, and this study can provide insightful information on item writing, the item writers and test-takers can be aware of the construct of reading.

The next practical implication to do with the construction of the test item is its quality. Item writing, which requires certain technical and procedural expertise, as well as experience from item writers, must be given special attention to developing high-quality tests. This is because item writing is a difficult undertaking that involves knowledge of both content and testing. As a result, item developers should be given enough exposure and training in the aspects, for example, item format, text type, readability index, test purpose, and test-taker characteristics, so that they can account for the effects of item characteristics on item difficulty.

This study utilized the selected response method. As there are many criticisms of the constructed response items, such as that all taught materials cannot be covered by a few items in a test; grading is inconsistent and more subjective; students with low writing abilities are underprivileged, etc., (Downing & Haladyna; 2006; Powell & Gillespie, 1990; Ventouras et al., 2010; Zeidner, 1987). Even though the selected response requires much more time to create (Powell & Gillespie, 1990), it covers most topics within a short period. Moreover, the writing speed of different students does not impact the reading performance, and the marking is consistent, fast, and cost-effective. This study further supports this claim and suggests having proper training for item development involving the construction of proper distractors as Downing (2002, p. 240) asserted that to produce valid MCQ items, "item writers must have the willingness to invest considerable time and effort into creating effective MCQs".

Another area that could benefit from the research is reading instruction. Reading is a complicated process, which makes teaching reading an even more difficult job for teachers. Because according to the study, high-level processing categories are difficult for students to learn, more attention should be placed on polishing these processes. In the classroom, there must be explicit instruction because classroom education focuses more on testing reading rather than teaching it. To make improvements, there must be a clear separation between the two features. Teachers' instruction and direction to utilize the English language outside the classroom will probably enhance the students' reading skills.

Although processes like word recognition and syntactic parsing were identified as the easiest processes, they are not always easy, and vice versa, as creating textual representation is not always a hard cognitive process as well. Moreover, these cognitive processes are equally important at different levels, although some are more important than others. However, clear instruction on reading and the cognitive processes involved in reading should be given to students. Similarly, an understanding of the purpose of reading, text types, item format, and item characteristics, which influence the item difficulty, ought to be emphasised by the teachers.

Furthermore, pre-testing or a pilot test allows test developers to examine the items and research the nature of test technique effects on item difficulty before they are given to test takers, as it was prescribed by Fulcher (1997). Although the items of the present study were developed by an experienced pool of item writers (of LRN), very few items were identified as inappropriate. Therefore, to have high-quality results in practice, piloting is recommended since it provides evidence for the *a posteriori* validation.

Despite having studied English for 11 to 13 years in schools similar to those in Malaysia (David et al., 2015), Sri Lankan undergraduates still have a long way to go in terms of language competency (Attanayake, 2017; Navaz, 2016; Rameez, 2019; Rathnayake, 2013; Walisundara & Hettiarachchi, 2016). Poor English proficiency among graduates has failed them in seeking vacancies in the job market (Dundar et al., 2017; Rameez, 2019); therefore, policymakers must take suitable measures to evaluate the student's performance in the English language, and especially their reading skills

should be monitored over time. The use of the item-person map in the Rasch MM allows policymakers, test designers, teachers, and students to have a complete view of what students have accomplished and what they still need to achieve in reading.

The common item linking procedures, and the item banking (160 items) have another practical implication to examine the performance or proficiency levels of the different cohorts of students in order to make comparable decisions. In addition, the four valid and reliable tests can be interchangeably utilized between different cohorts in the future without having security bias. Further, the banked items can also be used in the future to construct different forms of tests.

5.5 LIMITATIONS OF THE STUDY AND POINTERS FOR FURTHER RESEARCH

Several concerns were developed over the course of this research. As a result, its findings should be interpreted in the context of its limitations. These constraints may open up new opportunities for future research in the field of language evaluation in general, and reading assessment, in particular.

First, the present study focused highly on the cognitive validation of Khalifa and Weir (2009)'s socio-cognitive frameworks for reading, due to the detailed elements within the framework. Even within cognitive validation, the central processing core was studied in detail compared to metacognitive activities and the knowledge base. Although the framework has six components, this research discussed a little on test-taker characteristics, context validity, scoring validity, and consequential validity, but it did not evaluate the criterion validity in depth, the results on validating the tests are not consequential because of the absence of this component. However, to have a clear picture of validation in future research, consisting of all six components equally focused, would provide better results.

Secondly, this study employed only the objective type items of the selected response method rather than constructed responses. As Khalifa and Weir (2009) pointed out, constructed responses have more certainty that the results were due to comprehension than to any other factor like lucky guessing. However, further research

on both selected response and constructed response items, including other text variables like multiple text types, text length, passage difficulty, sentence length, and question language, can give more insightful results to the test development and validation research.

The third constraint is the unequal distribution of items examined for each cognitive process. Although the study ensured that all processes were examined in each test except for the process of CTR, it was difficult to assure an equal allocation of items for each cognitive process. The nature of the reading texts employed contributed to the uneven distribution of items. For example, only one item tested CTR out of four tests which had a mean logit of 0.33, whereas SP (having -0.55 mean logits) was measured by 34 items. Although it is challenging to create equal item distribution among the processing, employing this in future studies may provide unbiased results.

The next limitation is that this study used an uncommon IOC analysis for the quantified analysis of the expert judgment. It was difficult to obtain an agreement among some items of this study, as is usual in other studies that rely on expert judgment. Therefore, future research should focus on the items that did not achieve high agreement to see what elements of these items made them difficult for experts to agree on.

The maximum difficulty level of the tests applied in the study falls under the CEFR C1 level. When administering different tests with varying difficulty levels like the C2 level, the CPE, or multi-level tests like IELTS, TESOL, TEFOL, UTEL, or IELCA, in future tests, different findings may be achieved.

This study was a simulated test rather than a real test. Because of the high-stakes nature of the real test, and the fact that the reading portion was part of a larger test that includes grammar and writing, a real-time observation was deemed intrusive and could have influenced the validity and fairness of the test. Further, there is a claim that good readers are good writers as well as listeners. Studies examining the students' other English language skills are needed in future.

There is a need for further qualitative inquiry or analysis of the disparities in the performance of different faculty students to determine the elements that contribute to high and low performance. Thus, more detailed information about the performance of these students can be attained.

Further, it was observed that the item difficulty level was lower than the student's ability level. In this case, more difficult items are needed to be included or students having different ability levels should be targeted rather than examining the high ability students. However, despite these facts, the tests proved to be valid and reliable as per the measurement requirements of the Rasch MM.

5.6 RECOMMENDATIONS

According to the findings of the study, some recommendations are addressed. First, the Ministry of Higher Education should improve methods of assessing students' English literacy performance in nationwide universities. Although the UTEL provided this opportunity in Sri Lankan universities in 2015, it was not regularly practised in the following years, and there is a great need for implementing such tests in Sri Lanka, as the World Bank and the UGC Sri Lanka advocate and plan for it (Ministry of Higher Education and High ways, 2018).

The second recommendation is that this study validated the CEFR-aligned reading tests having equal difficulty levels, using the common item linking procedure, applying the Rasch MM. Such methods are recommended for validating nationwide English tests over the years. Since validation is a crucial factor in language testing, the findings of the research among Sri Lankan university students recommend having more validation studies to administer successful high stake tests in the country.

Further, it is necessary to investigate the elements that affect pupils' English language growth. It is recommended to help students use the English language outside the classroom, too. The research will help us better understand how students learn and may improve the efficacy of universities (faculty performance), instructors, the community, and the nation, in promoting English language competencies among

students. Although UTEL benchmarks can assist universities in keeping track of and focusing on underperforming students, due to the absence of such exams, these sorts of studies are recommended to measure the attainment of the students and their respective faculties.

It is suggested that item writers must receive the appropriate training in understanding the elements like text type, length, passage difficulty, reading purpose, item format, etc., that determine text and item difficulty in designing reading tests. Also, they must be trained in conceptualizing HOTS cognitive processes. Thus, in the future, test developers should also consider changing the table of specifications with multi-dimensional variables to have a stronger judgment of item difficulty.

Moreover, the Rasch MM is well-suited for ESL reading research, since it provides a reliable tool for researchers trying to identify the psychometric properties of tests as validity and reliability evidence, and latent variables. The shortcomings in the CTT urged the researcher to find an alternative way, and the IRT was identified as the best for this present study, and it applied the Rasch MM which is a popular model under the IRT. The concurrent analysis of the Wright item-person map empowers the stakeholders to understand the level of students' ability and the item difficulty on the same scale, which makes it easy to come to a conclusive decision on both the items and the students.

Finally, the concurrent analysis of the Rasch MM enables effective test equating procedures and helps compare the results of the different tests on the same scale. Therefore, this approach is recommended to analyse more items as the test-taker does not need to sit for all tests or answer all questions because of the fruitful use of common item linking.

5.7 CONCLUSION

Based on the findings of the study, the four tests proved to be valid and reliable. The adequacy of the reading tests was analysed through the analysis of individual tests, linking procedure, concurrent analysis, and defining the construct, as well as the analysis of the students' responses. Further, it is recommended to select more samples with multiple ability levels or to have more challenging items.

Even though many studies have been done to address the issues in reading construct, there is a lack of research in providing empirical evidence for the validation of Khalifa and Weir's (2009) cognitive processing, and to have conclusive decisions on their validation framework. As a result, this research was conducted as yet another attempt to provide empirical proof to the validation of reading tests applying Khalifa and Weir's reading model.

Although the reading model of Khalifa and Weir is considered a hierarchical model, the findings of the research indicated inconsistency in the order of cognitive processing involved to do the reading task. The content validation and empirical evidence of the research proved the componential approach to reading. Furthermore, WR and SP were identified to be the easiest items; however, this was not consistent among all individual items. Sometimes these processes were also identified as challenging in certain conditions. This is the same with the most challenging processes, too. To sum up, although the processing is hierarchical, it differs with the individual items, since item difficulty is influenced by many latent variables, like text length, sentence length, text type, item format, item language, and lexical and syntactic levels of both item and text, etc. Many of the cognitive processes can be used to measure across different difficulty levels.

However, the findings demonstrated that there is consistency in the order of low-level processing (LOT) and high-level processing (HOT). Clearly, not all low-order processes were easier than higher-order processes and vice versa, as item difficulty is influenced by item and text variables. Low-level processes, on the other hand, are generally less cognitively demanding than high-level processes. This implies that item difficulty is influenced by a variety of factors other than cognitive complexity. Therefore, item writers must be careful about all these processes and must

be given ample training on conceptualizing the processes like EPM, BMM, CTLS, and CITR.

Further, the majority of the students were categorized between the CEFR B1 and B2 levels according to the findings of this study. Only less than 1% of the total population was reported to score the A2 level, whereas about 6% of them recorded the C1 level. These results indicate that their English language proficiency is sufficient enough to cope with their English medium instruction at the university level according to the local requirement of language situations. Their achievements imply that they would successfully perform in their academic accomplishments if they can perform well in HOT processes, too.

However, students' reading performance in English varied from one group (faculty) to another. As anticipated by the researcher, the FE students scored the highest followed by FAS, FMC, and FAC students, and the findings are arranged in ranking order, according to their reading performance, respectively. Although the results were consistent, there were students in the FE who scored less than the ones who scored well in the FAC. Not all students in the FAC scored less, and not all the FE students did better. Individual items influence the item difficulty level of an entire test, individual test-taker characteristics also affect the performance of one group. Therefore, to make any conclusions about students' performance based only on faculty would be imprudent.

Several recommendations for theory, methodology, and practice were offered. There was a strong emphasis on the necessity for more research into test development involving item design, as well as the creation of multiple strategies for investigating item difficulty using the table of specifications, and validation features using the Rasch MM. To address the problems created by lexical and syntactical deficits in reading comprehension, it was suggested that more attention be paid to syntax and grammar in teaching and learning. Further, the findings of this investigation justify the necessity for better measurement for the improvement of students' performances over time.

The recommendations presented in this study are hoped to add to the scholarly discourse on test adaptation, test equating, test validation, item writing, testing reading skills in English studies, and the application of such studies to the testing of other English language skills. The findings of this investigation beckon evidence-based study would lead to improved ESL teaching and learning, and informed decision-making about students' performance, which would be the desirable outcomes of this research. There is high confidence that the findings of this study will lead to improvements in language teaching, learning, and assessment, allowing stakeholders to have more sound knowledge of English in real-life situations, while also contributing the best to the larger community.

REFERENCES

- A comparison of different readability scales.* (n.d.). Linguapress.
<https://linguapress.com/teachers/flesch-kincaid.htm>
- Abeywickrama, P. (2000). *Evaluation of the English placement test 1999 University of Colombo, Sri Lanka.*
- Abeywickrama, R. (2020). *Effectiveness and sustainability of the University Test of English Language (UTEL). April.*
- Aebersold, J. A., & Field, M. L. (1997). *From reader to reading teacher: Issues and strategies for second language classrooms.* Cambridge University Press.
- Alderson, J. C. (2000). *Assessing reading.* Cambridge Assessment English.
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment* (1st edition). Continuum.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge University Press.
- Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im) possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, 30(4), 535–556.
<https://doi.org/10.1177%2F0265532213489568>
- Alderson, J. C., & Lukmani, Y. (1989). *Cognition and reading: Cognitive levels as embodied in test questions.*
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129. <https://doi.org/10.1093/applin/14.2.115>
- Alfassi, M. (2004). Reading to learn: Effects of combined strategy instruction on high school students. *The Journal of Educational Research*, 97(4), 171–185.
<https://doi.org/10.3200/JOER.97.4.171-185>

- Alkialbi, A. S. (2015). *The Place of Reading Comprehension in Second Language Acquisition*. 6, 14–22.
- Allington, R. (2001). *What really matters for struggling readers: Designing research-based programs*. Longman.
- Almerino, P. M., Ocampo, L. A., Abellana, D. P. M., Almerino, J. G. F., Mamites, I. O., Pinili, L. C., Sitoy, R. E., Abelgas, L. J., & Peteros, E. D. (2020). Evaluating the Academic Performance of K-12 Students in the Philippines: A Standardized Evaluation Approach. *Education Research International*, 2020. <https://doi.org/10.1155/2020/8877712>
- Aloysius, M. (2015). *Problems of English teaching in Sri Lanka : how they affect teaching efficacy*. University of Bedfordshire.
- ALTE. (2002). The ALTE can do project. English version. *Framework*.
- American Marketing Association. (2001). First PCM test dates scheduled. *Marketing News*, 35, 33.
- American Psychological Association. (1985). *Standards for educational and psychological testing*. American Psychological Association.
- Anderson, N. J. (1999). *Exploring second language reading: Issues and strategies*. Heinle & Heinle.
- Andrew, A. (2017). *English Medium Instructions on English Language Proficiency*. *English Medium Instructions on English Language Proficiency*. November. <https://doi.org/10.9734/ARJASS/2017/37756>
- Andrich, D., & Godfrey, J. R. (1978). Hierarchies in the Skills of Davis' 'Reading Comprehension Test', Form D: An Empirical Investigation Using a Latent Trait Model. *Reading Research Quarterly*, 14(2), 182–200.
- Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing*, 33(4), 529–553.

- Aryadoustmcm, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 1–35. <https://doi.org/10.1177/026553222092748>
- Attanayake, A. U. (2017). *Undergraduate ELT in Sri Lanka*. Cambridge Scholars Publishing.
- Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe’s content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, 47(1), 79–86. <https://doi.org/10.1177/0748175613513808>
- Azeera, A. L. F., Nizla, M. L. F., & Kareema, M. I. F. (2016). Common drawbacks encountered in learning English language among the undergraduates of Eastern province. *6th International Symposium, 2016 on “Multidisciplinary Research for Sustainable Development in the Information Era”*, 7. <https://www.researchgate.net/profile/Fouzul-Kareema/publication/344160482>
- Babcock, B., & Hodge, K. J. (2020). Rasch Versus Classical Equating in the Context of Small Sample Sizes. *Educational and Psychological Measurement*, 80(3), 499–521.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university press.
- Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*. <https://doi.org/10.1177/026553220001700101>
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1

- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.
- Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*(1), 31–36.
- Badrasawi, K. J. I. (2012). *English reading literacy of Malaysian lower secondary students using Rasch measurement model* (Issue April) [International Islamic University Malaysia]. <http://studentrepo.iium.edu.my/handle/123456789/3793>
- Badrasawi, K., Yahefu, H., & Khalid, M. (2019). Challenges to parental involvement in children's education at a primary school: A Rasch analysis. *IIUM Journal of Educational Studies, 7*(1), 47–57. <https://doi.org/10.31436/ijes.v7i1.243>
- Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching & Research, 2*(5), 1052–1060. <https://doi.org/doi:10.4304/jltr.2.5.1052-1060>
- Baker, F. B. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement, 8*(3), 261–271. <https://doi.org/10.1177/014662168400800302>
- Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147–162. <https://doi.org/10.1111/j.1745-3984.1991.tb00350.x>
- Bannur, F. M., Abidin, S. A. Z., & Jamil, A. (2015a). A Validation Process of ESP Testing Using Weir's Socio Cognitive Framework (2005). *Procedia-Social and Behavioral Sciences, 202*, 199–208.
- Bannur, F. M., Abidin, S. A. Z., & Jamil, A. (2015b). A Validation Process of ESP Testing Using Weir's Socio Cognitive Framework (2005). *Procedia - Social and Behavioral Sciences, 202*, 199–208. <https://doi.org/10.1016/j.sbspro.2015.08.223>

- Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing*, 26(3), 349–379. <https://doi.org/10.1007/s11145-012-9372-9>
- Bax, S. (2012). *Text Inspector. Online text analysis tool*. <https://textinspector.com>
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4), 441–465. <https://doi.org/10.1177/0265532212473244>
- Bax, S., & Chan, S. H. C. (2016). Researching the cognitive validity of GEPT High-Intermediate and Advanced Reading: An eye-tracking and stimulated recall study. *LTTC-GEPT Research Reports*, 7, 1–47. www.ltcc.ntu.edu.tw/ltcc-gept-grants/RReport/RG07.pdf
- Benigay, M., Bordago, A., Carabuena, R., & Narcisco, R. (2018). *Critical Thinking Levels and Math Achievement of NORSU Senior High School Students*. https://www.researchgate.net/publication/327537816_
- Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a Framework for Constructed-Response Items*. <https://files.eric.ed.gov/fulltext/ED395032.pdf>
- Berk, R. (1984). Conducting the item analysis. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 97–143). Johns Hopkins University Press.
- Berkowitz, S., & Taylor, B. M. (1981). The effects of text type and familiarity on the nature of information recalled by readers. *Directions in Reading: Research and Instruction*, 157–161.
- Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, 25, 133–150. <https://doi.org/10.1017/S0267190505000073>
- Bickman, L., & Rog, D. J. (1998). *Handbook of Applied Social Research Methods*.

- Birch, B. M. (2007). *English L2 reading: Getting to the bottom*. Mahwah, NJ: L. Erlbaum Associates.
- Birch, B. M., & Fulop, S. (2020). *English L2 reading: Getting to the bottom*. Routledge.
- Blything, L. P., Hardie, A., & Cain, K. (2020). Question Asking During Reading Comprehension Instruction: A Corpus Study of How Question Type Influences the Linguistic Complexity of Primary School Students' Responses. *Reading Research Quarterly*, 55(3), 443–472. <https://doi.org/10.1002/rrq.279>
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model* (Third Edit). Routledge.
- Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, 15(4), 1–7. <https://doi.org/10.1187/cbe.16-04-0148>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer.
- Bornstein, R. F. (1996). *Face validity in psychological assessment: Implications for a unified model of validity*.
- Bramley, T., & Wilson, F. (2016). Maintaining test standards by expert judgement of item difficulty. *Research Matters: A Cambridge Assessment Publication*, 21, 48–54. <https://www.cambridgeassessment.org.uk/Images/374813-maintaining-test-standards-by-expert-judgement-of-item-difficulty.pdf>
- Brown, H. D. (2001). Teaching By Principles: An interactive approach to language pedagogy. In *Longman*.
- Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (Vol. 10). Pearson Education White Plains, NY.
- Brown, J. D. (2005). *Testing in language programs: a comprehensive guide to English language assesment*. McGraw-Hill College.

- Brunfaut, T., & Green, R. (2019). *TRANSFORM project English language assessment in Sri Lanka*. https://www.britishcouncil.lk/sites/default/files/elassess_output_2-report_on_current_national_english_language_assessment_in_sri_lanka_the_region_brunfautgreen_003.pdf
- Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study*. https://www.research.lancs.ac.uk/portal/files/84736502/Brunfaut_and_McCr
- Bryman, A. (2016). *Social Research Methods* (5th editio). Oxford University Press.
- Cambridge Assessment English. (2019). *Qualifications for higher education Cambridge English Qualifications for admissions purposes*. Cambridge Assessment English.
- Cambridge University Press. (2013). *Introductory Guide to the Common European Framework of Reference (CEFR) for English Language Teachers*. 11. <http://www.englishprofile.org/images/pdf/GuideToCEFR.pdf>
- Carlsen, C. H. (2018). The Adequacy of the B2 Level as University Entrance Requirement. *Language Assessment Quarterly*, 15(1), 75–89. <https://doi.org/10.1080/15434303.2017.1405962>
- Carrell, P. L., Devine, J., & Eskey, D. E. (2000). *Interactive approaches to second language reading*. Cambridge University Press.
- Carrell, P. L., Pharis, B. G., & Liberto, J. C. (1989). Metacognitive strategy training for ESL reading. *TESOL Quarterly*, 23(4), 647–678. <https://doi.org/10.2307/3587536>
- Cervetti, G. N., Bravo, M. A., Hiebert, E. H., Pearson, P. D., & Jaynes, C. A. (2009). Text genre and science content: Ease of reading, comprehension, and reader preference. *Reading Psychology*, 30(6), 487–511.

- Chalmers, H. (2019). The role of the first language in English medium instruction. *EMI Contexts and Multilingual Learners*, 1–38.
- Chalmers, J., & Walkinshaw, I. (2014). Reading strategies in IELTS tests: prevalence and impact on outcomes. *EA Journal*, 30(1), 24–39.
- Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, 20(4), 409–439. <https://doi.org/10.1191/0265532203lt266oa>
- Chapman, C. A. (1969). An Analysis of Three Theories of the Relationships Among Reading Comprehension Skills. *M*, 1–18. <https://files.eric.ed.gov/fulltext/ED043473.pdf>
- Cimmiyotti, C. B. (2013). *Impact of reading ability on academic performance at the primary level* [Dominican University of California]. <https://doi.org/10.33015/dominican.edu/2013.edu.18>
- Conrad, K. J., Conrad, K. H., Dennis, M. L., Riley, B. B., & Funk, R. (2011). *Validation of the Behavioural Complexity Scale (BCS) to the Rasch Model, GAIN Methods Report 1.2. 2012.*
- Coperahewa, S. (2009). The language planning situation in Sri Lanka. *Current Issues in Language Planning*, 10(1), 69–150.
- Council of Europe. (2001a). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. <https://rm.coe.int/1680459f97>
- Council of Europe. (2001b). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Structured overview of all CEFR scales*. <http://ebcl.eu.com/wp-content/uploads/2011/11/CEFR-all-scales-and-all-skills.pdf>
- Creswell, J. W. (2012). Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research. In *Pearson* (4th editio).

- Crocker, L. (2003). Teaching for the Test: Validity, Fairness, and Moral Action. *Educational Measurement: Issues and Practice*, 22(3), 5–11. <https://doi.org/10.1111/j.1745-3992.2003.tb00132.x>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.
- Crocker, L., Llabre, M., & Miller, M. D. (1988). The Generalizability of Content Validity Ratings. *Journal of Educational Measurement*, 25(4), 287–299. <https://doi.org/10.1111/j.1745-3984.1988.tb00309.x>
- Crocker, L. M., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, 2(2), 179–194. https://doi.org/10.1207/s15324818ame0202_6
- Cronbach, L. J. (1980). Validity on parole: How can we go straight. *New Directions for Testing and Measurement*, 5(1), 99–108.
- Curtis, D. D., & Boman, P. (2004). The identification of misfitting response patterns to, and their influences on the calibration of, attitude survey instruments. *12th International Objective Measurement Workshop, Cairns, QLD*.
- Curtis, D. D., & Boman, P. (2007). X-Ray Your Data with Rasch. *International Education Journal*, 8(2), 249–259. <https://files.eric.ed.gov/fulltext/EJ834248.pdf>
- Dabiri, A., & Kashefian-Naeeni, S. (2021). Towards a Thorough Framework in Strategic Reading Comprehension Instruction. *International Journal of Multicultural and Multireligious Understanding*, 8(12), 180–190. <https://doi.org/10.18415/ijmmu.v8i12.3268>
- David, A. R., Thang, S. M., & Azman, H. (2015). fAccommodating low proficiency ESL students' language learning needs through an online writing support system. *E-Bangi*, 12(4), 118–127. http://journalarticle.ukm.my/9355/1/118-127_LANGUAGE_LEARNING-Rowena.pdf

- Davis, D. R., & Boone, W. (2021). Using Rasch analysis to evaluate the psychometric functioning of the other-directed, lighthearted, intellectual, and whimsical (OLIW) adult playfulness scale. *International Journal of Educational Research Open*, 2, 100054.
- Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly*, 499–545.
- Davoudi, M., & Moghadam, H. R. H. (2015). Critical review of the models of reading comprehension with a focus on situation models. *International Journal of Linguistics*, 7(5), 172–187. <https://doi.org/10.5296/ijl.v7i5.8357>
- De Silva, R., & Devendra, D. (2014). Responding to English Language Needs of Undergraduates: Challenges and Constraints. *OUSL Journal*, 7(0), 1–24. <https://doi.org/10.4038/ouslj.v7i0.7305>
- DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, 15(1), 15–31.
- DeVellis, R. F. (2006). Classical test theory. In *Medical care*. JSTOR.
- Dissanayake, K. M., & Harun, R. N. S. R. (2012). Theory and practice of EAP in the Sri Lankan context. *Procedia-Social and Behavioral Sciences*, 66, 106–116. <https://doi.org/10.1016/j.sbspro.2012.11.252>
- Dissanayake, S. (2018). An analysis on reading skills of first year Engineering undergraduates: Skimming, Scanning and Vocabulary. *International Conference on the Humanities (ICH 2018/2019)*. <https://www.researchgate.net/publication/352283166>
- Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, 77(10), S103–S104.
- Downing, S. M., & Haladyna, T. (2006). *Handbook of Test Developing*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

- Dundar, H., Millot, B., Riboud, M., Shojo, M., Goyal, S., & Raju, D. (2017). *Sri Lanka education sector assessment: Achievements, challenges, and policy options*. World Bank Publications.
- Dunlea, J. (2015). *Validating a set of Japanese EFL proficiency tests: demonstrating locally designed tests meet international standard* [University of Bedfordshire Centre]. <http://hdl.handle.net/10547/618581>
- Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*(3), 515.
- Ebel, R. L., & Frisbie, D. A. (1972). *Essentials of educational measurement*. Prentice-Hall.
- Ebibi, J. O. (2014). The influence of text type on senior secondary students' reading comprehension. *IOSR Journal of Research & Method in Education, 4*(3), 1–5.
- Elder, C. (1998). What counts as bias in language testing. *Melbourne Papers in Language Testing, 7*(1), 1–42.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (2011). *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy*. Springer.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Lawrence Erlbaum Associates.
- Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework*. Educational Testing Service.
- Famularo, L. (2007). *The effect of response format and test taking strategies on item difficulty: a comparison of stem-equivalent multiple-choice and constructed-response test items* [Boston College]. <http://search.proquest.com/%0Adocview/304897354?accountid=14645> ProQuest Dissertations & Theses Full Text database. (3283877 Ph.D.),.

- Farhadi, H., & Hessami, G. H. R. (2005). Construct validity of L2 reading comprehension skills. *Iranian Journal of Applied Linguistics*, 8(2), 29–53. <https://www.sid.ir/en/Journal/ViewPaper.aspx?ID=61857>
- Farthing, D. W., Jones, D. M., & McPhee, D. (1998). Permutational multiple-choice questions: an objective and efficient alternative to essay-type examination questions. *ACM SIGCSE Bulletin*, 30(3), 81–85.
- Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. *IELTS Collected Papers*, 2, 391–453.
- Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, 66(4), 477–485.
- Fonseka, E. A. G. (2003). Sri Lankan English: Exploding the fallacy. *9th International Conference on Sri Lankan Studies, Matara*, 28–30. https://www.academia.edu/15261843/SRI_LANKAN_ENGLISH_EXPLODING_THE_FALLACY
- Fraenkel, J. R., Wallen, E. N., & Hyun, H. H. (2012). *How to design and evaluate research in education* (Eighth). McGraw Hill Companies.
- Fraenkel, J. R., & Wallen, N. (2006). E.(2006). In *How to design and evaluate research in education*.
- Freeman, F. S. (1962). *Theory and practice of psychological testing*. (3rd editio). Henry Holt.
- Fries, C. C. (1963). *Linguistics and reading* (Vol. 1). Holt Rinehart and Winston.
- Fulcher, G. (1997). An english language placement test: Issues in reliability and validity. *Language Testing*, 14(2), 113–139. <https://doi.org/10.1177/026553229701400201>
- Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*. <https://doi.org/10.1191/0265532203lt265oa>

- Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly: An International Journal*, 1(4), 253–266. https://doi.org/10.1207/s15434311laq0104_4
- Fulcher, G. (2010). Practical language testing. In *Practical Language Testing*. Hodder Education. <https://doi.org/10.4324/9780203767399>
- Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment*. Routledge.
- Gay, L., & Airasian, P. (2000). *Educational research: An introduction*. Upper Saddle River, NJ: Merrill.
- Geranpayeh, A. (2013). Scoring validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (Vol. 35, pp. 242–272).
- Ghaicha, A. (2016). Theoretical Framework for Educational Assessment: A Synoptic Review. *Journal of Education and Practice*, 7(24), 212–231. <https://files.eric.ed.gov/fulltext/EJ1112912.pdf>
- Giri, R. A. (2005). *The adaptation of language testing models to national testing of school graduates in Nepal: Processes, problems and emerging issues*. Victoria University of Technology.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1), 6–10. <https://doi.org/10.1177/074193258600700104>
- Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, 25(3), 375–406. <https://doi.org/10.2307/3586977>
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

- Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (Second Edi). Routledge.
- Granger, C. (2008). Rasch Analysis is important to understand and use for measurement. *Rasch Measurement Transactions*, 21(3), 1122-3. <https://www.rasch.org/rmt/rmt213d.htm>
- Granger, C., & Linacre, J. M. (2008). *What is measurement?* <https://www.yumpu.com/en/document/read/28306738/what-is-measurement-udsmr>
- Green, A., & Hawkey, R. (2011). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing*, 29(1), 109–129. <https://doi.org/10.1177/0265532211413445>
- Guerrero, M. D. (2000). The unified validity of the four skills exam: applying Messick's framework. *Language Testing*, 17(4), 397–421.
- Gunasekera, M. (2005). *The postcolonial identity of Sri Lankan English/Manique Gunesequera*. Katha Publishers.
- Gunawardana, A. A., & Karunarathna, J. A. M. B. (2017). International Symposium of Sabaragamuwa University of Sri Lanka (ICSUSL) – 2017 3. *International Symposium of Sabaragamuwa University of Sri Lanka (ICSUSL) – 2017, May*, 5–8.
- Hafiz, F. M., & Tudor, I. (1989). h. *ELT Journal*, 43(1), 4–13.
- Halek, M., Holle, D., & Bartholomeyczik, S. (2017). Development and evaluation of the content validity, practicability and feasibility of the Innovative dementia-oriented Assessment system for challenging behaviour in residents with dementia. *BMC Health Services Research*, 17(1). <https://doi.org/10.1186/s12913-017-2469-8>

- Hall, D. A., Zaragoza Domingo, S., Hamdache, L. Z., Manchaiah, V., Thammaiah, S., Evans, C., Wong, L. L. N., & NETwork, I. C. of R. A. and Tin. R. (2018). A good practice guide for translating and adapting hearing-related questionnaires for different languages and cultures. *International Journal of Audiology*, 57(3), 161–175. <https://doi.org/10.1080/14992027.2017.1393565>
- Hambleton, R. K. (1996). Guidelines for Adapting Educational and Psychological Tests. *Annual Meeting of the National Council on Measurement in Education*, 47. <https://files.eric.ed.gov/fulltext/ED399291.pdf>
- Hambleton, R. K., & Bollwark, J. (1991). *Adapting Tests for Use in Different Cultures: Technical Issues and Methods*. ERIC. <https://eric.ed.gov/?id=ED337481>
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.
- Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, 62(2), 143–157.
- Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, 32(3), 317–336. <https://doi.org/10.1177/0265532214564505>
- Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation*, 31(2–3), 145–161. <https://doi.org/10.1002/ase.1325>

- Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment*, 7(3), 238–247. <https://doi.org/10.1037/1040-3590.7.3.238>
- Hedgcock, J. S., & Ferris, D. R. (2009). *Teaching readers of English: students, texts, and contexts*. Routledge.
- Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, 2(2), 141–154. <https://doi.org/10.1177/026553228500200203>
- Hermida, J. (2009). the Importance of Teaching Academic Reading Skills. *The International Journal of Research and Review*, 3(Sept. 2009), 20–30, 3, 20–30. <http://julianhermida.com/teachingshowcasereading.pdf>
- Hickson, S., Reed, W. R., & Sander, N. (2012). Estimating the effect on grades of using multiple-choice versus constructive-response questions: Data from the classroom. *Educational Assessment*, 17(4), 200–213.
- Hidri, S. (2020). The IELCA and IELTS Exams: A Benchmark Report. *The Journal of Asia TEFL*, 17(2), 742–749. <https://doi.org/10.18823/asiatefl.2020.17.2.33.742>
- Hillocks, G., & Ludlow, L. H. (1984). A Taxonomy of Skills in Reading and Interpreting Fiction. *American Educational Research Journal*, 21(1), 7–24. <https://doi.org/10.3102/00028312021001007>
- Hoang, N. M. (2016). *The relationship between reading strategy use and reading proficiency of Vietnamese students in the UK* [Northumbria University, UK]. https://www.teachingenglish.org.uk/sites/teacheng/files/dissertation_design_for_publication_2016_northumbria_university.pdf
- Hollingworth, L., Beard, J. J., & Proctor, T. P. (2007). An investigation of item type in a standards-based assessment. *Practical Assessment, Research, and Evaluation*, 12(1), 18. <https://doi.org/doi.org/10.7275/6ggz-8837>

- Hooley, D. S., Tysseling, L. A., & Ray, B. (2013). Trapped in a cycle of low expectations: An exploration of high school seniors' perspectives about academic reading. *The High School Journal*, 96(4), 321–338. <https://doi.org/10.1353/hsj.2013.0018>
- Hoover, W. A., & Tunmer, W. E. (1993). *The components of reading*.
- Hosenfeld, C. (1977). A preliminary investigation of the reading strategies of successful and unsuccessful second language learners. *System*, 5(2), 110–123.
- Howard, P. J., Gorzycki, M., Desa, G., & Allen, D. D. (2018). Academic Reading: Comparing Students' and Faculty Perceptions of Its Value, Practice, and Pedagogy. *Journal of College Reading and Learning*, 48(3), 189–209. <https://doi.org/10.1080/10790195.2018.1472942>
- Huber, J. A. (2004). A closer look at SQ3R. *Reading Improvement*, 41(2), 108–113.
- Hubley, N. J. (2012). Assessing reading. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyhoff (Eds.), *The Cambridge guide to second language assessment* (pp. 211–217). Cambridge University Press.
- Hudson, T. (2007). *Teaching second language reading*. Oxford University Press Oxford.
- Hughes, A. (1989). *Testing for Language Teachers*. Cambridge University Press.
- Huy, N. Van, & Hamid, M. O. (2015). *Educational policy borrowing in a globalized world A case study of Common European Vietnamese University*. <https://doi.org/10.1108/ETPC-02-2015-0014>
- ielts.org. (n.d.). *IELTS and the CEFR*. <https://www.ielts.org/-/media/pdfs/ielts-and-the-cefr.ashx>
- Ingebo, G. S. (1997). *Probability in the Measure of Achievement*. Chicago: MESA.

- Isa, M., Bakar, M. A., Sipan, I., Hasim, M. S., Hashim, A. E., & Jalil, M. K. A. (2016). Measuring Instrument Constructs of Return Factors for Green Office Building Investments Variables Using Rasch Measurement Model. *MATEC Web of Conferences*, 1–9. <https://doi.org/10.1051/mateconf/20166600125>
- Jang, E. E. (2017). Cognitive aspects of language assessment. *Language Testing and Assessment*, 163–177. https://doi.org/10.1007/978-3-319-02326-7_11-1
- Jayasinghe, V. U., & Wijethunge, M. T. N. (2015). Study of English Language Skills of the First Year KDU Undergraduates with Special Reference to Faculty of Management, Social Sciences and Humanities. *Proceedings of 8th International Research Conference, KDU*, 80–83. <http://ir.kdu.ac.lk/bitstream/handle/345/1414/msh-015.pdf?sequence=1&isAllowed=y>
- Jiang, X. (2011). The Role of First Language Literacy and Second Language Proficiency in Second Language Reading Comprehension. *Reading Matrix: An International Online Journal*, 11(2), 177–190. <http://search.ebscohost.com/login.aspx?direct=true%5C&db=eric%5C&AN=EJ955195%5C&lang=ja%5C&site=ehost-live>
http://www.readingmatrix.com/articles/april%5C_2011/jiang.pdf%5Cnpapers3://publication/uuid/5574418D-A280-4706-8613-079D93D902D4
- Jiménez-Muñoz, A. J. (2014). Measuring the impact of CLIL on language skills: a CEFRbased approach for Higher Education. *Language Value*, 6, 28–50. <https://doi.org/10.6035/languagev.2014.6.4>
- Johns, J. L., & McNamara, L. P. (1980). The SQ3R study technique: A forgotten research target. *Journal of Reading*, 23(8), 705–708.
- Jusoh, Z. (2018). *A Rasch analysis of reading skill across text type and item format* (Issue August) [International Islamic University Malaysia]. studentrepo.iium.edu.my

- Kahraman, H. (2019). Reading as a Single Construct: A Process-Oriented Study. *Novitas-ROYAL*, 13(2), 206–220. <https://files.eric.ed.gov/fulltext/EJ1231980.pdf>
- Kamhi, A. G. (2007). Knowledge Deficits: The True Crisis in Education. *The ASHA Leader*, 12(7). <https://leader.pubs.asha.org/doi/full/10.1044/leader.FMP.12072007.28>
- Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3–17. <https://doi.org/10.1177/0265532211417210d>
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198–211. <https://doi.org/10.1080/0969594X.2015.1060192>
- Kareema, M. I. . (2013). *Patterns of Spelling Errors among the Second Language Learners of English*. University of Kelaniya.
- Kareema, M. I. F. (2016). Motivation for learning English among the students of South Eastern University of Sri Lanka. *6th International Symposium, 2016 on “Multidisciplinary Research for Sustainable Development in the Information Era”*, 7. <https://core.ac.uk/download/pdf/143655324.pdf>
- Karlin, O., & Karlin, S. (2018). Making Better Tests with the Rasch Measurement Model. *InSight: A Journal of Scholarly Teaching*, 13, 76–100. <https://files.eric.ed.gov/fulltext/EJ1184946.pdf>
- Kastner, M., & Stangl, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia - Social and Behavioral Sciences*, 12, 263–273. <https://doi.org/10.1016/j.sbspro.2011.02.035>
- Katalayi, G. B., & Sivasubramaniam, S. (2013). Careful reading versus expeditious reading: Investigating the construct validity of a multiple-choice reading test. *Theory and Practice in Language Studies*, 3(6), 877–884. <https://doi.org/10.4304/tpls.3.6.877-884>

- Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement*, 37(1), 39–57.
- Kedzierski, M. (2016). English as a medium of instruction in East Asia's higher education sector: a critical realist Cultural Political Economy analysis of underlying logics. *Comparative Education*, 52(3), 375–391. <https://doi.org/10.1080/03050068.2016.1185269>
- Keeves, J. P. & Alagumalai, S. (1999). New approaches to measurement. In G. N. Masters and J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 23–42). Pergamon.
- Keeves, J. P. (1990). *Educational research, methodology, and measurement: An international handbook*. Pergamon Press.
- Kektsidou, N., & Tsagari, D. (2019). Using DIALANG to track English language learners. *Papers in Language Testing and Assessment*, 8(1), 1–30. <https://www.researchgate.net/profile/Dina-Tsagari-2/publication/338165653>
- Kelly, T. L. (1927). *Interpretation of Educational Measurements*. Macmillan.
- Khalifa, H., & Weir, C. J. (2008). A cognitive processing approach towards defining reading comprehension. *Cambridge Esol*, 31, 2–36.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading: research and practice in assessing second language reading*. Cambridge University Press.
- Kim, H. (1996). Assessing attainment of Bloom's cognitive levels using testlets and multi categorical IRT. *ERA-AARE Joint Conference*.
- Kim, Y.-S. G. (2020). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology*, 112(4), 667–684. <https://doi.org/10.1037/edu0000407>

- Kirkpatrick. (2011). *2011 hong kong Internationalization or Englishization*. Centre for Governance and Citizenship, The Hong Kong Institute of Education.
- Klare, G. R. (1984). Readability. In *Handbook of reading research* (pp. 681–744).
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19(2), 193–220. <https://doi.org/10.1191/0265532202lt227oa>
- Kobayashi, M. (2009). *Hitting the mark: How can text organisation and response format affect reading test performance?* (Vol. 13). Peter Lang.
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.
- Kolen, M. J., & Brennan, R. L. (2013). *Test equating, scaling, and linking: methods and practices*. New York: Springer.
- Kor, L. K., & Teoh, S. H. (2009). *From literature review to developing a conceptual framework and to journal writing*. McGraw Hill.
- Krabbe, P. F. M. (2017). Classical Test Theory. In *The Measurement of Health and Health Status* (pp. 153–170). Elsevier. <https://doi.org/10.1016/B978-0-12-801504-9.00009-X>
- Krishnan, D. K. S. (2011). Careful versus expeditious reading: the case of the IELTS reading test. *Academic Research International*, 1(3), 25–35. [http://www.savap.org.pk/journals/ARInt./Vol.1\(3\)/2011\(1.3-03\).pdf](http://www.savap.org.pk/journals/ARInt./Vol.1(3)/2011(1.3-03).pdf)
- Kulasingham, R., Ilangakoon, S. ., Suraweera, D. P., Attanayake, A. M. A. ., & Ravindran, K. G. . (2012). *The Application of UTEL (A) Benchmarks Constructive Alignment for Course Design*. ELTU, University of Colombo, Sri Lanka. <https://arts.cmb.ac.lk/delt/wp-content/uploads/2018/08/BM-final-min.pdf>
- Kunnan, A. J. (2013). Approaches to validation in language assessment. In *Validation in language assessment* (pp. 15–30). Routledge.

- Laborda, J. G., Pizarro, M. A., Litzler, M. F., Esteban, S. G., & Otero, de J. N. (2017). *Student perceptions of the CEFR levels and the impact of guided practice on APTIS oral test performance*. British Council.
- Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. CA: Sage publications. <https://doi.org/10.4135/9781412963947>
- Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests* (Vol. 14). Cambridge university press.
- Learning Resource Network. (2015). *Qualification Specification- International English Language Competency Assessment (IELCA)*.
- León, A. B. (2008). Common-Item (or Common Person) Equating with Different Test Discriminations. *Rasch Measurement Transactions*, 22(3), 1172. <https://www.rasch.org/rmt/rmt223d.htm>
- Li, H., & Wilhelm, K. H. (2008). Exploring Pedagogical Reasoning: Reading Strategy Instruction From Two Teachers' Perspectives. *The Reading Matrix*, 8(1), 96–110.
- Li, S., & Munby, H. (1996). Metacognitive strategies in second language academic reading: A qualitative investigation. *English for Specific Purposes*, 15(3), 199–216. [https://doi.org/10.1016/0889-4906\(96\)00004-x](https://doi.org/10.1016/0889-4906(96)00004-x)
- Linacre, J. M. (2003). Constructing scientific measurement models. *Rasch Measurement Transactions*, 17(1), 907.
- Linacre, J. M. (2011). *Practical Rasch Measurement - Further Topics. Tutorial 4. Test Equating*. <https://doi.org/10.16194/j.cnki.31-1059/g4.2011.07.016>
- Linacre, J. M. (2018). *Detecting multidimensionality in Rasch data using Winsteps Table 23*. <https://www.youtube.com/watch?v=sna19QemE50>
- Linacre, J. M. (2020a). *A User's Guide to W I N S T E P S® M I N I S T E P Rasch-Model Computer Programs Program Manual 4.7.1*. winsteps.com.

- Linacre, J. M. (2020b). *Sample Size and Item Calibration [or Person Measure] Stability*. <https://www.rasch.org/rmt/rmt74m.htm>
- Linacre, J. M. (2020c). *Winsteps® (Version 4.7.1)* (Version 4.7.1; p. 763). Beaverton. www.winsteps.com/index.htm
- Linacre, J. M. (2021). *Fit diagnosis: infit outfit mean-square standardized*. <https://www.winsteps.com/winman/misfitdiagnosis.htm>
- Lunzer, E., Waite, M., & Dolan, T. (1979). Comprehension and comprehension tests. In E. Lunzer & K. Garner (Eds.), *The effective use of reading*. Heinemann Educational.
- Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching*, 51(1), 36–76. <https://doi.org/10.1017/S0261444817000350>
- Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, 1(1), 1–11. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1426043
- Mann, V. A. (1984). Reading skill and language skill. *Developmental Review*, 4(1), 1–15. [https://doi.org/10.1016/0273-2297\(84\)90014-5](https://doi.org/10.1016/0273-2297(84)90014-5)
- Martuza, V. R. (1977). *Applying norm-referenced and criterion-referenced measurement in education*. Allyn and Bacon.
- Masoumi, G. A., & Sadeghi, K. (2020). Impact of test format on vocabulary test performance of EFL learners: the role of gender. *Language Testing in Asia*, 10(1), 1–13.
- Matthews-López, J. L. (2003). *Best practices and technical issues in cross-lingual, cross-cultural assessments: An evaluation of a test adaptation*. Ohio University. https://etd.ohiolink.edu/apexprod/rws_etd/send_file/send?accession=ohiou1082054025&disposition=inline

- Mckee, S. (2012). Reading Comprehension, What We Know: A Review of Research 1995 to 2011. *Language Testing in Asia*, 2(1), 45–58. <https://doi.org/10.1186/2229-0443-2-1-45>
- McNamara, D. S., Jacovina, M. E., & Allen, L. K. (2015). Higher order thinking in comprehension. In *Handbook of Individual Differences in Reading* (pp. 182–194). Routledge.
- McNamara, T. F. (1996). *Measuring second language performance*. Longman Publishing Group.
- McNamara, T. F. (2006). Validity in Language Testing: The Challenge of Sam Messick ' s Legacy. *Language Assessment Quarterly*, 3(1), 31–51. https://doi.org/110.1207/s15434311laq0301_3
- McNamara, T., & Knoch, U. (2012). The Rasch Wars: The Emergence of Rasch Measurement in Language Testing. *Language Testing*, 29(4), 555–576. <https://doi.org/10.1177/0265532211430367c>
- Medina, A. L., & Pilonieta, P. (2006). *Once upon a Time: Comprehending Narrative Text*.
- Merenda, P. F. (2006). An overview of adapting educational and psychological assessment instruments: past and present. *Psychological Reports*, 99(2), 307–314. <https://doi.org/10.2466/pr0.99.2.307-314>
- Mermelstein, A. D. (2015). Reading Level Placement and Assessment for ESL/EFL Learners: The Reading Level Measurement Method. *ORTESOL Journal*, 32, 44–55.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35(11), 1012–1027.
- Messick, S. (1987). Validity. *ETS Research Report Series*, 2, i–208.

- Messick, S. (1989). Validity. *Educational Measurement (3rd Edition)*, R. L., 13–104.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241–256. <https://doi.org/10.1177/026553229601300302>
- Mikulecky, B. S. (2008). Teaching reading in a second language. *Teaching Reading in a Second Language*, 6. <https://www.researchgate.net/profile/Shahzad-Karim-2/publication/276247709>
- Miles, D. A. (2017). A taxonomy of research gaps: Identifying and defining the seven research gaps. *Doctoral Student Workshop: Finding Research Gaps-Research Methods and Strategies, Dallas, Texas*.
- Ministry of Higher Education and High ways. (2018). *AHEAD ELTA-ELSE DP Guidelines for Proposal Submission Faculty DP*. <https://esn.ac.lk/ots/downloads/ELTA-ELSE-Faculty-DP-guideliens.pdf>
- Ministry of Higher Education and Highways and World Bank. (2018). *AHEAD – RIC Grants: Guidelines for Proposal Submission*. <https://ahead.lk/wp-content/uploads/2018/05/AHEAD-RIC-Guidelines.pdf>
- Mokshein, S. E., Ishak, H., & Ahmad, H. (2019). The use of Rasch measurement model in English testing. *Jurnal Cakrawala Pendidikan*, 38(1), 16–32. <https://doi.org/doi: 10.21831/cp.v38i1.22750>
- Moore, T., Morton, J., & Price, S. (2012). Construct validity in the IELTS academic reading test: A comparison of reading requirements in IELTS test items and in university study. *IELTS Collected Papers, 2: Research in Reading and Listening Assessment*, 2, 120–150. www.ielts.org
- Moulton, M. (2015). *One ruler, many tests: a primer on test equating*. EDS Publications. http://www.eddata.com/resources/publications/EDS_APEC_Equating_Moulton.pdf.

- Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications*, 7(1), 1–12.
<https://doi.org/10.1186/s40488-020-00108-7>
- Mumin, Z. (2011). *Studies in Language Testing 29: Examining Reading: Research and Practice in Assessing Second Language Reading by Hana Khalifa and Cyril Weir*. 23(2), 225–230.
- Munby, J. (1978). *1978: Communicative syllabus design*. Cambridge: Cambridge University Press.
- Munby, John. (1968). *Read and think: training in intensive reading skills*. Longman.
- Mundrake, G. A. (2000). The evolution of assessment, testing, and evaluation. In J. Rucker (Ed.), *Assessment in business education* (Vol. 38). NBEA Yearbook.
- National Education Commission. (2016). Proposals for a National Policy Framework in General Education in Sri Lanka. In *National Education Commission*.
- National Reading Panel. (2000). *Report of the National Reading Panel*. National Institute of Child Health and Human Development.
- Natova, I. (2019). Estimating CEFR reading comprehension text complexity. *The Language Learning Journal*, 1–12.
<https://doi.org/https://doi.org/10.1080/09571736.2019.1665088>
- Navaz, M. M. A. (2016). Challenges faced by students in English medium undergraduate classes: an experience of a young university in Sri Lankaw. *Researchers World: Journal of Arts, Science & Commerce*, 7(4), 158–166.
[https://doi.org/10.18843/rwjasc/v7i4\(1\)/19](https://doi.org/10.18843/rwjasc/v7i4(1)/19)
- NETS, Department of Examination, S. L. (2016). *G . C . E .(O . L .) Examination - 2014 Evaluation Report 31 - English Language*. Department of Examinations, Sri Lanka.

- Neumann, K., Fischer, H. E., & Boone, W. J. (2014). *Quantitative Research Designs and Approaches*. https://www.researchgate.net/profile/Hans-Fischer/publication/266323766_Quantitative_Research_Designs_and_Approaches/links/560f971208ae6b29b49a5b1d/Quantitative-Research-Designs-and-Approaches.pdf
- Nixon, C., & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal*, 957–971. <https://doi.org/https://doi.org/10.2307/1061503>.
- North, B. (2000). Linking language assessments: An example in a low stakes context. *System*, 28(4), 555–577. [https://doi.org/10.1016/S0346-251X\(00\)00038-5](https://doi.org/10.1016/S0346-251X(00)00038-5)
- North, B. (2014a). Putting the common European framework of reference to good use. *Language Teaching*, 47(2), 228–249. <https://doi.org/10.1017/S0261444811000206>
- North, B. (2014b). *The CEFR in practice* (Vol. 4). Cambridge University Press.
- North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15(2), 217–262.
- Nuttall, C. (1985). Survey Reviews: Recent materials for the teaching of reading. *English Language Teaching Journal*, 39.
- Nuttall, C. (1996). *Teaching reading skills in a foreign language* (New Editio). Heinemann.
- O'Neill, S. (2009). Implementing NETPAW's diagnostic test of English proficiency in Australia: a case study. *Proceedings of the 13th International Conference on Multimedia Language Education (ROCMELIA 2009)*, 122–137. https://eprints.usq.edu.au/7188/2/O%27Neill_ROCMELIA_Keynote__2009_AV.pdf

- O'Sullivan, B. (2011). *Language Testing: Theories and Practices*. Palgrave Macmillan.
- O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: theories and practices* (pp. 13–32). Palgrave Macmillan.
- Ohata, K., & Fukao, A. (2014). L2 learners' conceptions of academic reading and themselves as academic readers. *System*, 42, 81–92. <https://doi.org/10.1016/j.system.2013.11.003>
- Olshavsky, J. E. (1977). Reading as problem solving: An investigation of strategies. *Reading Research Quarterly*, 12(4), 654–674. <https://doi.org/https://doi.org/10.2307/747446>
- Osterfind, S. J. (1997). *Constructing test items: Multiple-choice, constructed-response, performance and other formats*. Kluwer Academic Publishers.
- Owen, N. I. (2016). *An evidence-centred approach to Reverse Engineering: Comparative analysis of IELTS and TOEFL iBT reading sections*. University of Leicester.
- Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7th edition). Routledge.
- Parakrama, A., Navaz, A. M. M., & Rassool, R. (2021). English Language Teaching: A Historical Present. In P. de S. et Al. (Ed.), *Beyond Boundaries: One Hundred Years of Humanities and Social Sciences in Sri Lankan Universities - Volume I: Humanities* (p. 462). University Grants Commission, Sri Lanka.
- Paris, S. G., Wasik, B., & Turner, J. C. (1991). The development of strategic readers. *Handbook of Reading Research*, 2, 609–640.
- Paul, R., & Elder, L. (2008). *How to read a paragraph: The art of close reading*. CA:Foundation for Critical Thinking.

- Pearson, P., & Johnson, D. (1978). *Teaching reading comprehension*. New York: Holt, Rinehart & Winston.
- Pengruck, L., Boonphak, K., & Sisan, B. (2019). Early childhood education: A confirmatory factor analysis concerning thai administrators' creative administration. *Asia-Pacific Social Science Review*, 19(1), 17–32. <http://apsr.com/wp-content/uploads/2019/03/RA-2.pdf>
- Perfetti, C. A. (1985). Acquisition of Reading Skills. In *Acquisition of Reading Skills*. Routledge.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). *Scaling, norming, and equating*.
- Popham, W. J. (2000). Assessing mastery of wish-list content standards. *NASSP Bulletin*, 84(620), 30–36. <https://doi.org/10.1177/019263650008462004>
- Powell, J. L., & Gillespie, C. (1990). *Assessment: All Tests Are Not Created Equally*. <https://files.eric.ed.gov/fulltext/ED328908.pdf>
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3), 513–536. <https://doi.org/10.1111/1467-9922.00193>
- Rameez, A. (2019). English language proficiency and employability of university students: a sociological study of undergraduates at the faculty of arts and culture, South Eastern University of Sri Lanka (SEUSL). *International Journal of English Linguistics*, 9(2), 199–209. <https://doi.org/10.5539/ijel.v9n2p199>
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Mesa Press.
- Rathnayake, P. N. (2013). Clearing impediments to the use of English by the undergraduates: A Case study of the faculty of Humanities and Social Sciences at the university of Ruhuna, Sri Lanka. *Procedia-Social and Behavioral Sciences*, 93, 70–76. <https://doi.org/10.1016/j.sbspro.2013.09.154>

- Ratwatte, H. V. (2001, April). Issues and Concerns in Mass testing - University Test of English (UTEL). Paper presented at the UGC workshop on University Test of English. Colombo. April, SLFI No Title. *UGC, Sri Lanka*.
- Ratwatte, H. V. (2016). Employment as a reason to achieve fluency in English? The beliefs of Secondary School learners and undergraduate students of 21st century Sri Lanka. *VISTAS Journal of Humanities & Social Sciences*, 10, 102–143. http://repository.ou.ac.lk/bitstream/handle/94ousl/1527/Paper_5.pdf?sequence=1
- Rieben, L., & Perfetti, C. A. (2013). *Learning to read: Basic research and its implications*. Routledge.
- Roberts, D. F., Hornby, M. C., Hernandez-Ramos, P., & Bachen, C. M. (1984). Reading and television. *Communication Research*, 11(1), 9–49. <https://doi.org/10.1177/009365084011001002>
- Robinson, F., P. (1941). *Effective study*. Harper & Brothers Publishers.
- Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1155.
- Rogers, W. T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, 59(2), 234–247.
- Rogier, D. (2012). The effects of English-Medium Instruction on language proficiency of students enrolled in higher education in the UAE. *Thesis Doctoral PQDT-UK & Ireland*, U621250, <https://ore.exeter.ac.uk/repository/bitstream/handle/10036/4482/RogierD.pdf?sequence=2%0Ahttp://search.proquest.com/docview/1654741914?accountid=14542> <http://dn3nh3eq7d.search.serialssolutions.com/?genre=article&sid=ProQ:&atitle=The+effects+of+english-med>

- Rosenshine, B. V. (2017). Skill hierarchies in reading comprehension. In *Theoretical issues in reading comprehension* (pp. 535–554). Routledge.
- Rotfeld, H. (1998). Are we teachers or job trainers. *Academy of Marketing Science Quarterly*, 2(2).
- Rovinelli, R. J., & Hambleton, R. K. (1977). On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity. *Dutch Journal of Educational Research*, 2(49–60), 49–60.
- Rumelhart, D. E. (1977). Toward an interactive model of reading. In *Attention and performance* (pp. 573–603). Academy Press.
- Rusika, T. (2019). The factors influence on GCE A/L stream selection: a sociological research based on Mullaitivu district. *Proceedings of 9th International Symposium, South Eastern University of Sri Lanka*, 1032–1041. [http://192.248.66.13/bitstream/123456789/4087/1/Final Proceedings - Page 1051-1060.pdf](http://192.248.66.13/bitstream/123456789/4087/1/Final%20Proceedings%20-%20Page%201051-1060.pdf)
- Ryan, J., & Brockmann, F. (2009). A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory. In *Council of Chief State School Officers*. ERIC. <https://files.eric.ed.gov/fulltext/ED544690.pdf>
- Samson, D. M. M. (1983). *Rasch and reading*. In van Weeren, J., editor, *Practice and problems in language testing*, Arnhem: S. CITO.
- Saunders, B. M. (2007). *(Post) Colonial Language : English, Sinhala, and Tamil in Sri Lanka*. International Journal of Research in Engineering, IT and Social Sciences. <http://homes.chass.utoronto.ca/~cpercyc/courses/eng6365-saunders.htm>
- Schulz E. M. (1995). Construct Deficiency? *Rasch Measurement Transactions*, 9(3), 447.

- Senaratne, C. D. W. (2013). Are online assessment schemes of English skills successful?: A comparative analysis of the UTEL national assessment scheme and the Pre-orientation program (POP). *International Conference on Social Sciences 2013, University of Kelaniya, Sri Lanka*. <http://repository.kln.ac.lk/handle/123456789/2522>
- Sengupta, S. (2002). Developing academic reading at tertiary level: A longitudinal study tracing conceptual change. *The Reading Matrix*, 2(1), 1–37. <http://www.readingmatrix.com/articles/sengupta/article.pdf>
- Shaibah, H. S., & van der Vleuten, C. P. M. (2013). The validity of multiple choice practical examinations as an alternative to traditional free response examination formats in gross anatomy. *Anatomical Sciences Education*, 6(3), 149–156. <https://doi.org/10.1002/ase.1325>
- Sharma, S. R. (2000). *Modern teaching strategies*. Omsons Publications.
- Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge University Press.
- Shen, M.-Y. (2013). Toward an Understanding of Technical University EFL Learners' Academic Reading Difficulties, Strategies, and Needs. *Electronic Journal of Foreign Language Teaching*, 10(1), 70–79.
- Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1(2), 147–170.
- Skaggs, G., & Lissitz, R. W. (1986). IRT Test Equating: Relevant Issues and a Review of Recent Research. *Review of Educational Research*, 56(4), 495–529.
- Smith, E. V. J. (2001). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *Journal of Applied Measurement*, 2(3), 281–311. <https://www.researchgate.net/publication/11359121>
- Smith, R. M. (2003). *Rasch measurement models: Interpreting WINSTEPS and FACETS output*. Mesa Press.

- Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.
<https://litmedmod.ca/sites/default/files/pdf/snow-rand-mr1465.pdf>
- Snowling, M., Cain, K., Nation, K., & Oakhill, J. (2010). *Reading comprehension: nature, assessment and teaching*. Centre for Reading and Language.
<https://eprints.lancs.ac.uk/id/eprint/50134/1/ESRCcomprehensionbooklet.pdf>
- Spearitt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly*, 8(1), 92–111.
- Stanovich, K. E. (1982). Individual differences in the cognitive processes of reading: II. Text-level processes. *Journal of Learning Disabilities*, 15(9), 549–554.
<https://doi.org/10.1177/002221948201500908>
- Stauffer, R. G. (1967). Reading as a cognitive process. *Elementary English*, 44(4), 342–348. <https://www.jstor.org/stable/41386162>
- Suen, H. K. (2012). *Principles of test theories* (Second). Routledge.
- Sykes, R. C., & Yen, W. M. (2000). The Scaling of Mixed-Item-Format Tests With the One-Parameter and Two-Parameter Partial Credit Models. *Journal of Educational Measurement*, 37(3), 221–244. <https://doi.org/10.1111/j.1745-3984.2000.tb01084.x>
- Takwin, M., Pansri, O., Parnichparinchai, T., & Vibulrangson, S. (2018). Developing a Self-Assessment Instrument for Analysis of the Social and Personal Competencies of Teachers in Senior High Schools in Indonesia. *Journal of Physics: Conference Series*. <https://doi.org/10.1088/1742-6596/1028/1/012088>
- Taylor, G. (2009). *A student's writing guide: How to plan and write successful essays*. Cambridge University Press.
- Taylor, L. (2011). *Examining speaking: Research and practice in assessing second language speaking* (Vol. 30). Cambridge University Press.

- Thaidan, R. (2015). Washback in Language Testing. *Education Journal*.
<https://doi.org/10.11648/j.edu.20150401.12>
- Thamburaj, K. P., Sivanadhan, I., & Kumar, M. (2021). Improving Form 4 Student's Reading Comprehension Skills in Tamil Language By Using SQ3R Method. *Psychology and Education Journal*, 58(2), 2291–2295.
<https://doi.org/https://doi.org/10.17762/pae.v58i2.2394>
- Tian, G. S. (1991). Higher order reading comprehension skills in literature learning and teaching at the lower secondary school level in Singapore. *RELC Journal*, 22(2), 29–43.
- Traynor, R. (1985). The TOEFL: an appraisal. *ELT Journal*, 39(1), 43–47.
<https://doi.org/10.1093/elt/39.1.43>
- Tsai, T.-H., Hanson, B. A., Kolen, M. J., & Forsyth, and R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education*, 14(1), 17–30.
- Turner, R. C., & Carlson, L. (2003). Indexes of Item-Objective Congruence for Multidimensional Items. *International Journal of Testing*, 3(2), 163–171.
https://doi.org/10.1207/s15327574ijt0302_5
- UCAS. (2014). *International Qualifications: For Entry to University or College in 2015*. UCAS.
- Ulrich, T. (2021). *The Influence of the Foreign Service Institute on US Language Education: Critical Analysis of Historical Documentation*. Routledge.
- Umashankar, S. (2017). *Title Name Washback effects of speaking assessment of teaching English in Sri Lankan schools Singanayagam Umashankar*. August 2017 UNIVERSITY OF BEDFORDSHIRE.

- University Grants Commission. (2020a). Admission To Undergraduate Courses of the Universities in Sri Lanka. In *UGC Sri Lanka*.
- University Grants Commission. (2020b). *Fortyeth Annual report - 2018*.
- Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics*, *14*(2), 60–82. https://www.researchgate.net/profile/Michelle-Raquel-2/publication/292138200_
- Urquhart, A. H., & Weir, C. J. (1998). *Reading in a Second Language: Process, Product and Practice* (G.N. Candlin (Ed.)). Longman.
- Vacca, R T, & Vacca, J. L. (2008). *Content area reading: Literacy and learning across the curriculum*. Boston, MA: Pearson Education Group.
- Vacca, Richard T. (2002). From efficient decoders to strategic readers. *Educational Leadership*, *60*(3), 6–11. <https://www.researchgate.net/profile/Richard-Vacca/publication/296870969>
- van Steensel, R., Oostdam, R., & Van Gelderen, A. (2013). Assessing reading comprehension in adolescent low achievers: Subskills identification and task specificity. *Language Testing*, *30*(1), 3–21. <https://doi.org/10.1177%2F0265532212440950>
- Veeravagu, J., Muthusamy, C., Marimuthu, R., & Michael, A. S. (2010). Using Bloom's taxonomy to gauge students' reading comprehension performance. *Canadian Social Science*, *6*(3), 205–212. <https://doi.org/10.3968/j.css.1923669720100603.023>
- Ventouras, E., Triantis, D., Tsiakas, P., & Stergiopoulos, C. (2010). Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers. *Computers & Education*, *54*(2), 455–461.

- Wade-Woolley, L. (1999). First language influences on second language word reading: All roads lead to Rome. *Language Learning*, 49(3), 447–471. <https://doi.org/10.1111/0023-8333.00096>
- Walisundara, D. C., & Hettiarachchi, S. (2016). *English Language Policy and Planning in Sri Lanka : A Critical Overview. November 2015.*
- Waluyo, B. (2019). Thai First-Year University Students ' English Proficiency on CEFR Levels : A Case Study of Walailak University , Thailand. *The New English Teacher*, 13(2), 51–71. <http://www.assumptionjournal.au.edu/index.php/newEnglishTeacher/article/view/3651>
- Wang, W.-C., & Chen, C.-T. (2005). Item Parameter Recovery, Standard Error Estimates, and Fit Statistics of the Winsteps Program for the Family of Rasch Models. *Educational and Psychological Measurement*, 65(3), 376–404.
- Wazeema, T. M. F., & Kareema, M. I. F. (2017). Implication of multimedia audio-visual aids in the English language classroom. *Proceedings of 7th International Symposium, SEUSL, 7th & 8th December 2017*, 433–442. <http://192.248.66.13/handle/123456789/3028>
- Weaver, C. A., & Kintsch, W. (1991). *Expository text.*(In R. Barr, et. Al.: eds). *Handbook of reading Research. Vol. 2.* New York: Longman.
- Weir, C., Hawkey, R., Green, A., & Devi, S. (2012). The cognitive processes underlying the academic reading construct as measured by IELTS. *IELTS Collected Papers, 2: Research in Reading and Listening Assessment*, 2, 157–189. www.ielts.org
- Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. In *IELTS Research Reports Volume 9.* British Council and IELTS Australia. https://www.researchgate.net/publication/228904849_3

- Weir, C., Huizhong, Y., & Yan, J. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge University Press.
- Weir, C. J. (2005). *Language testing and validation: An Evidence-Based Approach*. Palgrave MacMillan.
- Weir, C. J., Hughes, A., & Porter, D. (1990). Reading skills: Hierarchies, implicational relationships and identifiability. *Reading in a Foreign Language*, 7(1), 505–510. https://scholarspace.manoa.hawaii.edu/bitstream/10125/67034/7_1_10125_67034_rf171weir.pdf
- Weir, C. J., & Porter, D. (1994). The multi-divisible or unitary nature of reading: The language tester between Scylla and Charybdis. *Reading in a Foreign Language*, 10(2), 1–19. https://scholarspace.manoa.hawaii.edu/bitstream/10125/66948/10_2_10125_66948_rf1102weir-sm.pdf
- Weiss, B., Gridling, G., Trödhandl, C., & Elmenreich, W. (2006). Embedded systems exams with true/false questions: A case study. *Real-Time Systems*, 182, 1.
- Wells, C. S., & Wollack, J. A. (2003). An instructor's guide to understanding test reliability. *Testing & Evaluation Services. University of Wisconsin*, 7. <https://testing.wisc.edu/Reliability.pdf>
- Wesche, M. B. (1987). Second language performance testing: The Ontario test of ESL as an example. *Language Testing*, 4(1), 28–47. <https://doi.org/10.1177/026553228700400103>
- Widdowson, H. G. (1979). *Explorations in applied linguistics* (Vol. 1). Oxford University Press, USA.
- Wijesekera, H. D. (2012). Dreams Deferred: English Language Teaching in Sri Lanka. *VISTAS Journal of Humanities & Social Sciences*, 7/8, 16–26. <https://doi.org/10.2139/ssrn.2438893>

- Wikramanayake, G. N., Karunartna, D. D., & Wettewe, D. S. (2012). Evaluation of English and Information Technology Skills of New Entrants to Sri Lankan Universities. *International Journal of Information and Education Technology*, 2(2), 171–174. <https://doi.org/10.7763/IJiet.2012.V2.103>
- Williams, E., & Moran, C. (1989). Reading in a foreign language at intermediate and advanced levels with particular reference to English. *Language Teaching*, 22(4), 217–228.
- Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, 45(3), 197–210. <https://doi.org/10.1177/0748175612440286>
- Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal*, 77(4), 473–489.
- Wolfe, E. W., & Smith, E. V. J. (2007). Instrument development tools and activities for measure validation using Rasch models: part II--validation activities. *Journal of Applied Measurement*, 8(2), 204–234.
- Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29(1), 23–48. <https://doi.org/10.1177/001316446902900102>
- Wright, B. D., & Stone, M. H. (1979). *Best Test design*. Mesa Press.
- Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. Wilmington, DE: Wide Range.
- Wright, B. D. (1978). The rasch model for test construction and person measurement. *Quinta Conferencia y Exhibición Anual de Medición y Evaluación, Universidad de Chicago*.

- Wright, B. D. (1993). Equitable test equating. *Rasch Measurement Transactions*, 7(2), 298–299.
- Wright, B D, & Linacre, J. M. (2001). Glossary of Rasch measurement terminology. *Rasch Measurement Transactions*, 15(2), 824–825.
- Wright, Benjamin D, & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Wright, Benjamin D, & Masters, G. N. (1982). *Rating scale analysis*. MESA press.
- Wright, Benjamin D, & Stone, M. H. (2004). *Making Measures*.
- Wu, J., & Wu, R. (2007). Using the CEFR in Taiwan: The perspective of a local examination board. *The Language Training and Testing Center Annual Report*, 56, 1–20.
- Wu, R. Y.-F. (2011). *Establishing the validity of the General English Proficiency Test reading component through a critical evaluation on alignment with the Common European Framework of Reference*. November. <http://uobrep.openrepository.com/uobrep/handle/10547/223000>
- www.lrnglobal.org. (n.d.-a). *Qualification Specification - LRN Level 2*. http://www.lrnglobal.org/web_docs/qualifications/esolint/C1
- www.lrnglobal.org. (n.d.-b). *Who we are?*
- Yu, C. H., & Osborn-Popp, S. E. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment, Research, and Evaluation*, 10(1), 4. <https://doi.org/10.7275/68dy-z131>
- Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: the student's perspective. *The Journal of Educational Research*, 80(6), 352–358.
- Zhang, S., & Duke, N. K. (2008). Strategies for Internet reading with different reading purposes: A descriptive study of twelve good Internet readers. *Journal of Literacy Research*, 40(1), 128–162. <https://doi.org/10.1080/10862960802070491>

Zubairi, A. M. (2001). *Reliability and validity in placement testing with reference to the English placement test at the International Islamic University Malaysia*. University of Surrey.

Zubairi, A. M., & Kassim, N. L. A. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research*, 2(1), 1–20. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.535.2955&rep=rep1&type=pdf>

APPENDIX A

FOUR READING TEST PAPERS

1. TEST 1 (with key)

South Eastern University of Sri Lanka
CEFR Multi-level Test of English Reading Skill

TEST 1

Reading 1

For questions 1-10, choose the best answer (A, B or C) and fill in the gaps.

ICE-CREAM

Most of us enjoy a cold ice-cream (1) a hot summer day, but have you ever (2) where ice-cream comes from? It is believed that ice-cream was invented (3) the Chinese. (4)was originally made by freezing the ingredients in a mixture of ice and salt and continued to be produced in this way until the (5) of the freezer in the 20th century.

Ice-cream (6) was made from milk was first eaten in Italy and later it was introduced to England in the 17th century. At that time, ice-cream had to be eaten immediately as there (7) no way to store it. It was during the late 19th century that ice-cream could be stored and (8) in great quantities. Since then, ice-cream has (9)one of the most popular desserts. Now it is available in (10) flavours and colours.

- | | | |
|-----------------------|---------------------|------------------|
| 1. <u>A. on</u> | B. in | C. at |
| 2. <u>A. wondered</u> | B. wondering | C. wonder |
| 3. A. from | B. of | C. <u>by</u> |
| 4. A. He | B. <u>It</u> | C. What |
| 5. A. inventing | B. <u>invention</u> | C. inventor |
| 6. A. what | B. whose | C. <u>which</u> |
| 7. A. is | B. <u>was</u> | C. will be |
| 8. A. <u>sold</u> | B. cleared | C. done |
| 9. A. became | B. becoming | C. <u>become</u> |
| 10. A. acceptable | B. accurate | C. <u>many</u> |

Key: 1.A 2.A 3.C 4.B 5.B 6.C 7.B 8.A 9.C 10.C

Reading 2

Read the text below about Jet Lag and then do the exercises that follow.
For questions 11-19, choose the best answer (A, B or C).

Jet Lag

Anyone who has been on a long-haul flight will be only too aware of the effect that jet lag has on their body rhythms. You get jet lag when you fly east or west across several time zones. There is now some evidence that flying westwards across the world causes less jet lag than flying eastwards. Some say that flying during the day also causes fewer symptoms. Adjustment to new time zones undoubtedly upsets the body rhythms and causes difficulties for the body's internal clock as it tries to compensate.

Symptoms include headaches, insomnia, disturbed appetite, upset stomach and lack of concentration. These symptoms are most noticeable if you are older and the farther you fly. Body temperature, heart rate and hormone rhythms can also be affected. When you reach your destination, you may find that you are unable to sleep even if you are exhausted. So, you are sleepy during the day and wide awake at night. However, there are some precautions and advice you can take if you are flying across the world.

Firstly, try to adapt to the new time zone in advance by getting up at a different time and adjusting your working day and mealtimes. For some days before you fly, make sure that you get plenty of exercise. Also, try not to be too stressed out, worried or even excited about your trip. If you get a cold, it might be worth postponing your trip for a few days because your condition will **deteriorate** on the plane. Last but not least, make sure that you get a good night's sleep on the night before your trip.

There are lots of things you can do once you are on the plane. At the start of a long flight, set your watch to the time zone you are travelling to and try to adjust accordingly by sleeping, resting and eating at the appropriate times. Drink plenty of water during the flight to stop dehydration caused by flying. Don't drink too much alcohol as this will only worsen the situation. Get as much exercise as you can. Walk up and down the aisle and do exercises while you are sitting down. When it is time to go to sleep on the flight, make yourself really comfortable. Wear ear plugs to shut out any noise and take off your shoes.

When your trip is finished and you eventually arrive home, be prepared for the same problems. However, you can make things easier by using the same guidelines you had followed before going away. It also might be a good idea to stay at home for 24 hours, resting and exercising and slowly **adjusting** to life and conditions back in your own country.

11. What is mentioned in the first paragraph about jet lag?

- A. It is better flying eastwards than westwards.
- B. Its effects may depend on what time of the day you travel.
- C. It depends on your personal body rhythms.

12. What is **TRUE** about the symptoms?

- A. They are not so bad if you fly farther.
- B. They include difficulty in concentrating.
- C. They cause you to have a bigger appetite.

13. One thing before a trip that will **NOT** make jet lag worse is
 A. stress. B. a cold. C. plenty of exercise.
14. The writer advises travellers to put off their trip for a few days if they
 A. are exhausted. B. have gone down with a cold. C. are too anxious about travelling.
15. Which of the following options can best replace the word **deteriorate** in the third paragraph?
 A. improve B. worsen C. relieve
16. One thing travellers shouldn't do while flying is
 A. drink lots of beer. B. drink plenty of water. C. set their watch to the new time zone.
17. During the flight, travellers are advised to
 A. exercise only in their seats. B. reduce noise levels while sleeping.
 C. sleep with their shoes on.
18. Once travellers arrive at their destination, it is advisable they should
 A. not go out of the house for a whole day. B. try to get used to things immediately.
 C. ignore the guidelines they followed before the trip.
19. What does the word **adjusting** in the last paragraph mean?
 A. confirming B. adapting C. accepting

Key: 11. B 12.B 13.C 14.B 15.B 16.A 17.B 18.A 19.B

Reading 3

Read the following two passages about Holidays.
 For questions **20-30**, choose the best answer (**A, B or C**).

Holidays

All-Inclusive Holidays

The main reason for the popularity of all-inclusive holidays is that they are very convenient and stress-free. Holidaymakers, whether they go on cheap all-inclusive holidays or on luxurious ones, pay in advance for holiday expenses such as transportation, accommodation and meals so they do not have to worry about overspending during their holidays. Nor do they need to be **concerned about** planning activities as the resorts they go to offer a variety of entertainment for all ages on site, which is included in the price. Package holidays also give you the chance to socialise with other holidaymakers.

These kinds of holidays do not come without drawbacks, however. For one thing, holidaymakers tend to spend their time on the resort grounds and so their interaction with local culture is either limited or non-existent. Those who do decide to

explore the culture and the sights of the area will pay extra for such activities. Moreover, the food served may not always be agreeable to all and the all-you-can-eat buffet can encourage guests to eat more than they should. Sometimes a package holiday simply fails to live up to expectations.

20. One reason all-inclusive holidays are popular is that they
- A. organise many excursions. B. are hassle-free. C. offer free meals.
21. What does the phrase ‘**concerned about**’ in the 1st paragraph mean?
- A. thrilled about B. doubtful about C. troubled about
22. What is **TRUE** about holidaymakers on an all-inclusive holiday?
- A. They hardly get to see the area around the resort.
B. They are not charged for activities outside the resort.
C. They often participate in local cultural events.
23. A holidaymaker on an all-inclusive holiday may be dissatisfied with
- A. the food provided. B. the small portions of the food served.
C. the luxury the resort offers.

Travelling Independently

The most important benefit of travelling independently is that you have the total freedom and flexibility to do whatever you please, whenever you want. As an independent traveller, you also benefit the local economy since you are not confined to the premises of one single resort. More important than anything else, however, is the magic that you feel when you travel independently- the sense of accomplishment from getting from place to place and the growth that occurs when you stretch your comfort zone.

However, organising your holiday on your own is not without pitfalls. First of all, looking for the best accommodation and transportation for your budget might be time-consuming and nerve-racking. You may also face an unpleasant surprise as the hotel you booked may not always be what the website claimed it to be. What is more, the cost of travelling independently can be much higher than that of an organised package holiday. One of the most serious disadvantages, though, is that as an independent traveller, you are not always safe. Not knowing the area, you may unexpectedly find yourself in a dangerous location.

24. What is **FALSE** about travelling independently?
- A. It allows travellers to do what they desire.
B. It is suggested primarily for long holidays.
C. It doesn’t oblige travellers to remain at their hotel.
25. What does the word **accomplishment**’ in the 1st paragraph mean?
- A. embarrassment B. achievement C. amusement

26. Organising your holiday on your own
- A. may cause you considerable anxiety.
 - B. will not usually take you much time.
 - C. is generally easier than you think.
27. One of the most serious pitfalls of travelling independently is that it may
- A. be as expensive as a package holiday.
 - B. put the traveller's safety at risk.
 - C. require online booking.

Questions 28-30 refer to BOTH Sections A and B.

28. In which passage(s) is inclusive entertainment mentioned?
- A. Passage A
 - B. Passage B
 - C. Passage A and B
29. The holidays in both passages
- A. are best suited for young people.
 - B. require you to book your own accommodation.
 - C. may lead to disappointment.

30. In which passage(s) is limited contact with the local community mentioned?
- A. Passage A
 - B. Passage B
 - C. Passage A and B

Key: 20.B 21.C 22.A 23.A 24.B 25.B 26.A 27.B 28.A
29.C 30.A

Reading 4

Read the 4 passages below and answer the questions that follow.

A

Ofcom, the UK government communications regulator, says one in three adults and most teenagers classify themselves as highly addicted to their smartphones such as iPhones, Blackberrys and Androids and in many cases are understood to be our 'closest companion'. Britons' appetite for Facebook and social networks on the go is driving a huge demand for smartphones – with 60% of teenagers describing themselves as "obsessed with" their device – according to new research by Ofcom. Almost half of teenagers and more than a quarter of adults now own a smartphone, with most using their iPhone or BlackBerry to browse Facebook and email. The study also shows that smartphones have begun to intrude on our most private moments, with 47% of teenagers being reluctant to own up to taking a sneak peek at their 'best friend' when retiring for the night. Only 22% of adults confessed to the same habit. Unsurprisingly, mobile-addicted teens are more likely than adults to be distracted by their phones over dinner and in the cinema – and a further number of this age bracket would answer their phone if it woke them up. Separate figures shared exclusively with the Guardian newspaper show that, for the first time, smartphone sales outstripped sales of regular mobiles in the first half of this year as the enormous demand continues to rise. Just

over half of the total 10.6m mobile sales from January to June 2011 were smartphones, according to research by GfK Retail and Technology UK. Of the new generation of smartphone users, 60% of teenagers classified themselves as "highly addicted" to their device, compared to 07% of adults. Ofcom surveyed 2,070 adults and 521 children and teenagers in March 2011. The regulator defines teenagers as aged between 12 and 15, with adults 16-years-old and above.

"Ofcom's report shows the influence that communication technology now has on our daily lives, and also on the way we behave and communicate with each other," said James Thickett, director of research for Ofcom. "Our research into the use of smartphones, in particular, reveals how quickly people become reliant on new technology – to the point of feeling addicted. As more and more people acquire smartphones, they are becoming an essential tool in peoples' social lives whether they are out with friends socialising or using Facebook on the move."

B

Facebook remains far and away the most frequented website for mobile users, with users spending almost four times the amount of time socialising online than using any other social networking website or browser – ie. Yahoo, Google, Ask Jeeves. Unsurprisingly, multi-tasking teenagers said they were less likely to read books if they owned a smartphone but they also said that owning a smartphone made them more likely to ditch games consoles like the PS0 and the computer, in favour of their pocket-sized handset." The research is saying that people are keeping their phones on longer and becoming addicted to them. This isn't a problem now but something we need to be aware of. Operators have responded by upgrading their networks so it is being coped with," Thickett said.

C

Despite being a nation of mobile addicts, Ofcom found that truisms still apply when it comes to more old fashioned, traditional media like TV and radio. Some of their other findings show that an increased amount of viewers are spending more time in front of the TV (at least four hours a day last year, compared to 0.8 hours in 2009). This is partly due to the rise of on demand viewing, most notably Sky+ where past programmes can be re-viewed within a set time window, and an increase in the number of homes with high-resolution TVs (HD) for substantially clearer viewing. Two newcomers to the HD market, Freeview HD and Freesat HD, have established themselves as more-affordable competitors to Virgin Media and Sky.

D

With regard to broadband, the new generation of broadband, enabling fast delivery through sophisticated fibre-optic cables, is now available for 57% of UK households – of which over 50% have adopted. Just over one in 10 said they browse the web via their games console, while 9% use it to watch BBC iPlayer. Finally, Britons sent an average of five text messages a day last year, contributing to a total of 129bn texts sent – up by 24% in 2009. However, Ofcom have warned that older Britons risk being left behind in the "digital revolution". While 90% of adults aged 05-44 have the internet at home, this falls to just a quarter of over 75s. Ofcom said that, for the first time, more than half of 65 to 74 year-olds have access to the internet at home, while just over three quarters own a mobile phone.

Questions 31-34

Choose the correct title for each paragraph A-D from the list below. Write the correct number i – iv.

List of Titles

- i. Ignoring some other forms of technology (B)
- ii. The need for smartphones and user behaviour (A)
- iii. Numbers of households that use technology (D)
- iv. Higher Definition recruits more audiences (C)

31. Paragraph A ___ ii _____
32. Paragraph B ___ i _____
33. Paragraph C ___ iv _____
34. Paragraph D ___ iii _____

Questions 35-40

Do the following statements agree with the view presented in the passage above?

Write TRUE if the statement is in agreement.

Write FALSE if the statement is not in agreement.

Write NOT GIVEN if the statement does not represent a view expressed in the passage.

Note questions are not necessarily in the same order as the text.

35. Over half of juveniles who own a smartphone confess to suffering a smartphone addiction. (T)
36. The demand for Google is higher than that for Facebook. (F)
37. Over 50% of 80-year-olds have broadband at home. (F)
38. TV and radio are becoming more popular. (NG)
39. Adults were too embarrassed to admit to using their smartphone while in bed. (F)
40. Technology has captured everyone's life without any age gap and its influence is visible over half of the amount of household devices is the best fitting summary of the passage. (T)

2. TEST 2 (with key)

South Eastern University of Sri Lanka
CEFR Aligned Test of English Reading Skills
TEST 2

Reading 1

For questions 1-10, choose the best answer (A, B or C) to fill in the gaps.
Mark your answers on the separate Answer Sheet.

Educational Programmes For Adults

A lot of institutions are (01) adult educational programmes nowadays. Most of the adult educational programmes are part-time, evening or summer courses and (02)..... designed to satisfy the needs and (03) of the students. The reasons why an adult may participate (04)..... such programmes vary. (05)..... example, some adults may attend a programme because they want to get a better position in their job. Some (06)..... may do it to get another degree in a different field. There are, of course, those who have (07)..... had any formal education and want to pursue a career. Since the 1990s, the number of participants in adult educational programmes (08)..... rapidly because the job market has become (09)..... competitive. Also, as technology is constantly changing, it is necessary for all kinds of workers to (10)..... the information and the skills required.

- | | | | |
|-----|--------------------|---------------------|-------------------------|
| 01. | A. <u>offering</u> | B. provided | C. given |
| 02. | A. has | B. can | C. <u>are</u> |
| 03. | A. benefits | B. <u>interests</u> | C. eagerness |
| 04. | A. on | B. <u>in</u> | C. at |
| 05. | A. Such an | B. As an | C. <u>For</u> |
| 06. | A. another | B. the others | C. <u>others</u> |
| 07. | A. always | B. <u>never</u> | C. just |
| 08. | A. increasing | B. was increased | C. <u>has increased</u> |
| 09. | A. <u>more</u> | B. much | C. a lot of |
| 10. | A. respond | B. ask | C. <u>have</u> |

Key: 1. A 2. C 3. B 4. B 5. C 6. C 7. B 8. C 9. A 10. C

Reading 2

For questions 11-19, choose the best answer (A, B or C).
Mark your answers on the separate Answer Sheet.

The Tradition of Coffee Drinking

Coffee drinking is an important part of daily life in many countries of the world. People rely on a cup of this **delicious** liquid to wake them up in the morning, and coffee shops provide important social centres in both cities and rural villages. Made from the bean of the coffee plant, coffee is a true gift of nature and its popularity has led to the growth of a global industry.

16. According to the text, which of the following statements is **TRUE**?
- A. Modern coffee shops may serve cakes and pastries with coffee.
 - B. Serving cakes and pastries with coffee is a modern custom.
 - C. Coffee served with cakes and pastries costs more.
17. The people who grow coffee
- A. make a lot of money.
 - B. are protected by “Fair-trade” organisations.
 - C. live in rich countries.
18. According to the text, which of the following statements is **FALSE**?
- A. Coffee shops in Europe offer a variety of coffee drinks.
 - B. Espresso is the only popular choice of coffee in Europe.
 - C. Cappuccino can be drunk with hot milk or cream.
19. What does the phrase “**catch up on**” in the last paragraph mean?
- A. take in
 - B. look for
 - C. learn about
- Key: 11. C 12.C 13.B 14.A 15.B 16.A 17.B 18.B
 19.C

Reading 3

Read the following two passages about Holidays.
 For questions **20-30**, choose the best answer (**A, B or C**).

Holidays

All-Inclusive Holidays

The main reason for the popularity of all-inclusive holidays is that they are very convenient and stress-free. Holidaymakers, whether they go on cheap all-inclusive holidays or on luxurious ones, pay in advance for holiday expenses such as transportation, accommodation and meals so they do not have to worry about overspending during their holidays. Nor do they need to be **concerned about** planning activities as the resorts they go to offer a variety of entertainment for all ages on site, which is included in the price. Package holidays also give you the chance to socialise with other holidaymakers.

These kinds of holidays do not come without drawbacks, however. For one thing, holidaymakers tend to spend their time on the resort grounds and so their interaction with local culture is either limited or non-existent. Those who do decide to explore the culture and the sights of the area will pay extra for such activities. Moreover, the food served may not always be agreeable to all and the all-you-can-eat buffet can encourage guests to eat more than they should. Sometimes a package holiday simply fails to live up to expectations.

20. One reason all-inclusive holidays are popular is that they
- A. organise many excursions.
 - B. are hassle-free.
 - C. offer free meals.

Questions 18-20 refer to BOTH Sections A and B.

28. In which passage(s) is inclusive entertainment mentioned?

- A. Passage A B. Passage B C. Passage A and B

29. The holidays in both passages

- A. are best suited for young people.
B. require you to book your own accommodation.
C. may lead to disappointment.

30. In which passage(s) is limited contact with the local community mentioned?

- A. Passage A B. Passage B C. Passage A and B

Key: 20.B 21.C 22.A 23.A 24.B 25.B 26.A 27.B 28.A
 29.C 30.A

Reading 4

Read the 4 passages below and answer the questions that follow.

A

Mind the gap, London famously reminds its residents and visitors when travelling on the Underground. But the narrow space between the Underground platform and Underground car was nothing compared with the gap that London had to “mind” in staging the planet’s biggest event: essentially 26 simultaneous world championships and two large-scale ceremonies over 17 days in a city of more than seven million people that is already bustling with enough challenges in the usual summer fortnight. But the lead-up to these Olympics was stressfully frantic with challenges: an economic downturn in Britain that made cost-cutting a leading theme for the new Conservative government; rioting of the previous summer that shook London’s sense of well-being and only contributed to the fast-climbing security budget for the Games themselves. As the preparations continued, one would not have been surprised to see more street riots over the online Olympic ticketing process, but at least Britons’ disenchantment with the ticketing reflected a mass interest in actually buying the tickets. That should have been a hint of the enthusiasm to come, but in the weeks immediately before the Games, the focus remained largely on dark doom clouds like missing security guards and on plenty of real clouds, as rain continued to pelt the soon-to-be Olympic city in such large amounts there was worry that some of the venues would sink into the mud.

B

Mind the gap indeed, but in the end it was all water under the many new bridges that decorate the vast Olympic Park. Despite the obstacles and the shadow of the successful, state-backed 2008 Games in Beijing, the London Organizing Committee, headed by Sebastian Coe, undeniably renowned for setting three world records in the mere space of 41 days, said “Today sees the closing of a wonderful Olympic Games in a wonderful city,” Coe said at the wrap-up ceremony on the Sunday night. “We lit the flame, and we lit up the world.” So they did with plenty of help from outsiders like the

Jamaican sprinter Usain Bolt, the American gymnast Gabby Douglas and the Kenyan middle-distance runner David Rudisha, who broke a world record in the 800 meters.

C

London 2012, unlike Beijing or the next Olympic host city, Rio de Janeiro, lacked an overarching geopolitical theme. The Beijing Games were a symbol of China's emergence as a global superpower. The Rio Games, the first in South America, should be a symbol of Brazil's rise and a continent's possibilities. London had to be content with putting on such a superb sporting event, and though there were the odd complaints to be heard — mostly concerning the empty seats in some venues despite voracious public demand — there was plenty of contentment to go around on the last Sunday. "I am such a grateful and happy man," said Jacques Rogge, who was attending his last Olympics as president of the International Olympic Committee. "London promised athletes the Olympic Games, and that is exactly what we got. A splendid village, state-of-the-art venues, 44 world records, 117 Olympic records and I would say that history has been written by many, many athletes". London Olympic Committee chairman, Sebastian Coe, said during the closing ceremonies, "To all the Olympians who came to London to compete, thank you...those of us who came to watch witnessed moments of heroism and heartache that will live long in the memory of the Olympics."

D

All good things must come to an end and so did the Games that were defying predictions of such gloom and fear. But ultimately, London's crowning achievement was that they were Games and only Games in the best sense of the term: happily devoid of a grand scandal that originally caused anxiety and distraction; happily devoid of terrorist activity or ancillary violence. This biggest show on the earth has crowned London to be the only city in the world in which the Olympic Games were held three times (1908, 1948, 2012) making it incomparable. As for London 2012, if it had a spiritual Olympic cousin, the closest would seem to be Sydney in 2000. The two shared popular fervour, a rich cultural attachment to sports, astute planning and a vast Olympic Park built on what was unused, contaminated land: Homebush in the west for Sydney; Stratford in the east for London.

Questions 31– 34

Choose the correct statement that best summarises each paragraph A-D from the list below. Write the correct number i-iv in the space given.

- i. A glorious accomplishment. A well-deserved applause for London 2012 for fulfilling its promises. (C)
 - ii. Proving all predictions wrong, London came out victorious and put on a show that earned it an unmatched status. (D)
 - iii. The grand show that lit up London and the world (B)
 - iv. There were clouds of doubt and fears at the start of the London Olympics 2012 (A)
31. Paragraph A ___iv_____
32. Paragraph B ___iii_____
33. Paragraph C ___i_____
34. Paragraph D ___ii_____

Questions 35-40

Do the following statements agree with the view presented in the passage above?

Write **TRUE** if the statement is in agreement.

Write **FALSE** if the statement is not in agreement.


Write **NOT GIVEN** if the statement does not represent a view expressed in the passage.

35. The London Olympics 2012 were continuous for 17 days. (T)
36. Britons showed the least interest in buying tickets because of the security concerns. (F)
37. The venues built for the Olympics actually sank into the muddy water just weeks before the opening ceremony. (F)
38. Following a bid headed by Sebastian Coe and Ken Livingstone, London was selected as the host city on 6 July 2005 during the 117th IOC Session in Singapore. (NG)
39. Jamaican sprinter Usain Bolt, the American gymnast Gabby Douglas and the Kenyan middle-distance runner David Rudisha were helpers of the London Olympics 2012. (T)
40. Forty-four Olympic records were made during the Olympics 2012. (F)

3. TEST 3 (as given in the google form)

Test of English Reading Skill - CEFR Multi-level South Eastern University of Sri Lanka

This test includes four reading passages. Please answer all questions.

mifkareema@gmail.com [Switch account](#) 

* Required

Email *

Your email _____

Choose your faculty *

Faculty of Arts and Culture

Faculty of Management and Commerce

Faculty of Technology

Faculty of Applied Sciences

Faculty of Engineering

Faculty of Islamic Studies and Arabic Language

Year *

First Year

Second Year

Third Year

Fourth Year

The medium of instruction in the university *

English

Sinhala

Tamil

Gender

Male

Female

Passage 1

Playing Outdoors

In the past, children used (01) outside much more than today. Nowadays, children spend most of their free time at home. They are more (02) in watching television, playing computer games or (03) with their friends on social network sites. (04), playing outdoors is very important for children of all ages. First of all, playing outside gives children the opportunity to exercise while they are having (05) Running, jumping, or riding their bikes (06) also improve their physical development. Secondly, when children play outside, they do activities (07) can make them feel happy and less stressed. If they (08) happier and calmer, they will be able to concentrate more on their schoolwork and (09), they will do better at school. Finally, when children play outside, they get Vitamin D, which is provided (10) the sun. Vitamin D promotes better moods, gives more energy, and improves memory.

For questions 1-10, choose the best answer.

Question 1.

1 point

- played
- to playing
- to play

Question 2.

1 point

- interested
- interesting
- interests

Question 3.

1 point

- discussing
- saying
- chatting

Question 4.

1 point

- Even if
- However
- But

Question 5.

1 point

- funnily
- funny
- fun

Question 6.

1 point

- it
- can
- which

Question 7.

1 point

- who
- but
- that

Question 8.

1 point

- had felt
- feel
- were feeling

Question 9.

1 point

- as a result
- despite
- as well

Question 10.

1 point

- of
- by
- in

Passage 2

Read the text below and then do the exercises that follow.

Supersonic Flight

24 October, 2003 marked the final flight of Concorde, one of the most iconic aeroplanes ever. One of only two supersonic airliners, it was by far the most successful having an outstanding 27-year service before retirement. Since then there have been no supersonic passenger flights, but now several airlines are planning to create new supersonic passenger planes.

A supersonic plane is one that is able to break the sound barrier, or in other words, to travel at speeds greater than the speed of sound. The first plane to do this was the U.S. Bell X-1 in 1947, piloted by Chuck Yeager. There had been several obstacles to overcome in order to achieve this. One obstacle was overheating and another was turbulence which led to planes being difficult to control; the problem of shock waves caused by such speeds also made the plane slow down and thus more powerful engines were required. Another way around this latter difficulty was to design a long and thin aircraft, which was the reason why Concorde had its distinctive shape and pointed nose. The numerous difficulties needed to be overcome to reach sonic speeds led to the popular notion of a 'sound barrier'. This concept was further reinforced by the sonic boom, the noise like an explosion which occurred when objects went faster than the speed of sound.

Sonic booms were a common phenomenon during the early 1970s as Concorde embarked on its commercial transatlantic flights between Britain and the USA at over twice the speed of sound. However, their loudness soon became a source of complaint among the general public, which resulted in Concorde flying supersonic only over the Atlantic. Despite this, it was still able to fly from London to New York in under 3 hours. Another criticism of the plane was that it was almost exclusively for the wealthy. Nevertheless, Concorde continued to operate for almost 3 decades until a variety of factors led to its retirement in 2003.

However, thanks to advances in technology, many companies are now planning to open supersonic routes worldwide as early as 2020. One company, Airbus, is designing a supersonic plane, named Concorde 2, which would be able to travel at more than twice the speed of the original Concorde. However, its turtle-like appearance bears no comparison to the sleek, classic ground-breaking design of its former namesake.

For questions 11-19, choose the best answer.

11. What is mentioned in the 1st paragraph about Concorde?

1 point

- It was the only supersonic airplane.
- It travelled further than any other plane.
- Its service record was exceptional.

12. What is **TRUE** of Chuck Yeager?

1 point

- He designed the Bell X-1.
- He flew the first supersonic plane.
- He discovered the sound barrier in 1947.

13. One phenomenon that is **NOT** an obstacle to breaking the sound barrier is

1 point

- excess heat.
- powerful engines.
- shock waves.

14. The writer says that one way to overcome the problem of shock wave was to

1 point

- slow the plane down.
- add more turbulence.
- make the aircraft slimmer.
- Other: _____

15. Which of the following options can best replace the word **notion** in the 2nd paragraph?

1 point

- idea
- experience
- discovery

16. What is a 'sonic boom'?

1 point

- a reinforced barrier
- a loud sound
- a fast-moving object

17. According to the 3rd paragraph, what was one result of the sonic boom?

1 point

- It allowed transatlantic flights.
- Many people protested about them.
- It stopped supersonic travel across the Atlantic.

18. What does the word **operate** in the 3rd paragraph mean?

1 point

- expand
- function
- prepare

19. One aspect of Concorde 2 that the author is critical of is its

1 point

- speed.
- name.
- look.

Passage 3

Answer questions 20- 23 by reading **Section a**

Holidays

a. All-Inclusive Holidays

The main reason for the popularity of all-inclusive holidays is that they are very convenient and stress-free. Holidaymakers, whether they go on cheap all-inclusive holidays or on luxurious ones, pay in advance for holiday expenses such as transportation, accommodation and meals so they do not have to worry about overspending during their holidays. Nor do they need to be **concerned about** planning activities as the resorts they go to offer a variety of entertainment for all ages on site, which is included in the price. Package holidays also give you the chance to socialise with other holidaymakers.

These kinds of holidays do not come without drawbacks, however. For one thing, holidaymakers tend to spend their time on the resort grounds and so their interaction with local culture is either limited or non-existent. Those who do decide to explore the culture and the sights of the area will pay extra for such activities. Moreover, the food served may not always be agreeable to all and the all-you- can-eat buffet can encourage guests to eat more than they should. Sometimes a package holiday simply fails to live up to expectations.

20. One reason all-inclusive holidays are popular is that they

1 point

- organise many excursions.
- are hassle-free.
- offer free meals.

21. What does the phrase '**concerned about**' in the 1st paragraph mean?

1 point

- thrilled about
- doubtful about
- troubled about

22. What is **TRUE** about holidaymakers on an all-inclusive holiday?

1 point

- They hardly get to see the area around the resort.
- They are not charged for activities outside the resort.
- They often participate in local cultural events.

23. A holidaymaker on an all-inclusive holiday may be dissatisfied with

1 point

- the food provided.
- the small portions of the food served.
- the luxury the resort offers.
- Other: _____

Answer questions 24- 27 by reading **Section b**.

b. Travelling Independently

The most important benefit of travelling independently is that you have the total freedom and flexibility to do whatever you please, whenever you want. As an independent traveller, you also benefit the local economy since you are not confined to the premises of one single resort. More important than anything else, however, is the magic that you feel when you travel independently- the sense of accomplishment from getting from place to place and the growth that occurs when you stretch your comfort zone.

However, organising your holiday on your own is not without pitfalls. First of all, looking for the best accommodation and transportation for your budget might be time-consuming and nerve-racking. You may also face an unpleasant surprise as the hotel you booked may not always be what the website claimed it to be. What is more, the cost of travelling independently can be much higher than that of an organised package holiday. One of the most serious disadvantages, though, is that as an independent traveller, you are not always safe. Not knowing the area, you may unexpectedly find yourself in a dangerous location.

24. What is **FALSE** about travelling independently?

1 point

- It allows travellers to do what they desire.
- It is suggested primarily for long holidays.
- It doesn't oblige travellers to remain at their hotel.

25. What does the word **accomplishment** in the 1st paragraph mean?

1 point

- embarrassment
- achievement
- amusement

26. Organising your holiday on your own

1 point

- may cause you considerable anxiety.
- will not usually take you much time.
- is generally easier than you think.

27. One of the most serious pitfalls of travelling independently is that it may

1 point

- be as expensive as a package holiday.
- put the traveller's safety at risk.
- require online booking.

Questions 28-30 refer to **BOTH** Sections A and B.

28. In which passage(s) is inclusive entertainment mentioned?

1 point

- Passage A
- Passage B
- Passage A and B

29. The holidays in both passages

1 point

- are best suited for young people.
- require you to book your own accommodation.
- may lead to disappointment.

30. In which passage(s) is limited contact with the local community mentioned?

1 point

- Passage A
- Passage B
- Passage A and B

Passage 4

Read the 4 passages below and answer the questions that follow.

A

Sir William Empson, professor of English literature at Sheffield University for nearly twenty years, is noted for revolutionising our ways of reading a poem. The school of literary criticism known as New Criticism gained important support from Empson's *Seven Types of Ambiguity: A Study of Its Effects on English Verse*. This work, together with his other published essays, has become part of the furniture of any good English or American critic's mind. This new approach to poetry appreciation, centred on the reader's close attention to the properties of poetic language, opened up a new field of literary criticism. This was a remarkable accomplishment, considering that Empson did so without proposing to alter previous methods of criticism. He neither revised the standards by which literature is traditionally judged, nor did he invent new ways to reclassify well-known works of literature.

B

In general usage, a word or reference is considered ambiguous if it has more than one possible meaning. In *Seven Types*, Empson proposed to use the word in an extended sense, thinking relevant to his subject any verbal shade, however slight, which gave room for alternative reactions to the same piece of language. Empson's seven types are briefly defined in his book's table of contents:

First-type ambiguities arise when a detail is effective in several ways at once. . . . In second-type ambiguities two or more alternative meanings are fully resolved into one. . . . The condition for the third type ambiguity is that two apparently unconnected meanings are given simultaneously. . . . In the fourth type the different meanings merge to illuminate a complicated state of mind in the author. . . . The fifth type is a fortunate confusion, as when the author is discovering his idea in the act of writing . . . or not holding it in mind all at once. . . . In the sixth type what is said is contradictory or irrelevant and the reader is forced to invent interpretations. . . . The seventh type is that of full contradiction, marking a division in the author's mind.

C

Ambiguity impedes communication when it results from the writer's indecision. Empson argued that it was not to be respected in so far as it was due to weakness or thinness of thought, obscured the matter at hand unnecessarily or when the interest of the passage was not focussed upon it. He regarded it as merely an opportunism in the handling of the material, if the reader failed to understand the ideas which were being shuffled and would be given a general impression of incoherence. However, the protean properties of words are a major component of poetic language. Empson said that being aware of how this facet of language operates was one of the pleasures of poetry. *Seven Types* is primarily an exercise intended to help the reader who has already felt the pleasure understand the nature of his response.

D

Some of Empson's early critics felt that he had simply written himself a license to search for multiple meanings with no awareness of the controlling context in which the local ambiguity appears. On the contrary, Empson guides critics to consider purpose, context, and person in addition to the critical principles of the author and of the public he is writing for when explaining meaning. Most discussions have picked on the book's least interesting aspects, its use of the word *ambiguity* and its ranging of the *types* along a scale of advancing logical disorder. But these matters are really minor. The book is not philosophical but literary, and its aim is to examine lines Empson finds beautiful and haunting. In at least fifteen places Empson shows that the aim of analysis is not so much understanding lines as uncovering whole tracts of the mind. The book is studded with the right things said about a poet or a historical period. In fact, certain passages of Empsonian exegesis have attained classic status, so that the text can't be intelligently considered without them. Empson had, though in lesser measure, Dr. Johnson's extraordinary gift for laying his finger on crucial literary moments; and that alone is likely to ensure him a measure of permanence.

Questions 31-34

Choose the correct title (i-iv) that best summarises each paragraph A-D from the list below.

Titles i – iv

- i. The justification behind the text
- ii. Neglected considerations
- iii. Characterisations of ambiguity
- iv. The reputation of Empson's writing

31. Paragraph A

1 point

- Title i
- Title ii
- Title iii
- Title iv

32. Paragraph B

1 point

- Title i
- Title ii
- Title iii
- Title iv

33. Paragraph C

1 point

- Title i
- Title ii
- Title iii
- Title iv

34. Paragraph D

1 point

- Title i
- Title ii
- Title iii
- Title iv

Questions 35-40

Do the following statements agree with the view presented in the passage above?

Select **"TRUE"** if the statement is in agreement.

Select **"FALSE"** if the statement is not in agreement.

Select **"NOT GIVEN"** if the statement does not represent a view expressed in the passage.

35. Empson proposed no major changes to the practice of literary criticism.

1 point

- True
- False
- Not Given

36. Certain passages in Empson's work have become fundamental to an understanding of the poems they concern. 1 point

- True
- False
- Not Given

37. Empson proposed more than seven types of ambiguity. 1 point

- True
- False
- Not Given

38. Second-type ambiguity concerns apparently different meanings. 1 point

- True
- False
- Not Given

39. Empson is unsuccessful at sorting various types of ambiguity based on the increasing levels of logical disorder. 1 point

- True
- False
- Not Given

40. Empson's use of ambiguity types can be predicted by the unique properties of the verse in which they occur. 1 point

- True
- False
- Not Given

Submit

Clear form

4. TEST 4(with key)

**South Eastern University of Sri Lanka
CEFR Multi-level Test of English Reading Skill**

Test 4

Reading 1

For questions 1-10, choose the best answer (A, B or C) to fill in the gaps.

Having Friends

Some people say that they can rely on no one else except themselves. They prefer (01) alone, without friends. It might be true that sometimes being alone (02) us to focus on ourselves and (03) us to enjoy our own company and love ourselves. However, (04) we don't have a circle of good friends in our lives, we can become isolated. We are social beings and it is difficult for us (05) without friends. By spending time with our friends, (06) learn more about our likes and dislikes. We also learn to accept other people's (07) and tolerate differences. With good friends, we can share the best of times and create happy memories. We might sometimes be (08) by people who we thought would be there for us but they let us down. This should not prevent us, though, from trying to meet new people. We can (09) who we wish to share our life with. With good friends we stop being selfish and we can learn that it is better to give than to (10)

- | | | | |
|-----|-------------------|------------------------|-------------------|
| 01. | A. be | B. <u>to be</u> | C. to being |
| 02. | A. <u>helps</u> | B. helping | C. to help |
| 03. | A. learns | B. says | C. <u>teaches</u> |
| 04. | A. <u>when</u> | B. which | C. where |
| 05. | A. living | B. live | C. <u>to live</u> |
| 06. | A. can | B. <u>we</u> | C. however |
| 07. | A. <u>beliefs</u> | B. believes | C. believing |
| 08. | A. disappoint | B. <u>disappointed</u> | C. disappoints |
| 09. | A. <u>choose</u> | B. pretend | C. stay |
| 10. | A. create | B. miss | C. <u>receive</u> |

Keys 1. B 2.A 3.C 4.A 5.C 6.B 7.A 8.B 9.A 10.C

Reading 2

Read the text below about 'The Magic of the Cinema' and then answer the questions that follow.

For questions 11-20, choose the best answer (A, B or C).

The Magic of the Cinema

Watching films is a popular free time activity for people of all ages. No matter what kind of film you watch, it is a way for you to relax and take a break from your daily routine. In the past, people could only go to the cinema to watch films but nowadays, thanks to modern technology, there are many other ways to enjoy a film: we can now download films from the Internet to watch on our laptops or watch DVDs on our home-cinemas with large TV screens and surround-sound. While some people prefer these modern alternatives as they are more economical and convenient, other film

lovers still believe that going to the cinema is the best way to enjoy the true magic of a film and that nothing else can match **this experience**.

It is certainly true that when you watch a film in the comfort of your home, you can avoid some of the difficulties you find at the cinema such as waiting in a queue to get your ticket or getting annoyed by someone talking on their mobile during the film or by someone else's head blocking your view. When you stop to think about it, though, there are many reasons to watch a film at the cinema. First of all, you can always rely on cinemas to show the latest films. It is also at the cinema that you get the chance to watch previews of all the exciting films coming out soon. And then there's the screen; you could never have such a big screen at home. Even if you found the money to buy one, you wouldn't have a wall in your home big enough to put it on! Also, **despite** the fact that you might find yourself sitting near some annoying people, watching a film as a member of the audience adds to the experience; comedies become funnier as the audience laughs together and horror movies seem scarier when everyone is screaming at the same time. Going to the cinema is also a good way to meet people, either while waiting in the queue to get a ticket, popcorn or refreshments, or while talking about the film when it is over.

If you watch a film at home, you will probably be interrupted by a phone call or you may end up stopping the film to do other things. At the cinema, you simply escape from everything for as long as the film lasts. And that is magic!

11. Which of the following statements is **FALSE**?
 - A. Children like watching films more than adults.
 - B. You can relax watching a film.
 - C. In the past, most people watched films at the cinema.
12. What makes it possible today to watch a film at other places except the cinema?
 - A. the internet connection at home
 - B. the low cost of internet connection
 - C. the development of technology
13. What does '**this experience**' in the 1st paragraph refer to?
 - A. watching a film on a DVD
 - B. going to the cinema
 - C. using digital technology
14. According to the text one disadvantage of watching a film at the cinema is
 - A. buying a ticket on the Internet.
 - B. sitting in the front row of seats.
 - C. not being able to see from your seat.
15. One good thing about going to the cinema is
 - A. getting ultimate experience.
 - B. watching film previews.
 - C. viewing 3D animations.
16. Which of the following can best replace the word '**Despite**' in the 2nd paragraph?
 - A. In spite of
 - B. However
 - C. In addition to
17. Going to the cinema gives us the opportunity to
 - A. stand in queues.
 - B. socialise with others.
 - C. watch funnier comedies.

18. According to the last paragraph, watching a film at home might not be such a good idea as
- A. you may have to stop watching the film to do other things.
 - B. the film may be boring.
 - C. you may feel uncomfortable.
19. The writer says that
- A. watching film at home is comfortable.
 - B. watching film at cinema is more comfortable than watching film at home.
 - C. neither watching film at cinema nor at home is comfortable.
- Key: 11. A 12.C 13.B 14.C 15.B 16.A 17.B 18.A 19.A

Reading 3

Read the following two passages about Holidays.
For questions **20-30**, choose the best answer (**A, B or C**).

Holidays

All-Inclusive Holidays

The main reason for the popularity of all-inclusive holidays is that they are very convenient and stress-free. Holidaymakers, whether they go on cheap all-inclusive holidays or on luxurious ones, pay in advance for holiday expenses such as transportation, accommodation and meals so they do not have to worry about overspending during their holidays. Nor do they need to be **concerned about** planning activities as the resorts they go to offer a variety of entertainment for all ages on site, which is included in the price. Package holidays also give you the chance to socialise with other holidaymakers.

These kinds of holidays do not come without drawbacks, however. For one thing, holidaymakers tend to spend their time on the resort grounds and so their interaction with local culture is either limited or non-existent. Those who do decide to explore the culture and the sights of the area will pay extra for such activities. Moreover, the food served may not always be agreeable to all and the all-you-can-eat buffet can encourage guests to eat more than they should. Sometimes a package holiday simply fails to live up to expectations.

20. One reason all-inclusive holidays are popular is that they
- A. organise many excursions.
 - B. are hassle-free.
 - C. offer free meals.
21. What does the phrase '**concerned about**' in the 1st paragraph mean?
- A. thrilled about
 - B. doubtful about
 - C. troubled about
22. What is **TRUE** about holidaymakers on an all-inclusive holiday?
- A. They hardly get to see the area around the resort.
 - B. They are not charged for activities outside the resort.
 - C. They often participate in local cultural events.

23. A holidaymaker on an all-inclusive holiday may be dissatisfied with
- A. the food provided.
 - B. the small portions of the food served.
 - C. the luxury the resort offers.

Travelling Independently

The most important benefit of travelling independently is that you have the total freedom and flexibility to do whatever you please, whenever you want. As an independent traveller, you also benefit the local economy since you are not confined to the premises of one single resort. More important than anything else, however, is the magic that you feel when you travel independently- the sense of accomplishment from getting from place to place and the growth that occurs when you stretch your comfort zone.

However, organising your holiday on your own is not without pitfalls. First of all, looking for the best accommodation and transportation for your budget might be time-consuming and nerve-racking. You may also face an unpleasant surprise as the hotel you booked may not always be what the website claimed it to be. What is more, the cost of travelling independently can be much higher than that of an organised package holiday. One of the most serious disadvantages, though, is that as an independent traveller, you are not always safe. Not knowing the area, you may unexpectedly find yourself in a dangerous location.

24. What is **FALSE** about travelling independently?
- A. It allows travellers to do what they desire.
 - B. It is suggested primarily for long holidays.
 - C. It doesn't oblige travellers to remain at their hotel.
25. What does the word **accomplishment**' in the 1st paragraph mean?
- A. embarrassment
 - B. achievement
 - C. amusement
26. Organising your holiday on your own
- A. may cause you considerable anxiety.
 - B. will not usually take you much time.
 - C. is generally easier than you think.
27. One of the most serious pitfalls of travelling independently is that it may
- A. be as expensive as a package holiday.
 - B. put the traveller's safety at risk.
 - C. require online booking.

Questions 18-20 refer to BOTH Sections A and B.

28. In which passage(s) is inclusive entertainment mentioned?
- A. Passage A
 - B. Passage B
 - C. Passage A and B
29. The holidays in both passages
- A. are best suited for young people.
 - B. require you to book your own accommodation.
 - C. may lead to disappointment.

D

Key objections to The Shard include concerns over the effect of the panorama of St Paul's Cathedral. Conservationists say it has already impacted on London's skyline. English Heritage has criticised the location of the Shard because of the "colossal" impact on one of the capital's most famous landmarks and questions whether it will be as venerated in 300 years as St Paul's is today. A spokeswoman for the conservation group said: "English Heritage is not against tall buildings, they are a part of London's skyline. The existing skyline is a positive but vulnerable asset which deserves care and respect. It should be managed sensitively."

E

The Shard is a symbol, but of what? Not of an ideal or a heroic event, obviously, but not exactly of the inexorable march of economics, either. It is not a pure expression of land values or of profit-and-loss calculations. It's more eccentric than that, something that popped through a gap in London's wonky, many-layered planning system. If anyone had sat down to plan the most sensible distribution of towers in London, they would not have come up with the Shard, standing alone on a crowded site in a location that is still a bit rickety, with little apparent relation to the things around it. But no one plans London like this and it's unlikely to happen any time soon. Meanwhile, the startling, part-graceful, part-clunky, impressive, slightly nutty Shard is a true monument to the city that made it.

Questions 31 – 35

Which statement (i – vii) is referred to in each paragraph A – E? Write the correct number i – v in the space given.

- i. The Shard is distasteful and favourless. C*
- ii. The Shard damages the overall appearance of London. D*
- iii. The Shard is designed to enlarge the appearance of London. A*
- iv. The Shard is intended to concentrate people near transport. B*
- v. The Shard is a consequence of London's planning system. E*

- 31. Passage A _____ III _____
- 32. Passage B _____ IV _____
- 33. Passage C _____ I _____
- 34. Passage D _____ II _____
- 35. Passage E _____ V _____

Questions 36- 41

Do the following statements agree with the view presented in the text?

*Write **TRUE** if the statement is in agreement.*

*Write **FALSE** if the statement is not in agreement.*

*Write **NOT GIVEN** if the statement does not represent a view expressed in the passage.*


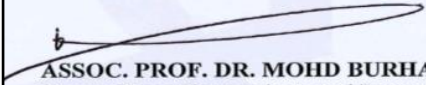


- 36. 'The Shard – Functional Elegance or Hideous Eyesore?' Can be best labelled as the title of the text.
- 37. One of the locals mentions that "The Shard has an unusual look".

38. English Heritage accepts the location of the tower.
39. The top 20 storeys of The Shard are lit blue in recognition of all the wonderful NHS staff and key workers who have kept the country safe during the pandemic and continue to do so.
40. It is likely that London's planning system will continue into the future.

Answer 36. T 37.T 38.F 39.NG 40.T

APPENDIX B

LETTER FOR EXPERT ASSISTANCE AND A SAMPLE OF RATER INFORMATION SHEET

	الجامعة الإسلامية العالمية ماليزيا INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA يُونَيْتِ رَيْسِيَّتِي اِسْلَامًا اَبْنَارًا بَخْسِيًا مِلْدِيَسِيًا <small>(Company No. 101067-F)</small>
KULLIYAH OF EDUCATION	
Our Reference	: IIUM/312/RNP/01/2
Date	: 18 th November 2020
AS PER ATTACHMENT	
<i>Assalamualaikum wrt. wbt.</i>	
Dear Sir/Madam,	
REQUEST FOR EXPERT ASSISTANCE	
NAME	: MOHAMED ISMAIL FOUZUL KAREEMA
STUDENT NO.	: G1918244
May this letter reaches you while you are in the best of <i>imān</i> and health by the grace of Allah <i>s.w.t.</i>	
This is to certify that Sr. Mohamed Ismail Fouzul Kareema (Matric No: G1918244) is a Ph.D student at Kulliyah of Education, IIUM.	
Currently she is writing a thesis entitled “ <i>Development and Validation of Multi Level English Reading test aligned with Common European Framework of Reference using Rasch Model</i> ” under the supervision of Prof. Dr. Ainol Madziah Zubairi. She is requesting your expertise to assist her in validating the contents of her test.	
Your kind considerations are sought and on behalf of the Kulliyah, I wish to thank you for the assistance rendered.	
Thank you. <i>Wassalam.</i>	
	
ASSOC. PROF. DR. MOHD BURHAN IBRAHIM Deputy Dean (Postgraduate and Research) Kulliyah of Education International Islamic University Malaysia	
Note	: <i>This letter is issued upon student's request.</i>
	
	
<i>Garden of Knowledge and Virtue</i>	
Office Address: Kulliyah of Education, International Islamic University Malaysia, Gombak Campus, Jalan Gombak, Selangor. Mailing Address: Kulliyah of Education, P.O. Box 10, 50728 Kuala Lumpur, Malaysia. Tel: +603 6196 5331 / 5334 / 5323 / 6356 / 6351 Fax: +603 6196 4851 / 5926 / 5927 / 6374 / 6375 Website: www.iium.edu.my/educ	

A SAMPLE OF RATER INFORMATION AND INSTRUCTIONS

Rater's demographic information

Name: Associate Professor Dr Ting Su Hie

Gender: Female

Institution: Universiti Malaysia Sarawak

Position: Lecturer

Higher qualifications: PhD (Applied Linguistics), University of Queensland

Specialization: Teaching of English as a Second Language, Sociolinguistics

Years of English language teaching and testing experience: 29 years

Instructions:

Please read the texts and the items (questions) following them and kindly do the following:

1. In the column 'Socio Cognitive Reading Skills'

Task: To what extent each item measure the intended cognitive skills?

Please assign a rating from 1 to 3 for each item to each objective. The rating indicates as follow:

- 1 definitely measure the objective
- 2 uncertain whether the item measures the objective
- 3 does not measure the objective

2. In the column 'type of reading'

Task: Which type of reading does a student apply when reading to find the answer for the item?

Please assign a rating from -1 to +1 for each item to each objective. The rating indicates as follow:

- 1 definitely measure the objective
- 2 uncertain whether the item measures the objective
- 3 does not measure the objective

APPENDIX C I

ITEM OBJECTIVE CONGRUENCE SHEET (TEST 1 AS A SAMPLE)

Test 1:Material: **Learning Resource Network (LRN) CERTIFICATE IN ESOL INTERNATIONAL CEFR B1, B2, C1, & IELCA B1-C2**

Text level & title	Item No	Item	* Please assign a rating from 1 to 3 **Socio-Cognitive Reading Skills								* Please assign a rating from 1 to 3 ***Type of Reading				Remarks
			WR	LA	SP	EPM	I	BMM	CTLS	CITR	Ex/Loc	Ex/Glob	Ca/Loc	Ca/Glob	
CEFR B1 Ice-Cream	1	For questions 1-10 , choose the best answer (A, B or C) and fill in the gaps a hot summer day A. on B. in C. at													
	2 where ice-cream comes from? A. wondered B. wondering C. wonder													
	3 the Chinese A. from B. of C. by													
	4was originally made A. So B. It C. When													
	5 of the freezer in the 20th century A. inventing B. invention C. inventor													
	6 was made from milk A. what B. whose C. which													
	7 no way to store A. is B. was C. will be													
	8 in great quantities A. sold B. cleared C. done													
	9one of the most popular desserts A. became B. becoming C. become													
	10 flavours and colours A. every B. a little C. many													

* Please assign a rating from 1 to 3 for each item for its socio-cognitive skill and type of reading.

1 refers to that the item definitely measures the objective

2 refers to uncertainty whether the item measures the objective

3 refers to that the item does not measure the objective

** Socio-cognitive reading skills

Word Recognition (WR), Lexical Access (LA), Syntactic Parsing (SP), Establishing Propositional Meaning (EPM), Inferencing (I), Building a Mental model (BMM), Creating a Text Level Structure (CTLS), Creating an inter-textual representation (CITR)

*** Expeditious Reading Local, Expeditious Reading Global, Careful Reading Local, Careful Reading Global

Text level & title	Item No	Item	* Please assign a rating from 1 to 3 ** Socio-Cognitive Reading Skills								* Please assign a rating from 1 to 3 *** Type of Reading				Remarks	
			WR	LA	SP	EPM	I	BMM	CTLS	CITR	Ex/Loc	Ex/Glob	Ca/Loc	Ca/Glob		
CEFR C1 Jet Lag	11	For questions 11-19 , choose the best answer (A, B or C). What is mentioned in the first paragraph about jet lag?														
	12	What is TRUE about the symptoms?														
	13	One thing before a trip that will NOT make jet lag worse is														
	14	The writer advises travellers to put off their trip for a few days if they														
	15	Which of the following options can best replace the word deteriorate in the third paragraph?														
	16	One thing travellers shouldn't do while flying is														
	17	During the flight, travellers are advised to														
	18	Once travellers arrive at their destination, it is advisable they should														
	19	What does the word adjusting in the last paragraph mean?														

25	25	-1.00	0.33	-1.00	-1.00	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
26	26	-1.00	-1.00	-1.00	-0.33	-0.33	0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
27	27	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
28	28	-1.00	-1.00	-1.00	-0.33	-1.00	-0.33	0.33	-1.00	1	0.333	-0.80952	8	0.556	Accepted
29	29	-1.00	-1.00	-1.00	-1.00	-0.33	-1.00	0.33	-0.33	1	0.333	-0.80952	8	0.556	Accepted
30	30	-1.00	-1.00	-1.00	-0.33	-1.00	-1.00	0.33	-0.33	1	0.333	-0.80952	8	0.556	Accepted
31	31	-1.00	-1.00	-1.00	-1.00	-1.00	-0.33	0.33	-1.00	1	0.333	-0.90476	8	0.6	Accepted
32	32	-1.00	-1.00	-1.00	-1.00	-1.00	-0.33	0.33	-1.00	1	0.333	-0.90476	8	0.6	Accepted
33	33	-1.00	-1.00	-1.00	-1.00	-1.00	-0.33	0.33	-1.00	1	0.333	-0.90476	8	0.6	Accepted
34	34	-1.00	-1.00	-1.00	-1.00	-1.00	-0.33	0.33	-1.00	1	0.333	-0.90476	8	0.6	Accepted
35	35	-1.00	-1.00	-1.00	0.33	-1.00	-0.33	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
36	36	-1.00	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
37	37	-1.00	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
38	38	-1.00	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
39	39	-1.00	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
40	40	-1.00	-1.00	-1.00	-1.00	-1.00	-0.33	0.33	-1.00	1	0.333	-0.90476	8	0.6	Accepted
41	1	-0.67	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.95238	8	0.978	Accepted
42	2	-0.67	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.95238	8	0.978	Accepted
43	3	1.00	-0.33	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.90476	8	0.956	Accepted
44	4	-1.00	-0.33	0.33	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
45	5	1.00	-1.00	-0.67	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.95238	8	0.978	Accepted
46	6	-0.67	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.95238	8	0.978	Accepted
47	7	-1.00	-1.00	0.67	-0.33	-1.00	-1.00	-1.00	-1.00	1	0.667	-0.90476	8	0.778	Accepted
48	8	-1.00	-1.00	-0.67	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.95238	8	0.978	Accepted
49	9	-0.67	-0.33	0.67	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.667	-0.85714	8	0.756	Accepted
50	10	-1.00	-0.67	1.00	-0.33	-1.00	-1.00	-1.00	-1.00	1	1	-0.85714	8	0.933	Accepted
51	11	1.00	-0.67	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.95238	8	0.978	Accepted
52	12	-0.67	-0.33	-1.00	0.33	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.85714	8	0.578	Accepted
53	13	-1.00	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
54	14	-1.00	-1.00	-1.00	-0.33	0.33	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
55	15	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
56	16	-1.00	-1.00	-1.00	-1.00	-0.67	1.00	-1.00	-1.00	1	1	-0.95238	8	0.978	Accepted
57	17	-1.00	-1.00	-1.00	-0.33	-1.00	0.33	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
58	18	-1.00	-0.33	-1.00	-0.67	-1.00	0.33	-1.00	-1.00	1	0.333	-0.85714	8	0.578	Accepted
59	19	-0.67	-0.33	-1.00	0.33	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.85714	8	0.578	Accepted

60	20	-1.00	-1.00	-0.33	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.90476	8	0.956	Accepted
61	21	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
62	22	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
63	23	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
64	24	-1.00	-1.00	-1.00	-0.33	-0.33	0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
65	25	-1.00	0.33	-1.00	-1.00	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
66	26	-1.00	-1.00	-1.00	-0.33	-0.33	0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
67	27	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
68	28	-1.00	-1.00	-1.00	-0.33	-1.00	-0.33	0.33	-1.00	1	0.333	-0.80952	8	0.556	Accepted
69	29	-1.00	-1.00	-1.00	-1.00	-0.33	-1.00	0.33	-0.33	1	0.333	-0.80952	8	0.556	Accepted
70	30	-1.00	-1.00	-1.00	-0.33	-1.00	-1.00	0.33	-0.33	1	0.333	-0.80952	8	0.556	Accepted
71	31	-1.00	-1.00	-1.00	-1.00	-0.33	0.67	-0.67	-0.33	1	0.667	-0.7619	8	0.711	Accepted
72	32	-1.00	-1.00	-1.00	-1.00	-0.33	0.67	-0.67	-0.33	1	0.667	-0.7619	8	0.711	Accepted
73	33	-1.00	-1.00	-1.00	-1.00	-0.33	0.67	-0.67	-0.33	1	0.667	-0.7619	8	0.711	Accepted
74	34	-1.00	-1.00	-1.00	-1.00	-0.33	0.67	-0.67	-0.33	1	0.667	-0.7619	8	0.711	Accepted
75	35	-1.00	-1.00	-1.00	-1.00	0.33	-1.00	-0.67	-0.33	1	0.333	-0.85714	8	0.578	Accepted
76	36	-1.00	-1.00	-1.00	-1.00	0.33	-1.00	-0.67	-0.33	1	0.333	-0.85714	8	0.578	Accepted
77	37	-1.00	-1.00	-1.00	-1.00	0.33	-1.00	-0.67	-0.33	1	0.333	-0.85714	8	0.578	Accepted
78	38	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.67	1.00	1	1	-0.95238	8	0.978	Accepted
79	39	-1.00	-1.00	-1.00	0.33	-1.00	-1.00	-0.67	-0.33	1	0.333	-0.85714	8	0.578	Accepted
80	40	-1.00	-1.00	0.33	-1.00	-1.00	-1.00	-0.67	-0.33	1	0	-0.85714	8	0.4	
81	1	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
82	2	-1.00	-0.33	0.33	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
83	3	0.33	-1.00	-0.33	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
84	4	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
85	5	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
86	6	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
87	7	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
88	8	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
89	9	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
90	10	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
91	11	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
92	12	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
93	13	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
94	14	-1.00	-1.00	-1.00	-0.33	0.33	-1.00	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted

95	15	-1.00	-1.00	-1.00	0.33	-1.00	1.00	-1.00	-1.00	2	0.667	-1	8	0.81	Accepted
96	16	-1.00	-1.00	-1.00	0.33	-0.33	0.33	-1.00	-1.00	2	0.333	-0.88889	8	0.571	Accepted
97	17	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
98	18	-1.00	1.00	-1.00	-0.33	-1.00	-1.00	-1.00	-1.00	1	1	-0.90476	8	0.956	Accepted
99	19	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
100	20	-1.00	-1.00	-0.33	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.90476	8	0.956	Accepted
101	21	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
102	22	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
103	23	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
104	24	-1.00	-1.00	-1.00	-0.33	-0.33	0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
105	25	-1.00	0.33	-1.00	-1.00	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
106	26	-1.00	-1.00	-1.00	-0.33	-0.33	0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
107	27	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
108	28	-1.00	-1.00	-1.00	-0.33	-1.00	-0.33	0.33	-1.00	1	0.333	-0.80952	8	0.556	Accepted
109	29	-1.00	-1.00	-1.00	-1.00	-0.33	-1.00	0.33	-0.33	1	0.333	-0.80952	8	0.556	Accepted
110	30	-1.00	-1.00	-1.00	-0.33	-1.00	-1.00	0.33	-0.33	1	0.333	-0.80952	8	0.556	Accepted
111	31	-1.00	-1.00	-1.00	-1.00	-1.00	-0.33	0.33	-1.00	1	0.333	-0.90476	8	0.6	Accepted
112	32	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
113	33	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
114	34	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
115	35	-1.00	-1.00	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
116	36	-1.00	-1.00	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
117	37	-1.00	-1.00	-1.00	-1.00	0.33	-0.33	-1.00	-1.00	1	0.333	-0.90476	8	0.6	Accepted
118	38	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
119	39	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
120	40	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
121	1	-0.67	-0.67	0.67	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.667	-0.90476	8	0.778	Accepted
122	2	-0.67	0.00	0.67	-1.00	-1.00	-1.00	-1.00	-1.00	2	0.333	-0.94444	8	0.595	Accepted
123	3	-0.33	-0.33	0.33	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
124	4	-1.00	0.33	0.00	-1.00	-1.00	-1.00	-1.00	-1.00	2	0.167	-1	8	0.524	Accepted
125	5	-0.33	-0.33	0.67	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.667	-0.80952	8	0.733	Accepted
126	6	-1.00	-0.33	0.33	-0.33	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
127	7	-0.67	0.00	0.33	-1.00	-1.00	-1.00	-1.00	-1.00	2	0.167	-0.94444	8	0.5	Accepted
128	8	-1.00	0.00	0.33	0.33	-1.00	-1.00	-1.00	-1.00	3	0.222	-1	8	0.521	Accepted
129	9	-0.33	1.00	-0.33	-0.33	-0.67	-1.00	-1.00	-1.00	1	1	-0.66667	8	0.844	Accepted

130	10	-0.33	1.00	-0.33	-0.33	-0.33	-1.00	-1.00	-1.00	1	1	-0.61905	8	0.822	Accepted
131	11	-1.00	-1.00	-1.00	1.00	-0.33	0.33	-1.00	-1.00	2	0.667	-0.88889	8	0.762	Accepted
132	12	-1.00	-1.00	-0.67	-0.33	0.33	-1.00	-1.00	-1.00	1	0.333	-0.85714	8	0.578	Accepted
133	13	-0.33	1.00	-0.33	-0.33	-1.00	-0.33	-1.00	-1.00	1	1	-0.61905	8	0.822	Accepted
134	14	-0.33	-0.33	-0.33	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.71429	8	0.867	Accepted
135	15	-0.33	-0.33	0.33	0.67	-1.00	-0.33	-1.00	-0.67	2	0.5	-0.61111	8	0.548	Accepted
136	16	-0.67	-0.67	-1.00	-0.33	0.33	-1.00	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
137	17	-0.33	1.00	0.00	-0.33	-1.00	-0.33	-1.00	-1.00	2	0.5	-0.66667	8	0.571	Accepted
138	18	-1.00	-0.33	-0.67	-1.00	-0.33	0.33	-1.00	-1.00	1	0.333	-0.7619	8	0.533	Accepted
139	19	-1.00	-1.00	-1.00	-0.67	-0.33	-0.33	0.33	-1.00	1	0.333	-0.7619	8	0.533	Accepted
140	20	-0.33	-0.33	-0.33	-0.33	-0.33	0.33	-0.33	-1.00	1	0.333	-0.42857	8	0.378	
141	20	-1.00	-1.00	-0.33	1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.90476	8	0.956	Accepted
142	21	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
143	22	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
144	23	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
145	24	-1.00	-1.00	-1.00	-0.33	-0.33	0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
146	25	-1.00	0.33	-1.00	-1.00	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
147	26	-1.00	-1.00	-1.00	-0.33	-0.33	0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
148	27	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	-1.00	1	0.333	-0.80952	8	0.556	Accepted
149	28	-1.00	-1.00	-1.00	-0.33	-1.00	-0.33	0.33	-1.00	1	0.333	-0.80952	8	0.556	Accepted
150	29	-1.00	-1.00	-1.00	-1.00	-0.33	-1.00	0.33	-0.33	1	0.333	-0.80952	8	0.556	Accepted
151	30	-1.00	-1.00	-1.00	-0.33	-1.00	-1.00	0.33	-0.33	1	0.333	-0.80952	8	0.556	Accepted
152	31	-1.00	-1.00	-1.00	-1.00	-0.67	0.33	-0.33	-0.33	1	0.333	-0.7619	8	0.533	Accepted
153	32	-1.00	-1.00	-1.00	-1.00	-0.67	0.33	-0.33	-0.33	1	0.333	-0.7619	8	0.533	Accepted
154	33	-1.00	-1.00	-1.00	-1.00	-0.67	0.33	-0.33	-0.33	1	0.333	-0.7619	8	0.533	Accepted
155	34	-1.00	-1.00	-1.00	-1.00	-0.67	0.33	-0.33	-0.33	1	0.333	-0.7619	8	0.533	Accepted
156	35	-1.00	-1.00	-1.00	-1.00	-0.67	0.33	-0.33	-0.33	1	0.333	-0.7619	8	0.533	Accepted
157	36	-1.00	-1.00	-1.00	-1.00	-1.00	0.00	1.00	-0.67	2	0.5	-0.94444	8	0.69	Accepted
158	37	-1.00	-1.00	-1.00	-1.00	-0.67	1.00	-0.33	-1.00	1	1	-0.85714	8	0.933	Accepted
159	38	-1.00	-1.00	-1.00	0.33	-1.00	-0.33	-0.33	-1.00	1	0.333	-0.80952	8	0.556	Accepted
160	39	-0.67	-0.67	-0.67	0.00	-1.00	-1.00	-0.33	-1.00	1	0	-0.7619	8	0.356	
161	40	-1.00	-1.00	-1.00	-1.00	0.33	-0.33	-0.33	-1.00	1	0.333	-0.80952	8	0.556	Accepted
162	41	-1.00	-1.00	-1.00	-0.33	-0.33	-0.33	-0.33	-1.00	0	0	-0.66667	8	0.333	

APPENDIX D

SAMPLES OF RATED ITEM OBJECTIVE CONGRUENCE SHEETS (RATED BY PROF.TING)

Test1- Material: **Learning Resource Network (LRN) CERTIFICATE IN ESOL INTERNATIONAL CEFR B1, B2, C1 & IELCA B1-C2**

Text level & title	Item No	Item	* Please assign a rating from 1 to 3 **Socio-Cognitive Reading Skills							* Please assign a rating from 1 to 3 ***Type of Reading				Remarks	
			WR	LA	SP	EPM	I	BMM	CTLS	CITR	Ex/Loc	Ex/Glob	Ca/Loc		Ca/Glob
CEFR B1 Ice-Cream	1	For questions 1-10 , choose the best answer (A, B or C) and fill in the gaps a hot summer day A. on B. in C. at													
	2 where ice-cream comes from? A. wondered B. wondering C. wonder													
	3 the Chinese A. from B. of C. by													
	4was originally made A. So B. It C. When													
	5 of the freezer in the 20th century A. inventing B. invention C. inventor													
	6 was made from milk A. what B. whose C. which													
	7 no way to store A. is B. was C. will be													
	8 in great quantities A. sold B. cleared C. done													
	9one of the most popular desserts A. became B. becoming C. become													
	10 flavours and colours A. every B. a little C. many													

* Please assign a rating from 1 to 3 for each item for its socio-cognitive skill and type of reading.

1 refers to that the item definitely measures the objective

2 refers to uncertainty whether the item measures the objective

4 refers to that the item does not measure the objective

** Socio-cognitive reading skills

Word Recognition (WR), Lexical Access (LA), Syntactic Parsing (SP), Establishing Propositional Meaning (EPM), Inferencing (I),

Building a Mental model (BMM), Creating a Text Level Structure (CTLS), Creating an inter-textual representation (CITR)

*** Expeditious Reading Local, Expeditious Reading Global, Careful Reading Local, Careful Reading Global

Text level & title	Item No	Item	* Please assign a rating from 1 to 3 ** Socio-Cognitive Reading Skills							* Please assign a rating from 1 to 3 *** Type of Reading				Remarks		
			WR	LA	SP	EPM	I	BMM	CTLS	CITR	Ex/Loc	Ex/Glob	Ca/Loc		Ca/Glob	
CEFR C1 Jet Lag	11	For questions 11-19 , choose the best answer (A, B or C). What is mentioned in the first paragraph about jet lag?														
	12	What is TRUE about the symptoms?														
	13	One thing before a trip that will NOT make jet lag worse is														
	14	The writer advises travellers to put off their trip for a few days if they														
	15	Which of the following options can best replace the word deteriorate in the third paragraph?														
	16	One thing travellers shouldn't do while flying is														
	17	During the flight, travellers are advised to														
	18	Once travellers arrive at their destination, it is advisable they should														
	19	What does the word adjusting in the last paragraph mean?														

Text level & title	Item No	Item	* Please assign a rating from 1 to 3 **Socio-Cognitive Reading Skills							* Please assign a rating from 1 to 3 ***Type of Reading				Remarks	
			WR	LA	SP	EPM	I	BMM	CTLS	CITR	Ex/Loc	Ex/Glob	Ca/Loc		Ca/Glob
IELCA B1-C2 Influence of technology	31	Choose the correct title for each paragraph A-D from the list below. Write the correct number i – iv. List of Titles i. Ignoring some other forms of technology ii. The need for smartphones and user behaviour iii. Numbers of households that use technology iv. Higher Definition recruits more audiences Paragraph A													
	32	Paragraph B													
	33	Paragraph C													
	34	Paragraph D													
	35	Questions 35-40 Do the following statements agree with the view presented in the passage above? True, false, not given Over half of teenagers who own a...													
	36	The demand for Google is higher...													
	37	Over 50% of 80 year olds have...													
	38	TV and radio are becoming...													
	39	Adults were too embarrassed to ...													
	40	Technology has captured ...													

Additional comments: 13, 16, 24 are not good questions because of the negative wording. 37 – another possible answer is NOT GIVEN. 40 – answer key needs rephrasing

Test 3 Material: **Learning Resource Network (LRN) CERTIFICATE IN ESOL INTERNATIONAL CEFR B1, B2, C1 & IELCA B1-C2**
Passage 1 of Test 3 Rated by Dr. Nor Liza bt Haji Ali (UTM)


Text level & title	Item No	Item	* Please assign a rating from 1 to 3 **Socio-Cognitive Reading Skills							* Please assign a rating from 1 to 3 ***Type of Reading				Remarks		
			WR	LA	SP	EPM	I	BMM	CTLS	CITR	Ex/Loc	Ex/Glob	Ca/Loc		Ca/Glob	
CEFR B1 Playing Outdoors	1	For questions 1-10 , choose the best answer (A, B or C) and fill in the gaps outside much more than A. played B. to playing C. to play			1									X		
	2 in watching television, A. interested B. interesting C. interests			1									X		
	3 with their friends A. discussing B. saying C. chatting			1									X		
	4 playing outdoors is very A. Even if B. However C. But				1								X		
	5 Running, jumping or riding A. funnily B. funny C. fun			1									X		
	6 also improve their physical A. it B. can C. which			1									X		
	7 can make them feel happy A. who B. but C. that			1									X		
	8 happier and calmer, A. had felt B. feel C. were feeling			1									X		
	9 they will do better at school A. as a result B. despite C. as well			1									X		
	10 the sun. A. of B. by C. in			1									X		

Test 2 Material: **Learning Resource Network (LRN) CERTIFICATE IN ESOL INTERNATIONAL CEFR B1, B2, C1 & IELCA B1-C2**
Passage 1 of Test 2 Rated by Dr. Nicola Latimer (CRELLA-University of Bedfordshire, UK)

Text level & title	Item No	Item	* Please assign a rating from 1 to 3 **Socio-Cognitive Reading Skills							* Please assign a rating from 1 to 3 ***Type of Reading				Remarks	
			WR	LA	SP	EPM	I	BMM	CTLS	CITR	Ex/Loc	Ex/Glob	Ca/Loc		Ca/Glob
CEFR B2	1	For questions 1-10 , choose the best answer (A, B or C) and fill in the gaps alone, without friends A. be B. to be C. to being	2	2	1	3	3	3	3	3	3	3	1	3	
Having friends	2 us to focus on ourselves, A. helps B. helping C. to help	2	2	1	3	3	3	3	3	3	3	1	3	
	3 us to enjoy our own A. learns B. says C. teaches	1	1	1	2	3	3	3	3	3	3	1	3	
	4 we don't have A. when B. which C. where	1	1	1	2	3	3	3	3	3	3	1	3	
	5 without friends. A. living B. live C. to live	1	1	1	3	3	3	3	3	3	3	1	3	
	6 learn more about A. can B. we C. however	1	1	1	2	3	3	3	3	3	3	1	3	
	7 and tolerate differences A. beliefs B. believes C. believing	2	2	1	3	3	3	3	3	3	3	1	3	
	8 by people who we A. disappointing B. disappointed C. disappointment	2	2	1	1	3	3	3	3	3	3	1	3	
	9 who we wish to share our life A. choose B. pretend C. stay	1	1	1	1	2	3	3	3	3	3	1	3	
	10	give than to (10) A. create B. miss C. receive	1	1	1	1	1	3	3	3	3	3	1	3	

APPENDIX E

APPROVAL LETTER TO COLLECT DATA

 **الجامعة الإسلامية العالمية ماليزيا**
INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ
Garden of Knowledge and Virtue

KULLIYAH OF EDUCATION

Our Reference : IIUM/312/RNP/01/2
Date : 25th May 2021

Vice Chancellor
South Eastern University of Sri Lanka
University Park
Oluvil, #32360
Sri Lanka.

Assalamualaikum wa. wbt.

Dear Sir/Madam,

PERMISSION TO CONDUCT RESEARCH AT YOUR OFFICE BY SR. MOHAMED ISMAIL FOUZUL KAREEMA (MATIC NO: G1918244)


May this letter reach you while you are in the best of *imān* and health by the grace of Allah *s.w.t*.

This is to certify that Sr. Mohamed Ismail Fouzul Kareema (Matric No: G1918244) is a Ph.D student at Kulliyah of Education, IIUM.




Currently she is writing a thesis entitled "*A Rasch Approach in Validation of CEFR Aligned Reading Tests Among EMI Undergraduates*" under the supervision of Prof. Dr. Ainol Madziah Zubairi. As part of the preparation, we would like to seek your good office to allow her to conduct the above mentioned research at your office

Any assistance rendered to her is greatly appreciated.

Thank you. *Wassalam.*


ASSOC. PROF. DR. MOHD BURHAN IBRAHIM
Deputy Dean (Postgraduate and Research)
Kulliyah of Education
International Islamic University Malaysia

Note : *This letter is issued upon student's request.*

Office Address: Kulliyah of Education, International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur.
Mailing Address: Kulliyah of Education, P.O. Box 13, 50725 Kuala Lumpur, Malaysia.
Tel: +603 6421 5351 / 5335 / 5334 / 5329 / 6058 / 6351 | Fax: +603 6421 4951 / 5894 / 5927 / 6374 / 6375 | Website: www.iiu.edu.my/iodu

APPENDIX F

PILOT STUDY DATA MATRIX: FIT STATISTICS FOR PILOT STUDY

1. Validity of 11 Common Items before Linking

a. Summary Statistics of 127 Measured Items and 124 Measured Persons

```

-----
Calculating Fit Statistics
>=====
Time for estimation: 0:0:1.756
Processing Table 0
concurrent analysis NEW LABEL TO ITEMS.x1sx
-----
| PERSON      124 INPUT      124 MEASURED      INFIT      OUTFIT
|              TOTAL      COUNT      MEASURE REALSE      IMNSQ ZSTD OMNSQ ZSTD|
| MEAN        6.3      11.0      .51      .88      1.02  .1  .92  -.0|
| P.SD        2.6      .0      1.58  .33      .26  .8  .37  -.6|
| REAL RMSE   .94 TRUE SD  1.27 SEPARATION  1.36 PERSON RELIABILITY .65|
|-----|
| ITEM        11 INPUT      11 MEASURED      INFIT      OUTFIT
|              TOTAL      COUNT      MEASURE REALSE      IMNSQ ZSTD OMNSQ ZSTD|
| MEAN       70.9     124.0      .00      .24      1.00  .1  .92  -.3|
| P.SD       19.7      .0      1.09  .04      .18  1.8  .30  1.3|
| REAL RMSE  .24 TRUE SD  1.06 SEPARATION  4.34 ITEM RELIABILITY .95|
|-----|
Output written to C:\Users\user\Desktop\aiium thesis\chapter 3\FINAL CONCURRENT ANALYSIS\ZOU759WS.TXT
CODES= 01
Measures constructed: use "Diagnosis" and "Output Tables" menus

```

b. Dimensionality Map of Common Items

```

TABLE 23.0 concurrent analysis NEW LABEL TO ITEM ZOU759WS.TXT May 20 2021 15:33
INPUT: 124 PERSON 11 ITEM REPORTED: 124 PERSON 11 ITEM 2 CATS WINSTEPS 4.4.7
-----
Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units
Eigenvalue Observed Expected
Total raw variance in observations = 15.7820 100.0% 100.0%
Raw variance explained by measures = 4.7820 30.3% 31.3%
Raw variance explained by persons = 2.3393 14.8% 15.3%
Raw Variance explained by items = 2.4427 15.5% 16.0%
Raw unexplained variance (total) = 11.0000 69.7% 100.0% 68.7%
Unexplnd variance in 1st contrast = 1.8157 11.5% 16.5%
Unexplnd variance in 2nd contrast = 1.4724 9.3% 13.4%
Unexplnd variance in 3rd contrast = 1.4040 8.9% 12.8%
Unexplnd variance in 4th contrast = 1.2379 7.8% 11.3%
Unexplnd variance in 5th contrast = 1.1865 7.5% 10.8%

```

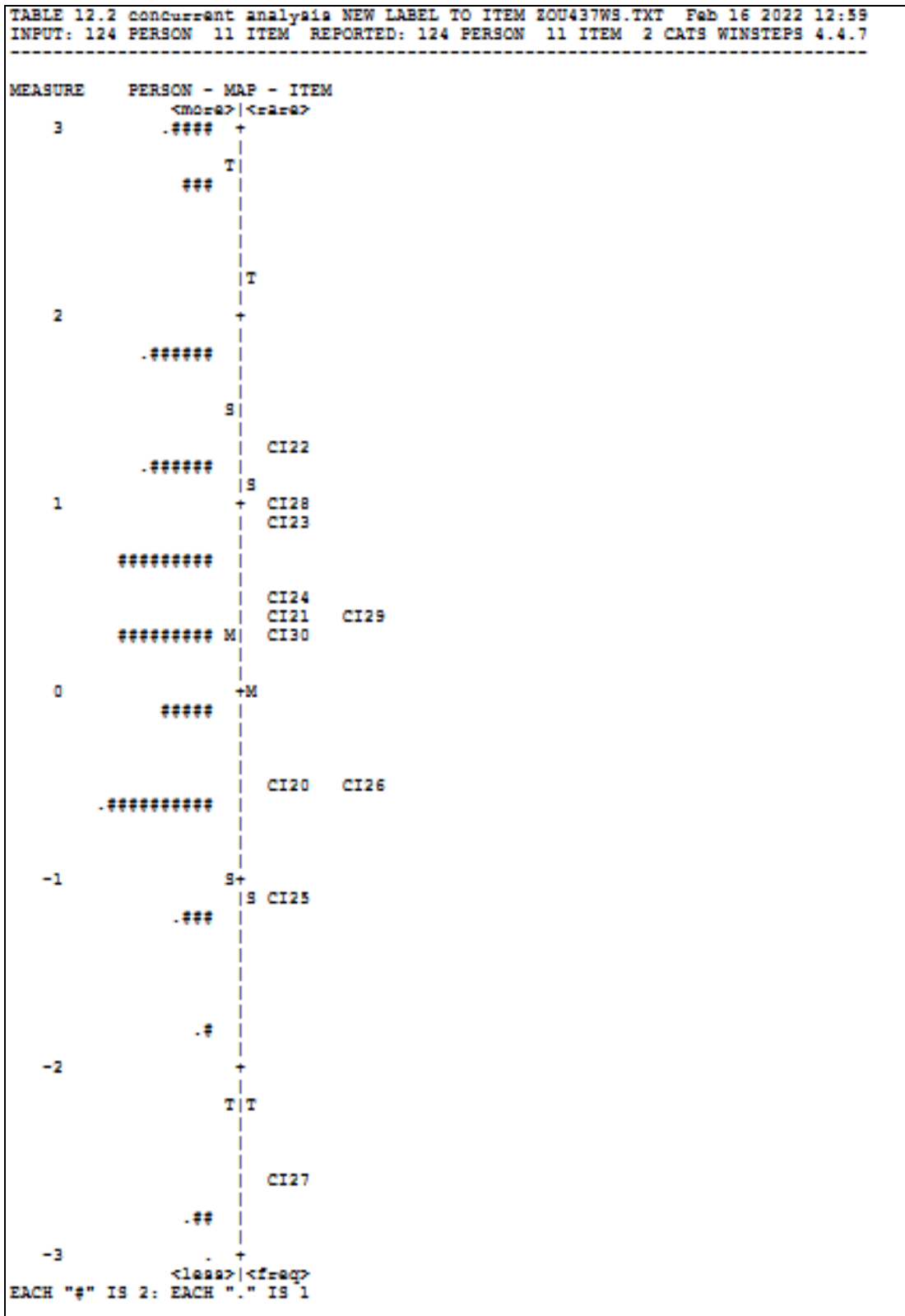
c. Item Fit Statistics: Correlation Order of Common Items

```

TABLE 26.1 concurrent analysis NEW LABEL TO ITEM ZOU666WS.TXT May 24 2021 10:10
INPUT: 124 PERSON 11 ITEM REPORTED: 124 PERSON 11 ITEM 2 CATS WINSTEPS 4.4.7
-----
PERSON: REAL SEP.: 1.36 REL.: .65 ... ITEM: REAL SEP.: 4.34 REL.: .95
ITEM STATISTICS: CORRELATION ORDER
-----
| ENTRY  TOTAL  TOTAL  MODEL  INFIT  OUTFIT  PTMEASUR-AL  EXACT  MATCH
| NUMBER SCORE  COUNT  MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP. | OBS% EXP%| ITEM
|-----|
| 9      51    124    1.00   .22|1.36 3.49|1.52 2.49| .35 .54| 57.9 72.4| CI28
| 4      54    124    .86   .22|1.24 2.56|1.16 .94| .42 .53| 57.0 71.4| CI23
| 1      84    124   -.54   .22|1.07 .71|1.01 .10| .46 .49| 72.8 73.4| CI20
| 7      83    124   -.49   .22|1.05 .53|1.01 .10| .47 .49| 70.2 73.2| CI26
| 5      61    124    .54   .21|1.05 .65|1.15 .98| .49 .53| 73.7 70.4| CI24
| 8     113    124  -2.65   .36|.80 -.74|.37 -1.34| .49 .38| 92.1 91.6| CI27
| 2      63    124    .44   .21|1.06 .71|1.00 .07| .50 .52| 66.7 70.2| CI21
| 6      95    124  -1.14   .25|.85 -1.16|.69 -1.35| .54 .46| 83.3 79.4| CI25
| 11     67    124    .26   .21|.89 -1.32|.80 -1.44| .58 .52| 71.9 70.0| CI30
| 10     65    124    .35   .21|.85 -1.84|.76 -1.79| .60 .52| 73.7 70.0| CI29
| 3      44    124    1.35   .23|.75 -2.58|.61 -1.96| .67 .54| 84.2 74.9| CI22
|-----|
| MEAN   70.9  124.0   .00   .23|1.00  .1| .92  -.3|      | 73.0 74.3|
| P.SD   19.7   .0   1.09  .04|.18  1.8|.30  1.3|      | 10.1  6.1|
|-----|

```

d. Wright Item Map for Common Items



2. Validity of Individual Tests

a. Summary Statistics of TEST1, TEST2, TEST3, and TEST4 individually

```

Calculating Fit Statistics
>=====<
Time for estimation: 0:0:0.859
Processing Table 0
pilot Arts test 1.xlsx
-----
| PERSON      30 INPUT      30 MEASURED      INFIT      OUTFIT
|      TOTAL    COUNT    MEASURE  REALSE    IMNSQ  ZSTD  OMNSQ  ZSTD
| MEAN      23.3    40.0      .44    .41      1.00   .0   .97   .0
| P.SD      5.4     .0      .80    .03      .19   1.1   .32   .9
| REAL RMSE .41 TRUE SD      .69 SEPARATION 1.69 PERSON RELIABILITY .74
|-----
| ITEM       40 INPUT      40 MEASURED      INFIT      OUTFIT
|      TOTAL    COUNT    MEASURE  REALSE    IMNSQ  ZSTD  OMNSQ  ZSTD
| MEAN      17.5    30.0     -1.1   .52      .99   .1   .97   .1
| P.SD      7.2     .0     1.53   .25      .14   .7   .27   .9
| REAL RMSE .58 TRUE SD      1.41 SEPARATION 2.44 ITEM RELIABILITY .86
|-----

```

```

Calculating Fit Statistics
>=====<
Time for estimation: 0:0:1.191
Processing Table 0
pilot test 2 30 samples.xlsx
-----
| PERSON      30 INPUT      30 MEASURED      INFIT      OUTFIT
|      TOTAL    COUNT    MEASURE  REALSE    IMNSQ  ZSTD  OMNSQ  ZSTD
| MEAN      27.6    40.0     1.26   .51      1.00  -.2   1.09  -.1
| P.SD      5.7     .0     1.28   .26      .35   1.8   .81   1.4
| REAL RMSE .57 TRUE SD      1.14 SEPARATION 2.00 PERSON RELIABILITY .80
|-----
| ITEM       40 INPUT      40 MEASURED      INFIT      OUTFIT
|      TOTAL    COUNT    MEASURE  REALSE    IMNSQ  ZSTD  OMNSQ  ZSTD
| MEAN      20.7    30.0     -1.0   .57      .97   .0   1.09   .3
| P.SD      6.9     .0     1.55   .25      .13   .6   .69   .9
| REAL RMSE .63 TRUE SD      1.42 SEPARATION 2.26 ITEM RELIABILITY .84
|-----

```

```

Calculating Fit Statistics
>=====<
Time for estimation: 0:0:1.80
Processing Table 0
pilot test 3 34 samples.xlsx
-----
| PERSON      34 INPUT      34 MEASURED      INFIT      OUTFIT
|      TOTAL    COUNT    MEASURE  REALSE    IMNSQ  ZSTD  OMNSQ  ZSTD
| MEAN      26.1    40.0     .97   .43      1.01   .1   .94   .0
| P.SD      6.6     .0     1.09   .13      .15   .9   .26   .8
| REAL RMSE .45 TRUE SD      .99 SEPARATION 2.22 PERSON RELIABILITY .83
|-----
| ITEM       40 INPUT      40 MEASURED      INFIT      OUTFIT
|      TOTAL    COUNT    MEASURE  REALSE    IMNSQ  ZSTD  OMNSQ  ZSTD
| MEAN      22.2    34.0     .00   .47      1.00   .1   .94   .0
| P.SD      6.1     .0     1.18   .15      .14   .8   .32   .8
| REAL RMSE .49 TRUE SD      1.07 SEPARATION 2.18 ITEM RELIABILITY .83
|-----

```

```

Calculating Fit Statistics
>=====<
Time for estimation: 0:0:1.444
Processing Table 0
Test 4 value new ID.xlsx
-----
| PERSON      30 INPUT      30 MEASURED      INFIT      OUTFIT
|      TOTAL    COUNT    MEASURE  REALSE    IMNSQ  ZSTD  OMNSQ  ZSTD
| MEAN      29.9    40.0     1.66   .52      .92  -.1   .96   .1
| P.SD      6.5     .0     1.28   .11      .36   1.4   .86   1.3
| REAL RMSE .53 TRUE SD      1.16 SEPARATION 2.19 PERSON RELIABILITY .83
|-----
| ITEM       40 INPUT      40 MEASURED      INFIT      OUTFIT
|      TOTAL    COUNT    MEASURE  REALSE    IMNSQ  ZSTD  OMNSQ  ZSTD
| MEAN      22.4    30.0     -1.0   .64      .96   .1   .96   .0
| P.SD      6.4     .0     1.61   .25      .29   1.1   1.04   1.2
| REAL RMSE .69 TRUE SD      1.45 SEPARATION 2.10 ITEM RELIABILITY .82
|-----

```


b. Dimensionality Map of TEST1, TEST2, TEST3, and TEST4 individually

TABLE 23.0 pilot Arts test 1.xlsx		ZOU196WS.TXT		May 21 2021 11: 6	
INPUT: 30 PERSON 40 ITEM REPORTED: 30 PERSON 40 ITEM 2 CATS		WINSTEPS 4.4.7			

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units					
		Eigenvalue	Observed	Expected	
Total raw variance in observations	=	56.3416	100.0%	100.0%	
Raw variance explained by measures	=	17.3416	30.8%	31.0%	
Raw variance explained by persons	=	5.1776	9.2%	9.2%	
Raw Variance explained by items	=	12.1640	21.6%	21.7%	
Raw unexplained variance (total)	=	39.0000	69.2%	100.0%	
Unexplned variance in 1st contrast	=	4.2266	7.5%	10.8%	
Unexplned variance in 2nd contrast	=	3.7913	6.7%	9.7%	
Unexplned variance in 3rd contrast	=	3.4523	6.1%	8.9%	
Unexplned variance in 4th contrast	=	3.0542	5.4%	7.8%	
Unexplned variance in 5th contrast	=	2.7265	4.8%	7.0%	

TABLE 23.0 pilot test 2 30 samples.xlsx		ZOU830WS.TXT		May 21 2021 11:44	
INPUT: 30 PERSON 40 ITEM REPORTED: 30 PERSON 40 ITEM 2 CATS		WINSTEPS 4.4.7			

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units					
		Eigenvalue	Observed	Expected	
Total raw variance in observations	=	60.7688	100.0%	100.0%	
Raw variance explained by measures	=	21.7688	35.8%	34.9%	
Raw variance explained by persons	=	7.2379	11.9%	11.6%	
Raw Variance explained by items	=	14.5309	23.9%	23.3%	
Raw unexplained variance (total)	=	39.0000	64.2%	100.0%	
Unexplned variance in 1st contrast	=	6.1581	10.1%	15.8%	
Unexplned variance in 2nd contrast	=	4.6986	7.7%	12.0%	
Unexplned variance in 3rd contrast	=	4.3998	7.2%	11.3%	
Unexplned variance in 4th contrast	=	3.1835	5.2%	8.2%	
Unexplned variance in 5th contrast	=	2.9224	4.8%	7.5%	

TABLE 23.0 pilot test 3 34 samples.xlsx		ZOU572WS.TXT		May 21 2021 11:58	
INPUT: 34 PERSON 40 ITEM REPORTED: 34 PERSON 40 ITEM 2 CATS		WINSTEPS 4.4.7			

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units					
		Eigenvalue	Observed	Expected	
Total raw variance in observations	=	54.9393	100.0%	100.0%	
Raw variance explained by measures	=	14.9393	27.2%	27.9%	
Raw variance explained by persons	=	6.0238	11.0%	11.2%	
Raw Variance explained by items	=	8.9156	16.2%	16.6%	
Raw unexplained variance (total)	=	40.0000	72.8%	100.0%	
Unexplned variance in 1st contrast	=	4.6420	8.4%	11.6%	
Unexplned variance in 2nd contrast	=	4.2136	7.7%	10.5%	
Unexplned variance in 3rd contrast	=	3.9834	7.3%	10.0%	
Unexplned variance in 4th contrast	=	3.0873	5.6%	7.7%	
Unexplned variance in 5th contrast	=	2.5768	4.7%	6.4%	

TABLE 23.0 Test 4 value new ID.xlsx		ZOU268WS.TXT		May 21 2021 12: 9	
INPUT: 30 PERSON 40 ITEM REPORTED: 30 PERSON 40 ITEM 2 CATS		WINSTEPS 4.4.7			

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units					
		Eigenvalue	Observed	Expected	
Total raw variance in observations	=	65.1711	100.0%	100.0%	
Raw variance explained by measures	=	26.1711	40.2%	40.1%	
Raw variance explained by persons	=	12.5783	19.3%	19.3%	
Raw Variance explained by items	=	13.5928	20.9%	20.9%	
Raw unexplained variance (total)	=	39.0000	59.8%	100.0%	
Unexplned variance in 1st contrast	=	6.2435	9.6%	16.0%	
Unexplned variance in 2nd contrast	=	4.9080	7.5%	12.6%	
Unexplned variance in 3rd contrast	=	4.2549	6.5%	10.9%	
Unexplned variance in 4th contrast	=	3.8198	5.9%	9.8%	
Unexplned variance in 5th contrast	=	3.2331	5.0%	8.3%	

3. Concurrent Analysis of all Four Tests

a. Summary Statistics of 127 Measured Items and 124 Measured Persons

TABLE 3.1 concurrent analysis NEW LABEL TO ITEMS ZOU322WS.TXT Feb 15 2022 13:30
 INPUT: 124 PERSON 127 ITEM REPORTED: 124 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7

SUMMARY OF 123 MEASURED (NON-EXTREME) PERSON								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	26.6	40.0	1.19	.43	.99	.00	.98	.01
SEM	.6	.0	.10	.01	.02	.11	.05	.10
P.SD	6.4	.0	1.14	.10	.23	1.17	.53	1.08
S.SD	6.4	.0	1.14	.10	.23	1.17	.54	1.09
MAX.	39.0	40.0	4.81	1.05	1.61	2.80	3.77	3.54
MIN.	11.0	40.0	-1.36	.35	.53	-3.10	.11	-2.25
REAL RMSE	.45	TRUE SD	1.04	SEPARATION	2.29	PERSON RELIABILITY	.84	
MODEL RMSE	.44	TRUE SD	1.05	SEPARATION	2.39	PERSON RELIABILITY	.85	
S.E. OF PERSON MEAN - .10								
MAXIMUM EXTREME SCORE: 1 PERSON .84								
SUMMARY OF 124 MEASURED (EXTREME AND NON-EXTREME) PERSON								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	26.7	40.0	1.23	.44				
SEM	.6	.0	.11	.01				
P.SD	6.5	.0	1.20	.16				
S.SD	6.5	.4	1.20	.16				
MAX.	40.0	40.0	5.44	1.84				
MIN.	11.0	40.0	-1.36	.35				
REAL RMSE	.48	TRUE SD	1.09	SEPARATION	2.27	PERSON RELIABILITY	.84	
MODEL RMSE	.47	TRUE SD	1.10	SEPARATION	2.35	PERSON RELIABILITY	.85	
S.E. OF PERSON MEAN - .11								
PERSON RAW SCORE-TO-MEASURE CORRELATION - .95								
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY - .51 SEM - 4.57								
SUMMARY OF 124 MEASURED (NON-EXTREME) ITEM								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	26.0	39.3	.00	.50	.99	.08	.99	.10
SEM	1.5	2.4	.12	.02	.02	.08	.06	.09
P.SD	16.5	26.5	1.39	.18	.18	.90	.65	1.03
S.SD	16.5	26.6	1.39	.18	.18	.90	.65	1.04
MAX.	113.0	124.0	3.54	1.07	1.60	3.45	5.17	4.07
MIN.	3.0	30.0	-3.15	.20	.50	-2.21	.16	-2.26
REAL RMSE	.55	TRUE SD	1.27	SEPARATION	2.33	ITEM RELIABILITY	.84	
MODEL RMSE	.53	TRUE SD	1.28	SEPARATION	2.40	ITEM RELIABILITY	.85	
S.E. OF ITEM MEAN - .12								
MAXIMUM EXTREME SCORE: 3 ITEM 2.44								
SUMMARY OF 127 MEASURED (EXTREME AND NON-EXTREME) ITEM								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	26.1	39.1	-.09	.53				
SEM	1.4	2.3	.13	.02				
P.SD	16.3	26.2	1.50	.27				
S.SD	16.1	26.1	1.50	.27				
MAX.	113.0	124.0	3.54	1.87				
MIN.	3.0	30.0	-4.40	.20				
REAL RMSE	.61	TRUE SD	1.37	SEPARATION	2.24	ITEM RELIABILITY	.83	
MODEL RMSE	.60	TRUE SD	1.37	SEPARATION	2.30	ITEM RELIABILITY	.84	
S.E. OF ITEM MEAN - .13								

b. Item Fit Statistics: Measure Order of 127 Items

TABLE 13.1 concurrent analysis NEW LABEL TO ITEM EOU322WS.TXT Feb 15 2022 13:30
 INPUT: 124 PERSON 127 ITEM REPORTED: 124 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7

PERSON: REAL SEP.: 2.27 REL.: .84 ... ITEM: REAL SEP.: 2.24 REL.: .83

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INPIT MNSQ	ESTD MNSQ	OUTPIT MNSQ	ESTD CORR.	PTMEASUR-AL EXP.	EXACT OBS%	MATCH EXP%	ITEM	
117	6	30	3.54	.49	1.60	1.91	5.17	3.71	-.31	.35	76.7	81.8	T4Q19
121	8	30	3.10	.45	1.14	.72	1.17	.49	.26	.38	76.7	76.2	T4Q34
38	3	30	2.95	.63	1.05	.26	1.01	.24	.16	.21	90.0	90.1	T1Q38
124	9	30	2.90	.44	1.36	1.83	3.89	4.07	-.10	.39	73.3	73.6	T4Q37
98	9	34	2.67	.43	1.29	1.25	1.64	1.69	.15	.43	70.6	77.5	T3Q40
88	10	34	2.49	.42	.95	-.20	.90	-.23	.47	.43	76.5	75.7	T3Q19
39	5	30	2.32	.51	1.08	.37	.93	.02	.21	.26	83.3	83.5	T1Q39
83	13	34	2.00	.39	1.07	.46	1.13	.61	.37	.43	64.7	71.3	T3Q14
86	13	34	2.00	.39	1.16	.98	1.16	.71	.30	.43	58.8	71.3	T3Q17
3	44	124	1.95	.21	.81	-2.21	.69	-2.26	.61	.46	77.2	72.8	CI22
127	15	30	1.87	.41	1.57	3.45	2.01	2.86	-.01	.44	46.7	68.0	T4Q40
42	10	30	1.86	.44	.81	-.90	.71	-1.04	.60	.46	82.8	73.8	T2Q2
66	10	30	1.86	.44	.91	-.39	.90	-.27	.52	.46	75.9	73.8	T2Q37
80	15	34	1.70	.38	1.00	.05	.91	-.33	.45	.43	64.7	69.2	T3Q11
22	8	30	1.67	.44	1.04	.26	1.05	.26	.26	.31	70.0	74.2	T1Q11
9	51	124	1.64	.21	1.24	2.64	1.32	2.27	.27	.46	57.7	71.1	CI28
82	16	34	1.56	.38	.95	-.34	.94	-.21	.47	.42	70.6	68.5	T3Q13
119	17	30	1.53	.42	.94	-.30	.83	-.52	.51	.46	73.3	71.1	T4Q32
4	54	124	1.52	.20	1.23	2.65	1.26	1.94	.28	.46	61.0	70.6	CI23
64	12	30	1.50	.42	.98	-.06	1.11	.56	.43	.44	79.3	70.5	T2Q35
96	17	34	1.41	.38	1.06	.48	1.04	.24	.37	.42	64.7	67.9	T3Q38
53	13	30	1.33	.41	.79	-1.36	.94	-.22	.55	.43	82.8	68.7	T2Q13
12	10	30	1.31	.41	1.27	1.53	1.36	1.45	.01	.33	53.3	70.0	T1Q1
92	18	34	1.27	.38	.82	-1.38	.73	-1.19	.58	.41	73.5	67.5	T3Q34
5	61	124	1.23	.20	.94	-.69	.95	-.40	.49	.45	74.8	69.5	CI24
122	19	30	1.17	.43	1.12	.65	1.14	.52	.38	.47	66.7	74.2	T4Q35
58	14	30	1.16	.41	.96	-.23	.93	-.31	.45	.42	72.4	67.3	T2Q18
2	63	124	1.15	.20	1.07	.89	1.07	.59	.40	.45	67.5	69.3	CI21
37	11	30	1.14	.40	1.13	.86	1.10	.54	.20	.34	63.3	68.4	T1Q37
89	19	34	1.12	.38	.96	-.25	1.79	2.64	.37	.41	67.6	67.5	T3Q31
10	65	124	1.07	.20	.87	-1.78	.78	-1.86	.56	.45	71.5	69.3	CI29
51	15	30	.99	.40	1.06	.44	1.06	.39	.37	.41	72.4	66.1	T2Q11
11	67	124	.98	.20	.86	-1.82	.80	-1.65	.55	.44	74.0	69.2	CI30
20	12	30	.98	.40	1.27	1.81	1.44	2.10	.00	.35	53.3	67.0	T1Q9
30	12	30	.98	.40	1.00	.05	1.11	.63	.31	.35	73.3	67.0	T1Q19
126	20	30	.98	.44	1.01	.12	1.04	.23	.45	.47	76.7	75.9	T4Q39
78	21	34	.83	.39	1.08	.61	1.05	.27	.33	.39	58.8	68.5	T3Q9
91	21	34	.83	.39	.72	-2.08	.62	-1.41	.63	.39	82.4	68.5	T3Q33
97	21	34	.83	.39	1.30	1.94	1.26	.93	.15	.39	47.1	68.5	T3Q39
114	21	30	.78	.46	1.01	.11	1.25	.74	.43	.47	80.0	77.7	T4Q16
125	21	30	.78	.46	.85	-.56	.71	-.69	.59	.47	80.0	77.7	T4Q38
28	14	30	.67	.39	1.04	.39	1.02	.21	.31	.35	63.3	65.1	T1Q17
85	23	34	.52	.40	1.17	1.05	1.30	.91	.20	.37	70.6	71.7	T3Q16
87	23	34	.52	.40	.94	-.34	.78	-.57	.45	.37	64.7	71.7	T3Q18
75	24	34	.36	.41	1.15	.89	1.47	1.20	.19	.36	70.6	73.6	T3Q6
95	24	34	.36	.41	.83	-.97	.70	-.74	.52	.36	82.4	73.6	T3Q37
47	19	30	.34	.41	1.00	.02	.88	-.43	.39	.36	55.2	67.4	T2Q7
7	83	124	.31	.21	.96	-.36	.86	-.74	.45	.41	68.3	72.3	CI26
1	84	124	.26	.21	1.04	.44	.98	-.03	.39	.41	69.9	72.7	CI20
93	25	34	.19	.42	1.18	.95	1.02	.19	.23	.34	67.6	75.6	T3Q35
45	20	30	.16	.42	.86	-.85	1.19	.76	.41	.35	82.8	69.3	T2Q5
118	24	30	.07	.52	.69	-.98	.52	-.82	.67	.45	90.0	83.9	T4Q31
120	24	30	.07	.52	.50	-1.80	.32	-1.46	.79	.45	90.0	83.9	T4Q33
23	18	30	.06	.40	.93	-.43	.91	-.47	.43	.35	63.3	66.5	T1Q12
32	18	30	.06	.40	.85	-1.09	.83	-.96	.53	.35	76.7	66.5	T1Q32
33	18	30	.06	.40	1.19	1.29	1.31	1.61	.10	.35	63.3	66.5	T1Q33
36	18	30	.06	.40	.89	-.73	.84	-.84	.48	.35	70.0	66.5	T1Q36
71	26	34	.01	.43	1.14	.67	1.13	.43	.22	.33	73.5	77.7	T3Q2
73	26	34	.01	.43	1.29	1.29	1.44	.98	.08	.33	67.6	77.7	T3Q4
81	26	34	.01	.43	.97	-.07	1.06	.28	.33	.33	85.3	77.7	T3Q12
90	26	34	.01	.43	.91	-.38	.77	-.37	.43	.33	79.4	77.7	T3Q32
94	26	34	.01	.43	.94	-.23	.72	-.50	.42	.33	73.5	77.7	T3Q36
60	21	30	-.02	.43	.92	-.43	.84	-.47	.40	.34	72.4	71.6	T2Q31
62	21	30	-.02	.43	1.04	.30	.95	-.07	.32	.34	65.5	71.6	T2Q33
67	21	30	-.02	.43	.91	-.49	.91	-.22	.40	.34	79.3	71.6	T2Q38
13	19	30	-.10	.40	1.17	1.10	1.21	1.02	.14	.35	66.7	68.0	T1Q2
19	19	30	-.10	.40	1.21	1.32	1.38	1.75	.07	.35	60.0	68.0	T1Q8
77	27	34	-.18	.45	1.19	.83	1.15	.44	.17	.31	73.5	79.9	T3Q8
105	25	30	-.22	.56	.81	-.45	.46	-.79	.61	.44	83.3	86.0	T4Q7
108	25	30	-.22	.56	.70	-.83	.40	-.93	.67	.44	90.0	86.0	T4Q10
16	20	30	-.27	.41	.90	-.55	.80	-.86	.48	.34	66.7	70.2	T1Q5
6	95	124	-.28	.23	.88	-1.03	.78	-.88	.46	.37	80.5	78.3	CI25
46	23	30	-.41	.46	.97	-.07	.98	.07	.32	.30	75.9	76.9	T2Q6

Item Fit Statistics: Measure Order of 127 Items (Continue)

46	23	30	-.41	.46	.97	-.07	.98	.07	.32	-.30	75.9	76.9	T2Q6
99	26	30	-.56	.61	.94	-.02	.85	.08	.47	-.42	86.7	88.3	T4Q1
100	26	30	-.56	.61	.71	-.70	.42	-.65	.63	-.42	83.3	88.3	T4Q2
101	26	30	-.56	.61	1.12	.43	.67	-.19	.41	-.42	86.7	88.3	T4Q3
104	26	30	-.56	.61	1.09	.36	.78	-.02	.41	-.42	86.7	88.3	T4Q6
111	26	30	-.56	.61	1.14	.48	.99	.25	.36	-.42	86.7	88.3	T4Q13
59	24	30	-.63	.48	1.20	.79	2.46	2.47	.02	-.29	82.8	79.8	T2Q19
61	24	30	-.63	.48	1.11	.49	1.06	-.29	.21	-.29	75.9	79.8	T2Q32
63	24	30	-.63	.48	.88	-.38	.79	-.34	.37	-.29	82.8	79.8	T2Q34
18	22	30	-.63	.44	.95	-.16	.87	-.37	.39	-.33	76.7	74.8	T1Q7
25	22	30	-.63	.44	1.05	.29	1.01	-.12	.28	-.33	76.7	74.8	T1Q14
29	22	30	-.63	.44	.91	-.38	.99	.07	.40	-.33	83.3	74.8	T1Q18
84	29	34	-.64	.51	.78	-.63	.49	-.71	.49	-.27	85.3	85.4	T3Q15
14	23	30	-.83	.45	.83	-.68	.69	-.91	.53	-.32	76.7	77.4	T1Q3
26	23	30	-.83	.45	.86	-.53	.74	-.74	.49	-.32	76.7	77.4	T1Q15
54	25	30	-.88	.51	.91	-.21	.92	.02	.32	-.26	86.2	82.8	T2Q14
65	25	30	-.88	.51	.93	-.14	1.00	.18	.30	-.26	79.3	82.8	T2Q36
70	30	34	-.92	.56	.94	-.04	.72	-.12	.32	-.25	88.2	88.3	T3Q1
74	30	34	-.92	.56	.96	.02	.85	.06	.29	-.25	88.2	88.3	T3Q5
76	30	34	-.92	.56	.97	.06	.69	-.18	.31	-.25	88.2	88.3	T3Q7
102	27	30	-.97	.68	1.15	.48	1.01	.35	.30	-.39	86.7	90.3	T4Q4
109	27	30	-.97	.68	1.36	.86	.88	.21	.24	-.39	86.7	90.3	T4Q11
112	27	30	-.97	.68	.69	-.62	.37	-.52	.59	-.39	93.3	90.3	T4Q14
115	27	30	-.97	.68	1.19	.55	.85	.18	.33	-.39	86.7	90.3	T4Q17
116	27	30	-.97	.68	.51	-1.16	.20	-.92	.70	-.39	93.3	90.3	T4Q18
123	27	30	-.97	.68	1.12	.40	1.00	.34	.31	-.39	93.3	90.3	T4Q36
34	24	30	-1.05	.48	1.18	.71	1.63	1.48	.02	-.30	76.7	80.3	T1Q34
50	26	30	-1.17	.56	.92	-.11	.70	-.33	.32	-.24	86.2	86.1	T2Q10
68	26	30	-1.17	.56	.88	-.21	.59	-.57	.36	-.24	86.2	86.1	T2Q39
24	25	30	-1.29	.51	.96	-.03	1.14	.45	.30	-.28	80.0	83.3	T1Q13
27	25	30	-1.29	.51	.71	-.91	.49	-1.22	.64	-.28	86.7	83.3	T1Q16
31	25	30	-1.29	.51	.80	-.58	.63	-.79	.53	-.28	86.7	83.3	T1Q31
103	28	30	-1.50	.80	.75	-.31	.23	-.61	.54	-.33	93.3	93.4	T4Q5
107	28	30	-1.50	.80	.97	.14	1.05	.43	.32	-.33	93.3	93.4	T4Q9
55	27	30	-1.52	.63	.75	-.47	.41	-.76	.42	-.21	89.7	89.6	T2Q15
56	27	30	-1.52	.63	1.09	.35	2.09	1.35	.07	-.21	89.7	89.6	T2Q16
57	27	30	-1.52	.63	1.00	.16	1.20	.51	.18	-.21	89.7	89.6	T2Q17
69	27	30	-1.52	.63	1.13	.43	3.69	2.41	-.03	-.21	89.7	89.6	T2Q40
17	26	30	-1.58	.56	.85	-.30	.62	-.64	.46	-.26	86.7	86.5	T1Q6
8	113	124	-1.58	.33	1.00	.08	.69	-.60	.28	-.25	91.1	91.0	CI27
72	32	34	-1.73	.75	.80	-.15	.30	-.62	.41	-.18	94.1	94.2	T3Q3
40	27	30	-1.93	.63	.98	.11	.64	-.40	.32	-.23	90.0	89.9	T1Q40
41	28	30	-1.99	.75	.84	-.08	.47	-.35	.32	-.18	93.1	93.1	T2Q1
44	28	30	-1.99	.75	.74	-.27	.31	-.67	.40	-.18	93.1	93.1	T2Q4
49	28	30	-1.99	.75	.91	.05	1.09	.41	.20	-.18	93.1	93.1	T2Q9
110	29	30	-2.33	1.07	.82	.06	.16	-.49	.42	-.25	96.7	96.7	T4Q12
113	29	30	-2.33	1.07	.82	.06	.16	-.49	.42	-.25	96.7	96.7	T4Q15
79	33	34	-2.48	1.03	.92	.22	.31	-.22	.27	-.13	97.1	97.1	T3Q10
48	29	30	-2.74	1.03	.82	.09	.24	-.34	.31	-.13	96.6	96.5	T2Q8
52	29	30	-2.74	1.03	.82	.09	.24	-.34	.31	-.13	96.6	96.5	T2Q12
15	29	30	-3.15	1.02	1.06	.36	1.24	.60	.02	-.15	96.7	96.6	T1Q4
21	29	30	-3.15	1.02	.94	.24	.42	-.15	.28	-.15	96.7	96.6	T1Q10
106	30	30	-3.63	1.87		MINIMUM MEASURE			.00	.00	100.0	100.0	T4Q8
43	30	30	-3.99	1.84		MINIMUM MEASURE			.00	.00	100.0	100.0	T2Q3
35	30	30	-4.40	1.83		MINIMUM MEASURE			.00	.00	100.0	100.0	T1Q35
MEAN	26.1	39.1	-.09	.53	.99	.1	.99	.1			78.4	79.0	
P.SD	16.3	26.2	1.50	.27	.18	.9	.65	1.0			11.6	9.4	

c. Item Fit Statistics: Correlation Order

TABLE 26.1 concurrent analysis NEW LABEL TO ITEM ZOU322WS.TXT Feb 15 2022 13:30
 INPUT: 124 PERSON 127 ITEM REPORTED: 124 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7
 PERSON: REAL SEP.: 2.27 REL.: .84 ... ITEM: REAL SEP.: 2.24 REL.: .83

ITEM STATISTICS: CORRELATION ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	OUTFIT ESTD/MNSQ	PTMEASUR-CORR.	AL-EXP.	EXACT OBS%	MATCH EXP%	ITEM		
117	6	30	3.54	.49	1.60	1.91	5.17	3.71	-.31	.35	76.7	81.8	T4Q19
124	9	30	2.90	.44	1.36	1.83	3.89	4.07	-.10	.39	73.3	73.6	T4Q37
69	27	30	-1.52	.63	1.13	.43	3.69	2.41	-.03	.21	89.7	89.6	T2Q40
127	15	30	1.87	.41	1.57	3.45	2.01	2.86	-.01	.44	46.7	68.0	T4Q40
20	12	30	.98	.40	1.27	1.81	1.44	2.10	.00	.35	53.3	67.0	T1Q9
12	10	30	1.31	.41	1.27	1.53	1.36	1.45	.01	.33	53.3	70.0	T1Q1
59	24	30	-.63	.48	1.20	.79	2.46	2.47	.02	-.29	82.8	79.8	T2Q19
15	29	30	-3.15	1.02	1.06	.36	1.24	.60	.02	-.15	96.7	96.6	T1Q4
34	24	30	-1.05	.48	1.18	.71	1.63	1.48	.02	-.30	76.7	80.3	T1Q34
19	19	30	-.10	.40	1.21	1.32	1.38	1.75	.07	.35	60.0	68.0	T1Q8
56	27	30	-1.52	.63	1.09	.35	2.09	1.35	.07	-.21	89.7	89.6	T2Q16
73	26	34	.01	.43	1.29	1.29	1.44	.98	.08	-.33	67.6	77.7	T3Q4
33	18	30	.06	.40	1.19	1.29	1.31	1.61	.10	-.35	63.3	66.5	T1Q33
13	19	30	-.10	.40	1.17	1.10	1.21	1.02	.14	-.35	66.7	68.0	T1Q2
98	9	34	2.67	.43	1.29	1.25	1.64	1.69	.15	.43	70.6	77.5	T3Q40
97	21	34	.83	.39	1.30	1.94	1.26	.93	.15	-.39	47.1	68.5	T3Q39
38	3	30	2.95	.63	1.05	.26	1.01	.24	.16	-.21	90.0	90.1	T1Q38
77	27	34	-.18	.45	1.19	.83	1.15	.44	.17	-.31	73.5	79.9	T3Q8

Item Fit Statistics: Correlation Order (Continue)

77	27	34	-.18	.45 1.19	.83 1.15	.44	.17	.31	73.5	79.9	T3Q8
57	27	30	-1.52	.63 1.00	.16 1.20	.51	.18	.21	89.7	89.6	T2Q17
75	24	34	.36	.41 1.15	.89 1.47	1.20	.19	.36	70.6	73.6	T3Q6
49	28	30	-1.99	.75 .91	.05 1.09	.41	.20	.18	93.1	93.1	T2Q9
85	23	34	.52	.40 1.17	1.05 1.30	.91	.20	.37	70.6	71.7	T3Q16
37	11	30	1.14	.40 1.13	.86 1.10	.54	.20	.34	63.3	68.4	T1Q37
61	24	30	-.63	.48 1.11	.49 1.06	.29	.21	.29	75.9	79.8	T2Q32
39	5	30	2.32	.51 1.08	.37 .93	.02	.21	.26	83.3	83.5	T1Q39
71	26	34	.01	.43 1.14	.67 1.13	.43	.22	.33	73.5	77.7	T3Q2
93	25	34	-.19	.42 1.18	.95 1.02	.19	.23	.34	67.6	75.6	T3Q35
109	27	30	-.97	.68 1.36	.86 .88	.21	.24	.39	86.7	90.3	T4Q11
121	8	30	3.10	.45 1.14	.72 1.17	.49	.26	.38	76.7	76.2	T4Q34
22	8	30	1.67	.44 1.04	.26 1.05	.26	.26	.31	70.0	74.2	T1Q11
79	33	34	-2.48	1.03 .92	.22 .31	-.22	.27	.13	97.1	97.1	T3Q10
9	51	124	1.64	.21 1.24	2.64 1.32	2.27	.27	.46	57.7	71.1	CI28
25	22	30	-.63	.44 1.05	.29 1.01	.12	.28	.33	76.7	74.8	T1Q14
21	29	30	-3.15	1.02 .94	.24 .42	-.15	.28	.15	96.7	96.6	T1Q10
4	54	124	1.52	.20 1.23	2.65 1.26	1.94	.28	.46	61.0	70.6	CI23
8	113	124	-1.58	.33 1.00	.08 .69	-.60	.28	.25	91.1	91.0	CI27
74	30	34	-.92	.56 .96	.02 .85	.06	.29	.25	88.2	88.3	T3Q5
24	25	30	-1.29	.51 .96	-.03 1.14	.45	.30	.28	80.0	83.3	T1Q13
102	27	30	-.97	.68 1.15	.48 1.01	.35	.30	.39	86.7	90.3	T4Q4
65	25	30	-.88	.51 .93	-.14 1.00	.18	.30	.26	79.3	82.8	T2Q36
86	13	34	2.00	.39 1.16	.98 1.16	.71	.30	.43	58.8	71.3	T3Q17
28	14	30	.67	.39 1.04	.39 1.02	.21	.31	.35	63.3	65.1	T1Q17
76	30	34	-.92	.56 .97	.06 .69	-.18	.31	.25	88.2	88.3	T3Q7
48	29	30	-2.74	1.03 .82	.09 .24	-.34	.31	.13	96.6	96.5	T2Q8
52	29	30	-2.74	1.03 .82	.09 .24	-.34	.31	.13	96.6	96.5	T2Q12
123	27	30	-.97	.68 1.12	.40 1.00	.34	.31	.39	93.3	90.3	T4Q36
30	12	30	.98	.40 1.00	.05 1.11	.63	.31	.35	73.3	67.0	T1Q19
62	21	30	-.02	.43 1.04	.30 .95	-.07	.32	.34	65.5	71.6	T2Q33
107	28	30	-1.50	.80 .97	.14 1.05	.43	.32	.33	93.3	93.4	T4Q9
54	25	30	-.88	.51 .91	-.21 .92	.02	.32	.26	86.2	82.8	T2Q14
41	28	30	-1.99	.75 .84	-.08 .47	-.35	.32	.18	93.1	93.1	T2Q1
46	23	30	-.41	.46 .97	-.07 .98	.07	.32	.30	75.9	76.9	T2Q6
40	27	30	-1.93	.63 .98	.11 .64	-.40	.32	.23	90.0	89.9	T1Q40
70	30	34	-.92	.56 .94	-.04 .72	-.12	.32	.25	88.2	88.3	T3Q1
50	26	30	-1.17	.56 .92	-.11 .70	-.33	.32	.24	86.2	86.1	T2Q10
115	27	30	-.97	.68 1.19	.55 .85	.18	.33	.39	86.7	90.3	T4Q17
78	21	34	.83	.39 1.08	.61 1.05	.27	.33	.39	58.8	68.5	T3Q9
81	26	34	.01	.43 .97	-.07 1.06	.28	.33	.33	85.3	77.7	T3Q12
111	26	30	-.56	.61 1.14	.48 .99	.25	.36	.42	86.7	88.3	T4Q13
68	26	30	-1.17	.56 .88	-.21 .59	-.57	.36	.24	86.2	86.1	T2Q39
83	13	34	2.00	.39 1.07	.46 1.13	.61	.37	.43	64.7	71.3	T3Q14
51	15	30	.99	.40 1.06	.44 1.06	.39	.37	.41	72.4	66.1	T2Q11
63	24	30	-.63	.48 .88	-.38 .79	-.34	.37	.29	82.8	79.8	T2Q34
89	19	34	1.12	.38 .96	-.25 1.79	2.64	.37	.41	67.6	67.5	T3Q31
96	17	34	1.41	.38 1.06	.48 1.04	.24	.37	.42	64.7	67.9	T3Q38
122	19	30	1.17	.43 1.12	.65 1.14	.52	.38	.47	66.7	74.2	T4Q35
1	84	124	.26	.21 1.04	.44 .98	-.03	.39	.41	69.9	72.7	CI20
47	19	30	.34	.41 1.00	.02 .88	-.43	.39	.36	55.2	67.4	T2Q7
18	22	30	-.63	.44 .95	-.16 .87	-.37	.39	.33	76.7	74.8	T1Q7
67	21	30	-.02	.43 .91	-.49 .91	-.22	.40	.34	79.3	71.6	T2Q38
2	63	124	1.15	.20 1.07	.89 1.07	.59	.40	.45	67.5	69.3	CI21
29	22	30	-.63	.44 .91	-.38 .99	.07	.40	.33	83.3	74.8	T1Q18
44	28	30	-1.99	.75 .74	-.27 .31	-.67	.40	.18	93.1	93.1	T2Q4
60	21	30	-.02	.43 .92	-.43 .84	-.47	.40	.34	72.4	71.6	T2Q31
104	26	30	-.56	.61 1.09	.36 .78	-.02	.41	.42	86.7	88.3	T4Q6
45	20	30	.16	.42 .86	-.85 1.19	.76	.41	.35	82.8	69.3	T2Q5
72	32	34	-1.73	.75 .80	-.15 .30	-.62	.41	.18	94.1	94.2	T3Q3
101	26	30	-.56	.61 1.12	.43 .67	-.19	.41	.42	86.7	88.3	T4Q3
110	29	30	-2.33	1.07 .82	.06 .16	-.49	.42	.25	96.7	96.7	T4Q12
113	29	30	-2.33	1.07 .82	.06 .16	-.49	.42	.25	96.7	96.7	T4Q15
94	26	34	.01	.43 .94	-.23 .72	-.50	.42	.33	73.5	77.7	T3Q36
55	27	30	-1.52	.63 .75	-.47 .41	-.76	.42	.21	89.7	89.6	T2Q15
90	26	34	.01	.43 .91	-.38 .77	-.37	.43	.33	79.4	77.7	T3Q32
64	12	30	1.50	.42 .98	-.06 1.11	.56	.43	.44	79.3	70.5	T2Q35
114	21	30	.78	.46 1.01	.11 1.25	.74	.43	.47	80.0	77.7	T4Q16
23	18	30	.06	.40 .93	-.43 .91	-.47	.43	.35	63.3	66.5	T1Q12
80	15	34	1.70	.38 1.00	.05 .91	-.33	.45	.43	64.7	69.2	T3Q11
7	83	124	.31	.21 .96	-.36 .86	-.74	.45	.41	68.3	72.3	CI26
87	23	34	.52	.40 .94	-.34 .78	-.57	.45	.37	64.7	71.7	T3Q18
126	20	30	.98	.44 1.01	.12 1.04	.23	.45	.47	76.7	75.9	T4Q39
58	14	30	1.16	.41 .96	-.23 .93	-.31	.45	.42	72.4	67.3	T2Q18
17	26	30	-1.58	.56 .85	-.30 .62	-.64	.46	.26	86.7	86.5	T1Q6
6	95	124	-.28	.23 .88	-1.03 .78	-.88	.46	.37	80.5	78.3	CI25
99	26	30	-.56	.61 .94	-.02 .85	.08	.47	.42	86.7	88.3	T4Q1
82	16	34	1.56	.38 .95	-.34 .94	-.21	.47	.42	70.6	68.5	T3Q13
88	10	34	2.49	.42 .95	-.20 .90	-.23	.47	.43	76.5	75.7	T3Q19
16	20	30	-.27	.41 .90	-.55 .80	-.86	.48	.34	66.7	70.2	T1Q5
36	18	30	.06	.40 .89	-.73 .84	-.84	.48	.35	70.0	66.5	T1Q36
26	23	30	-.83	.45 .86	-.53 .74	-.74	.49	.32	76.7	77.4	T1Q15
5	61	124	1.23	.20 .94	-.69 .95	-.40	.49	.45	74.8	69.5	CI24
84	29	34	-.64	.51 .78	-.63 .49	-.71	.49	.27	85.3	85.4	T3Q15

Item Fit Statistics: Correlation Order (Continue)

84	29	34	-.64	.51	.78	-.63	.49	-.71	.49	.27	85.3	85.4	T3Q15
119	17	30	1.53	.42	.94	-.30	.83	-.52	.51	.46	73.3	71.1	T4Q32
95	24	34	.36	.41	.83	-.97	.70	-.74	.52	.36	82.4	73.6	T3Q37
66	10	30	1.86	.44	.91	-.39	.90	-.27	.52	.46	75.9	73.8	T2Q37
31	25	30	-1.29	.51	.80	-.58	.63	-.79	.53	.28	86.7	83.3	T1Q31
32	18	30	.06	.40	.85	-1.09	.83	-.96	.53	.35	76.7	66.5	T1Q32
14	23	30	-.83	.45	.83	-.68	.69	-.91	.53	.32	76.7	77.4	T1Q3
103	28	30	-1.50	.80	.75	-.31	.23	-.61	.54	.33	93.3	93.4	T4Q5
11	67	124	.98	.20	.86	-1.82	.80	-1.65	.55	.44	74.0	69.2	CI30
53	13	30	1.33	.41	.79	-1.36	.94	-.22	.55	.43	82.8	68.7	T2Q13
10	65	124	1.07	.20	.87	-1.78	.78	-1.86	.56	.45	71.5	69.3	CI29
92	18	34	1.27	.38	.82	-1.38	.73	-1.19	.58	.41	73.5	67.5	T3Q34
125	21	30	.78	.46	.85	-.56	.71	-.69	.59	.47	80.0	77.7	T4Q38
112	27	30	-.97	.68	.69	-.62	.37	-.52	.59	.39	93.3	90.3	T4Q14
42	10	30	1.86	.44	.81	-.90	.71	-1.04	.60	.46	82.8	73.8	T2Q2
3	44	124	1.95	.21	.81	-2.21	.69	-2.26	.61	.46	77.2	72.8	CI22
105	25	30	-.22	.56	.81	-.45	.46	-.79	.61	.44	83.3	86.0	T4Q7
100	26	30	-.56	.61	.71	-.70	.42	-.65	.63	.42	93.3	88.3	T4Q2
91	21	34	.83	.39	.72	-2.08	.62	-1.41	.63	.39	82.4	68.5	T3Q33
27	25	30	-1.29	.51	.71	-.91	.49	-1.22	.64	.28	86.7	83.3	T1Q16
108	25	30	-.22	.56	.70	-.83	.40	-.93	.67	.44	90.0	86.0	T4Q10
118	24	30	.07	.52	.69	-.98	.52	-.82	.67	.45	90.0	83.9	T4Q31
116	27	30	-.97	.68	.51	-1.16	.20	-.92	.70	.39	93.3	90.3	T4Q18
120	24	30	.07	.52	.50	-1.80	.32	-1.46	.79	.45	90.0	83.9	T4Q33
MEAN	26.1	39.1	-.09	.53	.99	.1	.99	.1			78.4	79.0	
P.SD	16.3	26.2	1.50	.27	.18	.9	.65	1.0			11.6	9.4	

d. Item Fit Statistics: Misfit Order

TABLE 10.1 concurrent analysis NEW LABEL TO ITEM ZOU322WS.TXT Feb 15 2022 13:30
 INPUT: 124 PERSON 127 ITEM REPORTED: 124 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7
 PERSON: REAL SEP.: 2.27 REL.: .84 ... ITEM: REAL SEP.: 2.24 REL.: .83

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	OUTFIT MNSQ	PTMEASUR-AL CORR.	EXACT EXP%	MATCH EXP%	ITEM			
117	6	30	3.54	.49	1.60	1.91	5.17	3.71	A-.31	.35	76.7	81.8	T4Q19
124	9	30	2.90	.44	1.36	1.83	3.89	4.07	B-.10	.39	73.3	73.6	T4Q37
69	27	30	-1.52	.63	1.13	.43	3.69	2.41	C-.03	.21	89.7	89.6	T2Q40
59	24	30	-.63	.48	1.20	.79	2.46	2.47	D .02	.29	82.8	79.8	T2Q19
56	27	30	-1.52	.63	1.09	.35	2.09	1.35	E .07	.21	89.7	89.6	T2Q16
127	15	30	1.87	.41	1.57	3.45	2.01	2.86	F-.01	.44	46.7	68.0	T4Q40
89	19	34	1.12	.38	.96	-.25	1.79	2.64	G .37	.41	67.6	67.5	T3Q31
98	9	34	2.67	.43	1.29	1.25	1.64	1.69	H .15	.43	70.6	77.5	T3Q40
34	24	30	-1.05	.48	1.18	.71	1.63	1.48	I .02	.30	76.7	80.3	T1Q34
75	24	34	.36	.41	1.15	.89	1.47	1.20	J .19	.36	70.6	73.6	T3Q6
20	12	30	.98	.40	1.27	1.81	1.44	2.10	K .00	.35	53.3	67.0	T1Q9
73	26	34	.01	.43	1.29	1.29	1.44	.98	L .08	.33	67.6	77.7	T3Q4
19	19	30	-.10	.40	1.21	1.32	1.38	1.75	M .07	.35	60.0	68.0	T1Q8
12	10	30	1.31	.41	1.27	1.53	1.36	1.45	N .01	.33	53.3	70.0	T1Q1
109	27	30	-.97	.68	1.36	.86	.88	.21	O .24	.39	86.7	90.3	T4Q11
9	51	124	1.64	.21	1.24	2.64	1.32	2.27	P .27	.46	57.7	71.1	CI28
33	18	30	.06	.40	1.19	1.29	1.31	1.61	Q .10	.35	63.3	66.5	T1Q33
85	23	34	.52	.40	1.17	1.05	1.30	.91	R .20	.37	70.6	71.7	T3Q16
97	21	34	.83	.39	1.30	1.94	1.26	.93	S .15	.39	47.1	68.5	T3Q39
4	54	124	1.52	.20	1.23	2.65	1.26	1.94	T .28	.46	61.0	70.6	CI23
114	21	30	.78	.46	1.01	.11	1.25	.74	U .43	.47	80.0	77.7	T4Q16
15	29	30	-3.15	1.02	1.06	.36	1.24	.60	V .02	.15	96.7	96.6	T1Q4
13	19	30	-.10	.40	1.17	1.10	1.21	1.02	W .14	.35	66.7	68.0	T1Q2
57	27	30	-1.52	.63	1.00	.16	1.20	.51	X .18	.21	89.7	89.6	T2Q17
45	20	30	.16	.42	.86	-.85	1.19	.76	Y .41	.35	82.8	69.3	T2Q5
77	27	34	-.18	.45	1.19	.83	1.15	.44	Z .17	.31	73.5	79.9	T3Q8
101	26	30	-.56	.61	1.12	.43	.67	-.19	.41	.42	86.7	88.3	T4Q3
104	26	30	-.56	.61	1.09	.36	.78	-.02	.41	.42	86.7	88.3	T4Q6
8	113	124	-1.58	.33	1.00	.08	.69	-.60	.28	.25	91.1	91.0	CI27
40	27	30	-1.93	.63	.98	.11	.64	-.40	.32	.23	90.0	89.9	T1Q40
76	30	34	-.92	.56	.97	.06	.69	-.18	.31	.25	88.2	88.3	T3Q7
21	29	30	-3.15	1.02	.94	.24	.42	-.15	.28	.15	96.7	96.6	T1Q10
BETTER FITTING NOT SHOWN													
53	13	30	1.33	.41	.79	-1.36	.94	-.22	.55	.43	82.8	68.7	T2Q13
70	30	34	-.92	.56	.94	-.04	.72	-.12	.32	.25	88.2	88.3	T3Q1
87	23	34	.52	.40	.94	-.34	.78	-.57	.45	.37	64.7	71.7	T3Q18
94	26	34	.01	.43	.94	-.23	.72	-.50	.42	.33	73.5	77.7	T3Q36
50	26	30	-1.17	.56	.92	-.11	.70	-.33	.32	.24	86.2	86.1	T2Q10
79	33	34	-2.48	1.03	.92	.22	.31	-.22	.27	.13	97.1	97.1	T3Q10
90	26	34	.01	.43	.91	-.38	.77	-.37	.43	.33	79.4	77.7	T3Q32
6	95	124	-.28	.23	.88	-1.03	.78	-.88	.46	.37	80.5	78.3	CI25
63	24	30	-.63	.48	.88	-.38	.79	-.34	.37	.29	82.8	79.8	T2Q34
68	26	30	-1.17	.56	.88	-.21	.59	-.57	.36	.24	86.2	86.1	T2Q39
10	65	124	1.07	.20	.87	-1.78	.78	-1.86	.56	.45	71.5	69.3	CI29

Item Fit Statistics: Misfit Order (Continue)

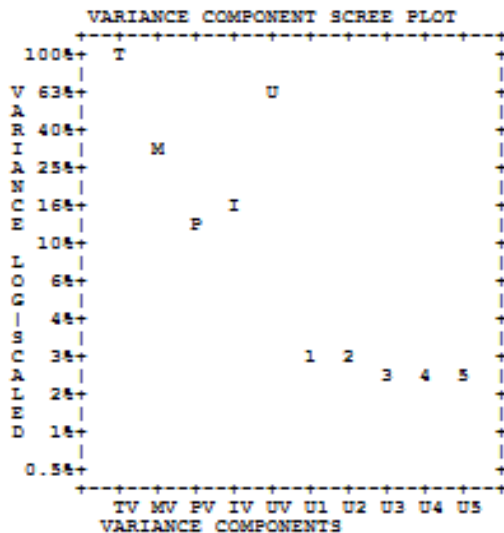
10	65	124	1.07	.20	.87	-1.78	.78	-1.86	.56	-.45	71.5	69.3	CI29
11	67	124	.98	.20	.86	-1.82	.80	-1.65	.55	-.44	74.0	69.2	CI30
26	23	30	-.83	.45	.86	-.53	.74	-.74	.49	-.32	76.7	77.4	T1Q15
17	26	30	-1.58	.56	.85	-.30	.62	-.64	.46	-.26	86.7	86.5	T1Q6
125	21	30	.78	.46	.85	-.56	.71	-.69	.59	-.47	80.0	77.7	T4Q38
41	28	30	-1.99	.75	.84	-.08	.47	-.35	.32	-.18	93.1	93.1	T2Q1
14	23	30	-.83	.45	.83	-.68	.69	-.91	.53	-.32	76.7	77.4	T1Q3
95	24	34	.36	.41	.83	-.97	.70	-.74	.52	-.36	82.4	73.6	T3Q37
48	29	30	-2.74	1.03	.82	-.09	.24	-.34	.31	-.13	96.6	96.5	T2Q8
52	29	30	-2.74	1.03	.82	-.09	.24	-.34	.31	-.13	96.6	96.5	T2Q12
92	18	34	1.27	.38	.82	-1.38	.73	-1.19	.58	-.41	73.5	67.5	T3Q34
110	29	30	-2.33	1.07	.82	-.06	.16	-.49	.42	-.25	96.7	96.7	T4Q12
113	29	30	-2.33	1.07	.82	-.06	.16	-.49	.42	-.25	96.7	96.7	T4Q15
3	44	124	1.95	.21	.81	-2.21	.69	-2.26	.61	-.46	77.2	72.8	CI22
42	10	30	1.86	.44	.81	-.90	.71	-1.04	.60	-.46	82.8	73.8	T2Q2
105	25	30	-.22	.56	.81	-.45	.46	-.79	.61	-.44	83.3	86.0	T4Q7
31	25	30	-1.29	.51	.80	-.58	.63	-.79	.53	-.28	86.7	83.3	T1Q31
72	32	34	-1.73	.75	.80	-.15	.30	-.62	.41	-.18	94.1	94.2	T3Q3
84	29	34	-.64	.51	.78	-.63	.49	-.71	.49	-.27	85.3	85.4	T3Q15
55	27	30	-1.52	.63	.75	-.47	.41	-.76	.42	-.21	89.7	89.6	T2Q15
103	28	30	-1.50	.80	.75	-.31	.23	-.61	.54	-.33	93.3	93.4	T4Q5
44	28	30	-1.99	.75	.74	-.27	.31	-.67	.40	-.18	93.1	93.1	T2Q4
91	21	34	.83	.39	.72	-2.08	.62	-1.41	.63	-.39	82.4	68.5	T3Q33
27	25	30	-1.29	.51	.71	-.91	.49	-1.22	.64	-.28	86.7	83.3	T1Q16
100	26	30	-.56	.61	.71	-.70	.42	-.65	.63	-.42	93.3	88.3	T4Q2
108	25	30	-.22	.56	.70	-.83	.40	-.93	.67	-.44	90.0	86.0	T4Q10
112	27	30	-.97	.68	.69	-.62	.37	-.52	.59	-.39	93.3	90.3	T4Q14
118	24	30	.07	.52	.69	-.98	.52	-.82	.67	-.45	90.0	83.9	T4Q31
116	27	30	-.97	.68	.51	-1.16	.20	-.92	.70	-.39	93.3	90.3	T4Q18
120	24	30	.07	.52	.50	-1.80	.32	-1.46	.79	-.45	90.0	83.9	T4Q33
MEAN	26.1	39.1	-.09	.53	.99	.1	.99	.1			78.4	79.0	
P.SD	16.3	26.2	1.50	.27	.18	.9	.65	1.0			11.6	9.4	

e. Dimensionality Map

TABLE 23.0 concurrent analysis NEW LABEL TO ITEM ZOU322WS.TXT Feb 15 2022 13:30
 INPUT: 124 PERSON 127 ITEM REPORTED: 124 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7

	Eigenvalue	Observed	Expected
Total raw variance in observations	183.0026	100.0%	100.0%
Raw variance explained by measures	59.0026	32.2%	32.4%
Raw variance explained by persons	25.2448	13.8%	13.9%
Raw Variance explained by items	33.7578	18.4%	18.5%
Raw unexplained variance (total)	124.0000	67.8%	100.0%
Unexplained variance in 1st contrast	5.2284	2.9%	4.2%
Unexplained variance in 2nd contrast	5.0874	2.8%	4.1%
Unexplained variance in 3rd contrast	4.3359	2.4%	3.5%
Unexplained variance in 4th contrast	4.2180	2.3%	3.4%
Unexplained variance in 5th contrast	4.0967	2.2%	3.3%

STANDARDIZED RESIDUAL VARIANCE SCREE PLOT



APPENDIX G

STATISTICS FOR FINAL DATA

1. Validity of Final Data

a. Summary Statistics of 127 Measured Items and 124 Measured Persons

TABLE 3.1 Pdeleted Concurrent final data analysi ZOU889WS.TXT Feb 17 2022 11:26LLS with C
INPUT: 902 PERSON 127 ITEM REPORTED: 902 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7

SUMMARY OF 902 MEASURED PERSON

	TOTAL		MEASURE	MODEL S.E.	INFIT		OUTFIT	
	SCORE	COUNT			MNSQ	ZSTD	MNSQ	ZSTD
MEAN	24.5	40.0	.77	.38	1.00	.02	.99	.03
SEM	.2	.0	.03	.00	.00	.03	.01	.03
P.SD	6.3	.0	.91	.08	.15	1.02	.29	1.01
S.SD	6.3	.0	.91	.08	.15	1.02	.29	1.01
MAX.	39.0	40.0	4.23	1.03	1.56	3.54	3.00	3.89
MIN.	7.0	40.0	-1.52	.34	.62	-3.36	.18	-3.01

REAL RMSE	.40	TRUE SD	.82	SEPARATION	2.05	PERSON RELIABILITY	.81	
MODEL RMSE	.39	TRUE SD	.83	SEPARATION	2.11	PERSON RELIABILITY	.82	
S.E. OF PERSON MEAN = .03								

PERSON RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .37 SEM = 5.02

SUMMARY OF 127 MEASURED ITEM

	TOTAL		MEASURE	MODEL S.E.	INFIT		OUTFIT	
	SCORE	COUNT			MNSQ	ZSTD	MNSQ	ZSTD
MEAN	173.7	284.1	.00	.16	.99	-.06	.98	-.08
SEM	9.7	17.2	.09	.00	.01	.16	.02	.15
P.SD	109.0	193.1	1.06	.05	.10	1.76	.19	1.71
S.SD	109.4	193.9	1.06	.05	.10	1.77	.20	1.72
MAX.	693.0	902.0	2.66	.42	1.37	5.68	1.76	5.29
MIN.	27.0	179.0	-3.24	.07	.78	-5.36	.48	-4.38

REAL RMSE	.17	TRUE SD	1.05	SEPARATION	6.02	ITEM RELIABILITY	.97	
MODEL RMSE	.17	TRUE SD	1.05	SEPARATION	6.11	ITEM RELIABILITY	.97	
S.E. OF ITEM MEAN = .09								

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.22
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

b. Dimensionality Map

TABLE 23.0 Pdeleted Concurrent final data analys ZOU889WS.TXT Feb 17 2022 11:26ILLS with C
INPUT: 902 PERSON 127 ITEM REPORTED: 902 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

		Eigenvalue	Observed	Expected
Total raw variance in observations	=	166.8381	100.0%	100.0%
Raw variance explained by measures	=	39.8381	23.9%	24.0%
Raw variance explained by persons	=	18.3713	11.0%	11.1%
Raw Variance explained by items	=	21.4668	12.9%	12.9%
Raw unexplained variance (total)	=	127.0000	76.1%	100.0%
Unexplnd variance in 1st contrast	=	3.7745	2.3%	3.0%
Unexplnd variance in 2nd contrast	=	2.6759	1.6%	2.1%
Unexplnd variance in 3rd contrast	=	2.2488	1.3%	1.8%

c. Item Fit Statistics: Misfit Order of 127 items

TABLE 10.1 Deleted Concurrent final data analysis ZOU889WS.TXT Feb 17 2022 11:26ILLS with C
 INPUT: 902 PERSON 127 ITEM REPORTED: 902 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7

PERSON: REAL SEP.: 2.05 REL.: .81 ... ITEM: REAL SEP.: 6.02 REL.: .97

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	OUTFIT ZSTD MNSQ	PTMEASUR-AL ZSTD CORR.	EXP.	OBS%	EXP%	ITEM	
117	42	179	2.36	.19	1.36	3.37 1.76	3.89 A-.02	.37	73.7	77.8	T4Q19 B2 CTLS	
127	115	179	.23	.17	1.37	4.71 1.60	4.70 B .02	.40	59.8	70.5	T4Q40 C1 I	
115	160	179	-1.53	.25	1.07	.45 1.53	1.56 C .14	.25	89.4	89.4	T4Q17 B2 LA	
40	169	208	-1.10	.18	1.12	1.10 1.49	2.65 D .06	.26	81.3	81.2	T1Q40 C2 CTLS	
124	61	179	1.74	.17	1.26	3.33 1.36	2.95 E .16	.41	59.2	71.3	T4Q37 C1 EMM	
85	175	268	.10	.14	1.09	1.70 1.31	2.80 F .25	.35	67.2	69.6	T3Q16 C1 EPM	
75	195	268	-.30	.15	1.07	1.07 1.29	2.09 G .22	.32	73.9	74.1	T3Q6 B1 SP	
45	178	247	-.36	.15	1.05	.83 1.26	2.20 H .20	.31	74.5	73.4	T2Q5 B2 WR	
59	147	247	.27	.14	1.12	2.59 1.26	3.28 I .18	.34	64.0	65.9	T2Q19 B2 EPM	
93	161	268	.36	.13	1.14	2.80 1.26	2.78 J .23	.37	59.3	67.4	T3Q35 C2 I	
66	81	247	1.54	.15	1.14	2.20 1.25	2.77 K .18	.36	69.2	71.6	T2Q37 C2 I	
98	92	268	1.61	.14	1.13	2.12 1.25	2.76 L .25	.40	69.4	71.6	T3Q40 C2 EMM	
121	59	179	1.80	.17	1.22	2.79 1.23	1.90 M .21	.40	59.2	71.8	T4Q34 C1 EMM	
97	115	268	1.18	.13	1.16	3.18 1.21	2.78 N .24	.39	61.2	67.4	T3Q39 C2 EMM	
9	391	902	1.06	.07	1.15	5.68 1.20	5.29 O .22	.38	61.8	67.0	CI28 C1 CTLS	
15	202	208	-3.24	.42	1.04	.22 1.20	.54 P .02	.12	97.1	97.1	T1Q4 B1 SP	
35	172	208	-1.20	.19	1.06	.56 1.17	1.01 Q .15	.26	82.2	82.6	T1Q35 C2 EPM	
64	109	247	.98	.14	1.15	3.25 1.17	2.64 R .18	.36	57.9	65.4	T2Q35 C2 I	
73	213	268	-.71	.16	1.14	1.53 1.17	1.03 S .16	.29	78.7	79.9	T3Q4 B1 EPM	
86	109	268	1.29	.14	1.15	2.81 1.17	2.20 T .25	.40	61.9	68.2	T3Q17 C1 EPM	
12	96	208	.71	.15	1.11	2.44 1.15	2.35 U .19	.34	60.6	64.1	T1Q1 B1 SP	
56	172	247	-.23	.15	1.02	.32 1.15	1.40 V .27	.31	70.0	71.6	T2Q16 B2 EMM	
4	405	902	.99	.07	1.11	4.24 1.14	3.87 W .26	.38	61.5	66.5	CI23 C1 EPM	
105	111	179	.35	.17	1.08	1.15 1.14	1.31 X .33	.41	64.2	69.8	T4Q7 B2 SP	
109	166	179	-1.97	.30	1.01	.11 1.14	.47 Y .19	.21	92.7	92.7	T4Q11 B2 EPM	
46	201	247	-.94	.17	1.04	.39 1.13	.84 Z .20	.26	81.8	81.5	T2Q6 B2 SP	
51	148	247	.25	.14	1.13	2.81 1.09	1.23	.20	.34	54.7	66.0	T2Q11 B2 WR
58	132	247	.56	.14	1.12	2.81 1.11	1.72	.22	.35	52.2	64.4	T2Q18 B2 EMM
2	446	902	.78	.07	1.10	3.89 1.11	3.17	.28	.38	59.4	65.7	CI21 C1 LA
BETTER FITTING NOT SHOWN												
67	144	247	.33	.14	.90	-2.19 .95	-.63 .44	.34	72.9	65.4	T2Q38 C2 CITR	
110	159	179	-1.47	.25	.95	-.26 .65	-1.23 .35	.26	88.8	88.8	T4Q12 B2 I	
7	540	902	.29	.07	.94	-2.39 .91	-2.24 .43	.36	70.4	66.9	CI26 C1 EMM	
72	253	268	-2.29	.27	.94	-.22 .79	-.51 .25	.17	94.4	94.4	T3Q3 B1 WR	
49	218	247	-1.52	.20	.93	-.47 .74	-1.15 .33	.22	88.3	88.2	T2Q9 B2 SP	
79	247	268	-1.92	.23	.93	-.33 .72	-.90 .29	.20	92.2	92.2	T3Q10 B1 SP	
17	183	208	-1.66	.22	.92	-.49 .76	-1.08 .34	.22	88.0	87.9	T1Q6 B1 SP	
55	209	247	-1.19	.18	.92	-.70 .72	-1.59 .37	.25	84.2	84.6	T2Q15 B2 EPM	
27	179	208	-1.48	.21	.91	-.63 .79	-1.02 .35	.24	86.1	86.0	T1Q16 C1 I	
32	117	208	.25	.15	.91	-2.05 .87	-2.00 .45	.33	70.7	64.4	T1Q32 C2 CTLS	
57	218	247	-1.52	.20	.91	-.59 .72	-1.29 .35	.22	88.3	88.2	T2Q17 B2 EMM	
61	157	247	.08	.14	.91	-1.88 .84	-1.97 .45	.33	68.8	67.8	T2Q32 C2 EMM	
71	189	268	-.17	.14	.91	-1.46 .89	-.87 .42	.33	74.3	72.6	T3Q2 B1 SP	
76	219	268	-.87	.17	.91	-.88 .82	-1.00 .37	.28	82.1	82.0	T3Q7 B1 SP	
100	140	179	-.58	.19	.91	-.89 .77	-1.30 .44	.34	81.0	78.7	T4Q2 B2 SP	

Item Fit Statistics: Misfit Order (continue)

	100	140	179	-.58	.19	.91	-.89	.77	-1.30	v	.44	.34	81.0	78.7	T4Q2	B2	SP	
	101	136	179	-.44	.19	.91	-.95	.75	-1.56	u	.46	.35	76.5	76.8	T4Q3	B2	SP	
	119	90	179	.93	.17	.91	-1.46	.86	-1.76	t	.51	.42	71.5	68.9	T4Q32	C1	EMM	
	14	170	208	-1.13	.19	.90	-.92	.74	-1.66	s	.41	.26	80.8	81.7	T1Q3	B1	SP	
	26	132	208	-.09	.15	.90	-1.79	.86	-1.79	r	.44	.32	70.2	67.5	T1Q15	C1	WR	
	48	170	247	-.18	.15	.90	-1.63	.85	-1.58	q	.43	.32	73.3	71.1	T2Q8	B2	EPM	
	43	168	247	-.14	.14	.89	-1.88	.82	-1.95	p	.45	.32	72.9	70.5	T2Q3	B2	WR	
	52	163	247	-.04	.14	.89	-2.17	.86	-1.64	o	.45	.32	75.7	69.2	T2Q12	B2	EPM	
	54	173	247	-.25	.15	.89	-1.90	.80	-2.00	n	.45	.31	75.3	71.9	T2Q14	B2	I	
	65	143	247	.35	.14	.89	-2.62	.87	-2.00	m	.47	.34	73.3	65.3	T2Q36	C2	I	
	116	166	179	-1.97	.30	.89	-.44	.48	-1.55	l	.37	.21	92.7	92.7	T4Q18	B2	EMM	
	122	87	179	1.01	.17	.89	-1.73	.87	-1.64	k	.52	.42	71.5	68.9	T4Q35	C1	EMM	
	44	196	247	-.80	.16	.87	-1.48	.73	-1.96	j	.44	.28	79.8	79.5	T2Q4	B2	SP	
	69	133	247	.54	.14	.87	-3.31	.83	-2.76	i	.50	.35	75.3	64.4	T2Q40	C2	SP	
	74	219	268	-.87	.17	.87	-1.42	.70	-1.73	h	.43	.28	82.1	82.0	T3Q5	B1	SP	
	113	128	179	-.17	.18	.86	-1.78	.85	-1.08	g	.51	.38	78.2	73.8	T4Q15	B2	EPM	
	41	196	247	-.80	.16	.84	-1.83	.70	-2.19	f	.47	.28	78.9	79.5	T2Q1	B2	EPM	
	63	134	247	.52	.14	.84	-4.08	.80	-3.37	e	.53	.35	75.7	64.4	T2Q34	C2	EMM	
	112	139	179	-.54	.19	.84	-1.70	.76	-1.44	d	.50	.34	80.4	78.2	T4Q14	B2	EPM	
	118	109	179	.40	.17	.83	-2.73	.76	-2.66	c	.58	.41	78.8	69.5	T4Q31	C1	EMM	
	60	133	247	.54	.14	.79	-5.36	.74	-4.38	b	.59	.35	84.2	64.4	T2Q31	C2	EMM	
	120	99	179	.68	.17	.78	-3.69	.74	-3.25	a	.62	.42	79.3	68.8	T4Q33	C1	EMM	

	MEAN	173.7	284.1	.00	.16	.99	-.1	.98	-.1				73.8	73.7				
	P.SD	109.0	193.1	1.06	.05	.10	1.8	.19	1.7				9.4	8.5				

d. Person Fit Statistics of 902 students

TABLE 6.1 Pdeleted Concurrent final data analysis ZOU889WS.TXT Feb 17 2022 11:26LLS with C INPUT: 902 PERSON 127 ITEM REPORTED: 902 PERSON 127 ITEM 2 CATS WINSTEPS 4.4.7

PERSON: REAL SEP.: 2.05 REL.: .81 ... ITEM: REAL SEP.: 6.02 REL.: .97

PERSON STATISTICS: MISFIT ORDER

ENTRY	TOTAL	TOTAL		MODEL	INFIT	OUTFIT	PTMEASUR	AL	EXACT	MATCH									
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	PERSON						
	716	39	40	4.23	1.02	1.06	.37	3.00	1.48	A	-.15	.10	97.5	97.5	FAS261				
	871	34	40	2.14	.47	1.16	.63	2.98	2.56	B	.05	.33	82.5	85.3	FE148				
	688	32	40	1.85	.41	1.06	.33	2.44	2.78	C	.08	.27	80.0	79.9	FAS233				
	199	25	40	.70	.37	1.10	.71	2.30	3.09	D	.28	.44	75.0	71.4	FAC199				
	192	25	40	.70	.37	1.12	.85	2.29	3.08	E	.27	.44	70.0	71.4	FAC192				
	844	36	40	2.66	.55	1.18	.57	2.26	1.52	F	.06	.28	90.0	90.0	FE121				
	164	25	40	.70	.37	1.09	.65	2.19	2.88	G	.31	.44	70.0	71.4	FAC164				
	727	14	40	-.74	.37	1.51	2.73	2.05	3.30	H	-.04	.44	62.5	72.7	FE004				
	177	22	40	.31	.36	1.18	1.26	2.00	2.95	I	.26	.45	65.0	69.7	FAC177				
	143	24	40	.57	.36	1.03	.28	1.97	2.62	J	.36	.44	70.0	70.8	FAC143				
	57	23	40	.44	.36	.96	-.24	1.85	2.49	K	.40	.45	82.5	70.2	FAC057				
	130	21	40	.18	.36	1.24	1.64	1.82	2.60	L	.25	.46	57.5	69.6	FAC130				
	678	22	40	.52	.34	1.30	2.47	1.80	3.89	M	-.04	.36	60.0	65.9	FAS223				
	725	15	40	-.60	.37	1.38	2.18	1.80	2.85	N	.09	.44	65.0	71.7	FE002				
	51	22	40	.31	.36	1.08	.59	1.74	2.32	O	.35	.45	65.0	69.7	FAC051				
	186	21	40	.18	.36	1.40	2.65	1.73	2.38	P	.14	.46	57.5	69.6	FAC186				
	804	16	40	-.47	.36	1.56	3.16	1.72	2.79	Q	-.02	.45	50.0	71.0	FE081				
	121	21	40	.18	.36	1.09	.71	1.71	2.32	R	.34	.46	72.5	69.6	FAC121				
	768	28	40	1.09	.38	1.21	1.20	1.71	2.07	S	.18	.41	67.5	73.8	FE045				
	532	16	40	-.18	.35	1.51	3.37	1.70	3.44	T	-.19	.38	52.5	68.0	FAS077				
	866	33	40	1.93	.45	1.16	.67	1.66	1.29	U	.15	.35	85.0	83.0	FE143				
	882	29	40	1.24	.39	1.09	.55	1.66	1.80	V	.26	.40	75.0	75.5	FE159				
	227	14	40	-.53	.35	1.45	2.82	1.63	2.99	W	-.22	.35	47.5	69.5	EMC019				
	854	12	40	-1.02	.38	1.52	2.57	1.62	1.84	X	.00	.42	57.5	74.9	FE131				
	426	17	40	-.16	.34	1.48	3.54	1.61	3.63	Y	-.25	.35	47.5	66.4	EMC218				
	599	37	40	3.05	.61	1.10	.36	1.61	.92	Z	-.03	.17	92.5	92.5	FAS144				
	166	18	40	-.20	.36	1.32	2.09	1.59	2.06		.22	.46	55.0	70.5	FAC166				
	813	27	40	.95	.37	1.38	2.18	1.59	1.92		.08	.42	60.0	72.4	FE090				

Person Fit Statistics of 902 students (continue)

813	27	40	.95	.37 1.38	2.18 1.59	1.92	.08	.42	60.0	72.4	FE090
474	9	40	-1.15	.41 1.25	1.08 1.57	1.64	.07	.36	72.5	79.8	FAS019
597	39	40	4.23	1.02 1.04	.36 1.57	.80	-.03	.10	97.5	97.5	FAS142
189	32	40	1.78	.43 1.40	1.56 1.56	1.02	.09	.37	75.0	82.0	FAC189
283	35	40	2.33	.49 1.19	.64 1.56	1.13	-.09	.22	87.5	87.5	FMC075
182	23	40	.44	.36 1.23	1.55 1.55	1.75	.24	.45	67.5	70.2	FAC182
694	13	40	-.56	.37 1.40	2.28 1.55	2.27	-.06	.38	62.5	71.8	FAS239
831	19	40	-.09	.36 1.40	2.48 1.55	2.52	.12	.45	50.0	70.4	FE108
837	19	40	-.09	.36 1.29	1.87 1.55	2.50	.18	.45	60.0	70.4	FE114
777	30	40	1.40	.40 1.32	1.55 1.54	1.42	.10	.39	72.5	77.1	FE054
785	33	40	1.93	.45 1.17	.73 1.53	1.09	.16	.35	80.0	83.0	FE062
796	16	40	-.47	.36 1.37	2.19 1.52	2.14	.13	.45	55.0	71.0	FE073
202	36	40	2.74	.56 1.00	.13 1.51	.80	.26	.29	90.0	90.0	FAC202
155	13	40	-.89	.38 1.33	1.73 1.50	1.47	.22	.46	67.5	74.8	FAC155
549	29	40	1.40	.37	-.15 1.50	1.57	.26	.31	75.0	72.7	FAS094
261	24	40	.64	.34 1.31	2.46 1.49	2.70	-.08	.34	62.5	65.9	FMC053
408	14	40	-.53	.35 1.32	2.08 1.49	2.41	-.07	.35	62.5	69.5	FMC200
420	15	40	-.40	.35 1.40	2.71 1.49	2.64	-.15	.35	55.0	68.3	FMC212
185	23	40	.44	.36 1.33	2.15 1.47	1.53	.20	.45	62.5	70.2	FAC185
191	26	40	.84	.37 1.23	1.42 1.47	1.31	.24	.43	65.0	72.3	FAC191
790	19	40	-.09	.36 1.37	2.30 1.47	2.21	.15	.45	50.0	70.4	FE067
828	19	40	-.09	.36 1.24	1.59 1.47	2.19	.22	.45	65.0	70.4	FE105
201	28	40	1.12	.38 1.19	1.08 1.46	1.14	.25	.42	72.5	74.7	FAC201
290	32	40	1.73	.41 1.10	.49 1.46	1.28	.10	.27	80.0	80.0	FMC082
226	16	40	-.28	.34 1.35	2.55 1.45	2.61	-.08	.35	47.5	67.2	FMC018
608	24	40	.76	.35 1.07	.64 1.45	2.12	.22	.35	62.5	67.0	FAS153
633	8	40	-1.33	.43 1.32	1.24 1.45	1.23	.01	.36	80.0	81.9	FAS178
746	20	40	.04	.36 1.22	1.46 1.45	2.16	.24	.45	67.5	70.2	FE023
2	16	40	-.47	.37 1.27	1.71 1.44	1.53	.26	.46	60.0	71.6	FAC002
111	21	40	.18	.36 1.27	1.87 1.44	1.55	.25	.46	62.5	69.6	FAC111
311	30	40	1.41	.38 1.20	1.10 1.44	1.47	.02	.29	70.0	75.2	FMC103
72	24	40	.57	.36 1.34	2.20 1.43	1.35	.20	.44	50.0	70.8	FAC072
183	31	40	1.60	.42 1.14	.69 1.43	.90	.25	.39	80.0	80.1	FAC183
193	29	40	1.28	.39 1.14	.79 1.43	1.01	.26	.41	80.0	76.4	FAC193
302	31	40	1.57	.40 1.15	.76 1.43	1.32	.03	.28	77.5	77.5	FMC094
590	23	40	.64	.34 1.13	1.15 1.43	2.20	.17	.36	62.5	66.3	FAS135
598	32	40	1.85	.41 1.15	.71 1.43	1.10	.05	.27	80.0	79.9	FAS143
864	23	40	.42	.36 1.39	2.47 1.43	1.90	.13	.45	52.5	70.2	FE141
127	12	40	-1.04	.39 1.21	1.13 1.42	1.21	.29	.46	70.0	76.1	FAC127
782	23	40	.42	.36 1.17	1.15 1.42	1.85	.27	.45	67.5	70.2	FE059
852	24	40	.55	.36 1.24	1.59 1.42	1.77	.21	.44	65.0	70.4	FE129
857	30	40	1.40	.40 1.20	1.02 1.42	1.16	.21	.39	67.5	77.1	FE134
452	13	40	-.65	.36 1.32	1.93 1.41	1.90	-.05	.34	57.5	70.9	FMC244
720	14	40	-.43	.36 1.29	1.83 1.40	1.88	.06	.38	60.0	70.4	FAS265
70	15	40	-.60	.37 1.24	1.45 1.39	1.35	.28	.46	65.0	72.2	FAC070
145	28	40	1.12	.38 1.13	.78 1.39	1.00	.29	.42	77.5	74.7	FAC145
881	24	40	.55	.36 1.33	2.06 1.39	1.68	.17	.44	55.0	70.4	FE158
205	26	40	.84	.37 1.38	2.24 1.38	1.11	.17	.43	60.0	72.3	FAC205
660	13	40	-.56	.37 1.28	1.66 1.38	1.66	.08	.38	62.5	71.8	FAS205
787	26	40	.82	.37 1.34	2.01 1.38	1.41	.15	.43	62.5	71.4	FE064
843	31	40	1.56	.41 1.24	1.13 1.38	.99	.17	.38	75.0	78.9	FE120
137	18	40	-.20	.36 1.17	1.18 1.36	1.37	.33	.46	65.0	70.5	FAC137
364	21	40	.30	.34 1.29	2.48 1.36	2.47	-.01	.35	50.0	64.7	FMC156
488	29	40	1.40	.37	-.42 1.36	1.20	.33	.31	75.0	72.7	FAS043
291	31	40	1.57	.40 1.12	.66 1.35	1.13	.07	.28	77.5	77.5	FMC083
314	12	40	-.78	.37 1.15	.93 1.35	1.51	.11	.34	70.0	72.5	FMC106
656	13	40	-.56	.37 1.17	1.05 1.35	1.55	.16	.38	72.5	71.8	FAS201
437	12	40	-.78	.37 1.24	1.42 1.34	1.47	.03	.34	65.0	72.5	FMC229
566	29	40	1.40	.37 1.16	.97 1.34	1.14	.10	.31	70.0	72.7	FAS111
770	20	40	.04	.36 1.23	1.52 1.34	1.66	.25	.45	62.5	70.2	FE047
94	25	40	.70	.37 1.30	1.87 1.33	1.04	.23	.44	60.0	71.4	FAC094
123	15	40	-.60	.37 1.33	1.94 1.23	.87	.26	.46	55.0	72.2	FAC123
293	22	40	.41	.34 1.21	1.81 1.33	2.16	.07	.35	57.5	64.9	FMC085
565	23	40	.64	.34 1.31	2.47 1.33	1.72	.03	.36	52.5	66.3	FAS110
733	14	40	-.74	.37 1.15	.95 1.33	1.28	.29	.44	72.5	72.7	FE010
794	17	40	-.34	.36 1.33	2.06 1.31	1.45	.19	.45	57.5	70.7	FE071
90	17	40	-.33	.36 1.26	1.67 1.32	1.20	.28	.46	62.5	71.0	FAC090
511	24	40	.76	.35	-.48 1.32	1.59	.35	.35	72.5	67.0	FAS056
793	23	40	.42	.36 1.32	2.04 1.30	1.39	.20	.45	52.5	70.2	FE070
190	29	40	1.28	.39 1.27	1.39 1.31	.78	.22	.41	75.0	76.4	FAC190
268	10	40	-1.07	.39 1.18	.95 1.31	1.13	.09	.32	72.5	76.2	FMC060
281	25	40	.76	.35 1.11	.93 1.31	1.67	.15	.33	62.5	66.8	FMC073
313	13	40	-.65	.36 1.19	1.23 1.31	1.50	.09	.34	62.5	70.9	FMC105
349	21	40	.30	.34 1.23	2.03 1.31	2.12	.05	.35	60.0	64.7	FMC141
702	16	40	-.18	.35 1.15	1.11 1.31	1.70	.19	.38	62.5	68.0	FAS247
49	19	40	-.07	.36 1.30	2.01 1.22	.91	.27	.46	57.5	70.1	FAC049
77	10	40	-1.36	.41 1.02	.18 1.30	.80	.39	.45	77.5	79.2	FAC077
374	22	40	.41	.34 1.18	1.58 1.30	2.01	.10	.35	62.5	64.9	FMC166
377	19	40	.07	.34 1.22	1.86 1.30	2.09	.07	.35	57.5	65.3	FMC169
466	13	40	-.56	.37 1.19	1.17 1.30	1.34	.16	.38	67.5	71.8	FAS011
500	19	40	.17	.34 1.21	1.78 1.30	1.80	.13	.37	57.5	65.9	FAS045
631	20	40	.29	.34 1.18	1.57 1.30	1.78	.15	.37	57.5	65.5	FAS176
10	13	40	-.89	.38 1.20	1.14 1.29	.94	.31	.46	67.5	74.8	FAC010
50	16	40	-.47	.37 1.27	1.70 1.29	1.10	.27	.46	65.0	71.6	FAC050

Person Fit Statistics of 902 students (continue)

50	16	40	-.47	.37 1.27	1.70 1.29	1.10	.27	.46	65.0	71.6	FAC050		
204	30	40	1.43	.40 1.07	.41 1.29	.71	.32	.40	82.5	78.1	FAC204		
244	16	40	-.28	.34 1.21	1.64 1.29	1.77	.08	.35	57.5	67.2	FMC036		
76	19	40	-.07	.36 1.28	1.87 1.28	1.10	.27	.46	57.5	70.1	FAC076		
138	23	40	.44	.36 1.17	1.20 1.28	1.00	.31	.45	67.5	70.2	FAC138		
447	14	40	-.53	.35 1.22	1.51 1.28	1.46	.07	.35	62.5	69.5	FMC239		
736	31	40	1.56	.41	.97	-.08 1.28	.78	.35	.38	85.0	78.9	FE013	
841	24	40	.55	.36 1.28	1.81 1.28	1.26	.22	.44	60.0	70.4	FE118		
299	30	40	1.41	.38 1.09	.56 1.27	.98	.16	.29	70.0	75.2	FMC091		
303	36	40	2.59	.54 1.12	.42 1.27	.62	.01	.20	90.0	90.0	FMC095		
455	22	40	.41	.34 1.27	2.30 1.26	1.76	.03	.35	42.5	64.9	FMC247		
730	28	40	1.09	.38 1.05	.33 1.27	.93	.36	.41	67.5	73.8	FE007		
317	33	40	1.90	.43 1.03	.19 1.26	.74	.16	.25	82.5	82.5	FMC109		
410	26	40	.88	.35 1.11	.87 1.26	1.33	.16	.33	65.0	68.0	FMC202		
610	21	40	.41	.34 1.15	1.34 1.26	1.55	.19	.37	55.0	65.5	FAS155		
622	15	40	-.31	.35 1.23	1.60 1.26	1.40	.14	.38	62.5	69.1	FAS167		
821	18	40	-.21	.36 1.09	.61 1.26	1.29	.36	.45	70.0	70.4	FE098		
884	30	40	1.40	.40 1.09	.54 1.26	.78	.30	.39	72.5	77.1	FE161		
91	19	40	-.07	.36 1.25	1.71 1.20	.83	.30	.46	57.5	70.1	FAC091		
430	18	40	-.05	.34 1.20	1.69 1.25	1.76	.10	.35	57.5	65.8	FMC222		
671	32	40	1.85	.41 1.07	.38 1.25	.73	.15	.27	80.0	79.9	FAS216		
847	26	40	.82	.37 1.14	.91 1.25	1.00	.30	.43	72.5	71.4	FE124		
93	20	40	.05	.36 1.24	1.67 1.19	.77	.30	.46	55.0	69.8	FAC093		
308	18	40	-.05	.34 1.21	1.73 1.24	1.66	.10	.35	57.5	65.8	FMC100		
326	14	40	-.53	.35 1.17	1.20 1.24	1.29	.13	.35	62.5	69.5	FMC118		
404	13	40	-.65	.36 1.10	.70 1.24	1.19	.19	.34	72.5	70.9	FMC196		
459	15	40	-.31	.35 1.21	1.45 1.24	1.29	.16	.38	57.5	69.1	FAS004		
470	39	40	4.23	1.02 1.03	.35 1.24	.59	.01	.10	97.5	97.5	FAS015		
591	24	40	.76	.35 1.03	.28 1.24	1.23	.28	.35	67.5	67.0	FAS136		
626	24	40	.76	.35 1.24	1.90 1.24	1.20	.10	.35	52.5	67.0	FAS171		
655	30	40	1.54	.38 1.16	.90 1.24	.79	.10	.30	72.5	75.0	FAS200		
676	19	40	.17	.34 1.22	1.84 1.24	1.47	.14	.37	52.5	65.9	FAS221		
754	34	40	2.14	.47	.90	-.27 1.24	.58	.35	.33	87.5	85.3	FE031	
888	27	40	.95	.37 1.24	1.41 1.20	.77	.24	.42	60.0	72.4	FE165		
208	24	40	.57	.36 1.23	1.53 1.23	.82	.28	.44	60.0	70.8	FAC208		
288	33	40	1.90	.43	.98	.01 1.23	.68	.22	.25	82.5	82.5	FMC080	
391	17	40	-.16	.34 1.14	1.12 1.23	1.54	.16	.35	67.5	66.4	FMC183		
461	15	40	-.31	.35 1.17	1.20 1.23	1.23	.19	.38	62.5	69.1	FAS006		
849	31	40	1.56	.41 1.00	.09 1.23	.68	.34	.38	80.0	78.9	FE126		
869	34	40	2.14	.47 1.09	.39 1.23	.57	.22	.33	87.5	85.3	FE146		
149	17	40	-.33	.36	.91	-.60 1.22	.87	.49	.46	77.5	71.0	FAC149	
292	26	40	.88	.35 1.11	.88 1.22	1.14	.17	.33	65.0	68.0	FMC084		
418	17	40	-.16	.34 1.16	1.28 1.22	1.48	.15	.35	62.5	66.4	FMC210		
567	25	40	.88	.35 1.15	1.22 1.22	1.04	.16	.34	62.5	67.7	FAS112		
662	28	40	1.26	.37 1.17	1.09 1.22	.85	.13	.32	65.0	70.9	FAS207		
59	31	40	1.60	.42 1.21	.95 1.01	.19	.28	.39	75.0	80.1	FAC059		
102	32	40	1.78	.43 1.05	.28 1.21	.54	.33	.37	80.0	82.0	FAC102		
366	18	40	-.05	.34 1.17	1.41 1.21	1.47	.15	.35	52.5	65.8	FMC158		
367	20	40	.18	.34 1.16	1.41 1.21	1.54	.15	.35	50.0	65.0	FMC159		
575	15	40	-.31	.35 1.16	1.14 1.21	1.13	.20	.38	57.5	69.1	FAS120		
606	26	40	1.00	.35 1.21	1.57 1.16	.77	.13	.34	52.5	68.5	FAS151		
695	39	40	4.23	1.02 1.03	.35 1.21	.58	.02	.10	97.5	97.5	FAS240		
763	27	40	.95	.37 1.12	.79 1.21	.79	.31	.42	65.0	72.4	FE040		
809	23	40	.42	.36 1.16	1.09 1.21	1.03	.31	.45	67.5	70.2	FE086		
862	26	40	.82	.37	.98	-.09 1.21	.88	.41	.43	77.5	71.4	FE139	
188	36	40	2.74	.56 1.05	.26 1.20	.51	.24	.29	90.0	90.0	FAC188		
338	20	40	.18	.34 1.16	1.46 1.20	1.48	.15	.35	55.0	65.0	FMC130		
414	28	40	1.14	.36 1.09	.60 1.20	.91	.18	.31	72.5	71.2	FMC206		
483	11	40	-.84	.38 1.20	1.08 1.18	.75	.18	.37	62.5	75.4	FAS028		
507	14	40	-.43	.36 1.15	1.00 1.20	1.02	.22	.38	60.0	70.4	FAS052		
196	36	40	2.74	.56 1.15	.49	.70	-.10	.26	.29	90.0	90.0	FAC196	
31	30	40	1.43	.40 1.02	.16	.77	-.36	.42	.40	72.5	78.1	FAC031	
32	31	40	1.60	.42 1.01	.10	.78	-.29	.41	.39	80.0	80.1	FAC032	
747	35	40	2.38	.51 1.01	.15	.73	-.26	.33	.30	87.5	87.5	FE024	
696	38	40	3.49	.73	.99	.19	.68	-.04	.19	.14	95.0	95.0	FAS241
686	36	40	2.72	.54	.97	.06	.75	-.21	.26	.20	90.0	90.0	FAS231
179	37	40	3.09	.64	.96	.08	.66	-.10	.30	.26	92.5	92.5	FAC179
478	36	40	2.72	.54	.96	.02	.73	-.25	.27	.20	90.0	90.0	FAS023
585	37	40	3.05	.61	.96	.08	.64	-.27	.26	.17	92.5	92.5	FAS130
BETTER FITTING NOT SHOWN													
748	33	40	1.93	.45	.96	-.06	.68	-.58	.42	.35	80.0	83.0	FE025
195	29	40	1.28	.39	.95	-.23	.79	-.38	.46	.41	80.0	76.4	FAC195
269	33	40	1.90	.43	.95	-.13	.75	-.59	.36	.25	82.5	82.5	FMC061
690	38	40	3.49	.73	.95	-.13	.52	-.27	.26	.14	95.0	95.0	FAS235
304	39	40	4.09	1.02	.94	.25	.39	-.19	.25	.10	97.5	97.5	FMC096
557	32	40	1.85	.41	.94	-.20	.80	-.42	.35	.27	80.0	79.9	FAS102
649	31	40	1.69	.40	.94	-.23	.80	-.50	.37	.28	77.5	77.4	FAS194
693	37	40	3.05	.61	.94	-.04	.63	-.30	.28	.17	92.5	92.5	FAS238
699	34	40	2.23	.46	.94	-.12	.75	-.38	.33	.24	85.0	84.9	FAS244
158	34	40	2.20	.48	.93	-.15	.58	-.45	.43	.34	85.0	86.1	FAC158
255	34	40	2.10	.46	.93	-.16	.71	-.59	.37	.24	85.0	85.0	FMC047
745	34	40	2.14	.47	.93	-.15	.63	-.60	.42	.33	82.5	85.3	FE022
773	36	40	2.66	.55	.93	-.06	.69	-.22	.34	.28	90.0	90.0	FE050
899	29	40	1.24	.39	.93	-.33	.75	-.72	.49	.40	75.0	75.5	FE176

Person Fit Statistics of 902 students (continue)

899	29	40	1.24	.39	.93	-.33	.75	-.72	.49	.40	75.0	75.5	FE176
900	31	40	1.56	.41	.93	-.25	.73	-.61	.46	.38	80.0	78.9	FE177
901	34	40	2.14	.47	.93	-.17	.78	-.24	.39	.33	87.5	85.3	FE178
207	29	40	1.28	.39	.92	-.35	.78	-.40	.47	.41	80.0	76.4	FAC207
272	32	40	1.73	.41	.92	-.30	.74	-.72	.41	.27	80.0	80.0	FMC064
611	37	40	3.05	.61	.92	-.01	.54	-.46	.31	.17	92.5	92.5	FAS156
639	31	40	1.69	.40	.92	-.33	.78	-.54	.39	.28	77.5	77.4	FAS184
653	7	40	-1.52	.45	.92	-.20	.69	-.74	.48	.35	82.5	84.0	FAS198
761	28	40	1.09	.38	.92	-.44	.76	-.77	.50	.41	72.5	73.8	FE038
298	37	40	2.92	.61	.91	-.04	.51	-.62	.36	.17	92.5	92.5	FMC090
530	34	40	2.23	.46	.91	-.23	.73	-.43	.35	.24	85.0	84.9	FAS075
534	27	40	1.13	.36	.91	-.64	.79	-.86	.45	.33	67.5	69.5	FAS079
546	31	40	1.69	.40	.91	-.38	.77	-.57	.39	.28	77.5	77.4	FAS091
767	32	40	1.74	.43	.91	-.31	.71	-.59	.46	.36	77.5	80.8	FE044
776	37	40	3.00	.62	.91	-.04	.58	-.26	.34	.24	92.5	92.5	FE053
812	26	40	.82	.37	.91	-.55	.79	-.80	.52	.43	67.5	71.4	FE089
842	28	40	1.09	.38	.91	-.47	.78	-.69	.50	.41	77.5	73.8	FE119
55	30	40	1.43	.40	.90	-.44	.72	-.49	.49	.40	82.5	78.1	FAC055
56	30	40	1.43	.40	.90	-.44	.72	-.49	.49	.40	82.5	78.1	FAC056
85	14	40	-.74	.38	.90	-.59	.79	-.69	.54	.46	75.0	73.4	FAC085
141	24	40	.57	.36	.90	-.67	.79	-.65	.53	.44	75.0	70.8	FAC141
330	36	40	2.59	.54	.90	-.14	.61	-.58	.37	.20	90.0	90.0	FMC122
9	18	40	-.20	.36	.89	-.76	.79	-.80	.55	.46	70.0	70.5	FAC009
23	24	40	.57	.36	.89	-.72	.78	-.67	.53	.44	75.0	70.8	FAC023
37	20	40	.05	.36	.89	-.77	.80	-.77	.54	.46	75.0	69.8	FAC037
97	32	40	1.78	.43	.89	-.41	.69	-.41	.47	.37	85.0	82.0	FAC097
150	24	40	.57	.36	.89	-.72	.78	-.70	.53	.44	75.0	70.8	FAC150
301	31	40	1.57	.40	.89	-.51	.73	-.87	.45	.28	77.5	77.5	FMC093
477	28	40	1.26	.37	.89	-.71	.78	-.79	.45	.32	70.0	70.9	FAS022
595	37	40	3.05	.61	.89	-.07	.50	-.54	.34	.17	92.5	92.5	FAS140
800	34	40	2.14	.47	.89	-.29	.63	-.58	.44	.33	87.5	85.3	FE077
836	31	40	1.56	.41	.89	-.48	.72	-.65	.48	.38	85.0	78.9	FE113
38	31	40	1.60	.42	.88	-.51	.66	-.56	.49	.39	80.0	80.1	FAC038
52	28	40	1.12	.38	.88	-.63	.71	-.66	.52	.42	77.5	74.7	FAC052
62	19	40	-.07	.36	.88	-.85	.79	-.80	.55	.46	77.5	70.1	FAC062
229	27	40	1.01	.36	.88	-.86	.79	-1.01	.48	.32	72.5	69.3	FMC021
740	34	40	2.14	.47	.88	-.34	.62	-.61	.45	.33	87.5	85.3	FE017
855	29	40	1.24	.39	.88	-.62	.75	-.73	.51	.40	80.0	75.5	FE132
40	24	40	.57	.36	.87	-.87	.76	-.79	.55	.44	75.0	70.8	FAC040
41	27	40	.98	.38	.87	-.75	.75	-.62	.53	.42	72.5	73.2	FAC041
42	22	40	.31	.36	.87	-.91	.77	-.83	.55	.45	80.0	69.7	FAC042
75	21	40	.18	.36	.87	-.97	.75	-.94	.56	.46	72.5	69.6	FAC075
82	20	40	.05	.36	.87	-.98	.80	-.78	.56	.46	75.0	69.8	FAC082
559	36	40	2.72	.54	.87	-.21	.51	-.72	.39	.20	90.0	90.0	FAS104
823	25	40	.68	.36	.87	-.88	.77	-1.00	.56	.44	77.5	70.7	FE100
848	19	40	-.09	.36	.87	-.86	.80	-1.06	.57	.45	75.0	70.4	FE125
867	29	40	1.24	.39	.87	-.71	.70	-.91	.53	.40	80.0	75.5	FE144
24	23	40	.44	.36	.86	-1.04	.78	-.76	.56	.45	77.5	70.2	FAC024
44	30	40	1.43	.40	.86	-.68	.66	-.65	.52	.40	82.5	78.1	FAC044
53	21	40	.18	.36	.86	-1.02	.76	-.93	.56	.46	77.5	69.6	FAC053
68	29	40	1.28	.39	.86	-.70	.73	-.53	.51	.41	80.0	76.4	FAC068
78	26	40	.84	.37	.86	-.92	.80	-.51	.53	.43	75.0	72.3	FAC078
95	29	40	1.28	.39	.86	-.70	.73	-.53	.51	.41	80.0	76.4	FAC095
267	33	40	1.90	.43	.86	-.49	.65	-.89	.46	.25	82.5	82.5	FMC059
600	37	40	3.05	.61	.86	-.13	.45	-.64	.38	.17	92.5	92.5	FAS145
735	33	40	1.93	.45	.86	-.48	.78	-.31	.45	.35	85.0	83.0	FE012
744	35	40	2.38	.51	.86	-.34	.59	-.55	.44	.30	87.5	87.5	FE021
786	31	40	1.56	.41	.86	-.61	.67	-.80	.51	.38	80.0	78.9	FE063
792	33	40	1.93	.45	.86	-.48	.78	-.31	.45	.35	85.0	83.0	FE069
66	20	40	.05	.36	.85	-1.14	.78	-.85	.57	.46	75.0	69.8	FAC066
156	22	40	.31	.36	.85	-1.13	.74	-.95	.57	.45	80.0	69.7	FAC156
525	29	40	1.40	.37	.85	-.87	.72	-.97	.48	.31	75.0	72.7	FAS070
594	30	40	1.54	.38	.85	-.79	.77	-.66	.46	.30	77.5	75.0	FAS139
728	30	40	1.40	.40	.85	-.73	.66	-.97	.54	.39	82.5	77.1	FE005
760	31	40	1.56	.41	.85	-.66	.76	-.52	.50	.38	85.0	78.9	FE037
788	35	40	2.38	.51	.85	-.38	.53	-.69	.46	.30	87.5	87.5	FE065
865	29	40	1.24	.39	.85	-.85	.64	-1.17	.56	.40	75.0	75.5	FE142
887	13	40	-.88	.38	.85	-.89	.74	-.98	.57	.43	75.0	73.8	FE164
12	25	40	-.70	.37	.84	-1.05	.74	-.77	.55	.44	80.0	71.4	FAC012
89	18	40	-.20	.36	.84	-1.15	.74	-1.03	.58	.46	75.0	70.5	FAC089
165	21	40	.18	.36	.84	-1.19	.74	-1.02	.58	.46	77.5	69.6	FAC165
170	19	40	-.07	.36	.84	-1.20	.74	-1.07	.58	.46	77.5	70.1	FAC170
175	22	40	.31	.36	.84	-1.22	.78	-.80	.57	.45	75.0	69.7	FAC175
203	37	40	3.09	.64	.84	-.20	.39	-.55	.41	.26	92.5	92.5	FAC203
296	30	40	1.41	.38	.84	-.84	.70	-1.11	.51	.29	80.0	75.2	FMC088
484	23	40	.64	.34	.84	-1.44	.80	-1.15	.53	.36	72.5	66.3	FAS029
531	25	40	.88	.35	.84	-1.33	.75	-1.24	.53	.34	72.5	67.7	FAS076
561	33	40	2.03	.43	.84	-.59	.65	-.78	.45	.25	82.5	82.4	FAS106
646	29	40	1.40	.37	.84	-.99	.72	-.98	.50	.31	75.0	72.7	FAS191
672	15	40	-.31	.35	.84	-1.19	.79	-1.19	.56	.38	72.5	69.1	FAS217
1	21	40	.18	.36	.83	-1.31	.72	-1.10	.59	.46	77.5	69.6	FAC001
74	31	40	1.60	.42	.83	-.73	.73	-.40	.50	.39	85.0	80.1	FAC074
80	28	40	1.12	.38	.83	-.96	.74	-.59	.54	.42	77.5	74.7	FAC080
88	15	40	-.60	.37	.83	-1.09	.70	-1.11	.59	.46	80.0	72.2	FAC088

Person Fit Statistics of 902 students (continue)

88	15	40	-.60	.37	.83	-1.09	.70	-1.11	.59	.46	80.0	72.2	FAC088
151	31	40	1.60	.42	.83	-.74	.66	-.57	.52	.39	80.0	80.1	FAC151
163	21	40	.18	.36	.83	-1.30	.73	-1.07	.59	.46	77.5	69.6	FAC163
169	28	40	1.12	.38	.83	-.99	.68	-.77	.55	.42	77.5	74.7	FAC169
176	27	40	.98	.38	.83	-1.06	.72	-.72	.55	.42	82.5	73.2	FAC176
180	24	40	.57	.36	.83	-1.20	.72	-.95	.57	.44	75.0	70.8	FAC180
247	29	40	1.27	.37	.83	-1.06	.70	-1.24	.54	.30	75.0	73.2	FMC039
248	25	40	.76	.35	.83	-1.48	.76	-1.43	.55	.33	77.5	66.8	FMC040
259	29	40	1.27	.37	.83	-1.02	.74	-1.06	.52	.30	75.0	73.2	FMC051
480	20	40	.29	.34	.83	-1.54	.80	-1.34	.55	.37	82.5	65.5	FAS025
496	26	40	1.00	.35	.83	-1.33	.74	-1.21	.53	.34	77.5	68.5	FAS041
513	23	40	.64	.34	.83	-1.52	.77	-1.32	.54	.36	72.5	66.3	FAS058
524	23	40	.64	.34	.83	-1.58	.76	-1.38	.55	.36	72.5	66.3	FAS069
623	34	40	2.23	.46	.83	-.54	.59	-.82	.45	.24	85.0	84.9	FAS168
755	32	40	1.74	.43	.83	-.70	.57	-1.01	.53	.36	82.5	80.8	FE032
853	15	40	-.60	.37	.83	-1.07	.75	-1.10	.59	.44	75.0	71.7	FE130
898	16	40	-.47	.36	.83	-1.09	.74	-1.27	.60	.45	75.0	71.0	FE175
17	21	40	.18	.36	.82	-1.35	.71	-1.15	.59	.46	77.5	69.6	FAC017
60	28	40	1.12	.38	.82	-1.06	.65	-.86	.56	.42	77.5	74.7	FAC060
167	22	40	.31	.36	.82	-1.33	.73	-1.01	.58	.45	80.0	69.7	FAC167
212	26	40	.88	.35	.82	-1.43	.80	-1.07	.54	.33	75.0	68.0	FAC004
214	25	40	.76	.35	.82	-1.57	.74	-1.59	.57	.33	72.5	66.8	FMC006
253	32	40	1.73	.41	.82	-.78	.61	-1.18	.53	.27	80.0	80.0	FMC045
263	34	40	2.10	.46	.82	-.56	.56	-1.06	.50	.24	85.0	85.0	FMC055
497	28	40	1.26	.37	.82	-1.22	.71	-1.12	.52	.32	75.0	70.9	FAS042
552	33	40	2.03	.43	.82	-.67	.61	-.89	.47	.25	82.5	82.4	FAS097
578	28	40	1.26	.37	.82	-1.25	.70	-1.17	.53	.32	80.0	70.9	FAS123
602	25	40	.88	.35	.82	-1.55	.74	-1.30	.55	.34	77.5	67.7	FAS147
739	33	40	1.93	.45	.82	-.66	.69	-.55	.49	.35	85.0	83.0	FE016
752	39	40	4.21	1.03	.82	.11	.18	-.45	.34	.15	97.5	97.5	FE029
771	32	40	1.74	.43	.82	-.77	.60	-.92	.53	.36	87.5	80.8	FE048
11	18	40	-.20	.36	.81	-1.36	.71	-1.19	.60	.46	75.0	70.5	FAC011
99	30	40	1.43	.40	.81	-.95	.68	-.59	.54	.40	82.5	78.1	FAC099
264	35	40	2.33	.49	.81	-.47	.50	-1.06	.50	.22	87.5	87.5	FMC056
276	35	40	2.33	.49	.81	-.47	.50	-1.06	.50	.22	87.5	87.5	FMC068
458	25	40	.88	.35	.81	-1.56	.78	-1.08	.54	.34	82.5	67.7	FAS003
607	30	40	1.54	.38	.81	-1.08	.72	-.86	.50	.30	77.5	75.0	FAS152
673	25	40	.88	.35	.81	-1.64	.77	-1.15	.55	.34	82.5	67.7	FAS218
750	32	40	1.74	.43	.81	-.81	.55	-1.09	.55	.36	82.5	80.8	FE027
764	32	40	1.74	.43	.81	-.81	.55	-1.09	.55	.36	82.5	80.8	FE041
4	26	40	.84	.37	.80	-1.35	.65	-1.05	.59	.43	75.0	72.3	FAC004
16	22	40	.31	.36	.80	-1.49	.69	-1.17	.60	.45	75.0	69.7	FAC016
34	26	40	.84	.37	.80	-1.35	.68	-.95	.58	.43	80.0	72.3	FAC034
47	24	40	.57	.36	.80	-1.48	.69	-1.06	.60	.44	75.0	70.8	FAC047
54	30	40	1.43	.40	.80	-1.01	.58	-.89	.56	.40	77.5	78.1	FAC054
64	26	40	.84	.37	.80	-1.34	.64	-1.07	.59	.43	75.0	72.3	FAC064
119	21	40	.18	.36	.80	-1.48	.71	-1.13	.60	.46	77.5	69.6	FAC119
216	28	40	1.14	.36	.80	-1.37	.69	-1.45	.57	.31	72.5	71.2	FMC008
237	25	40	.76	.35	.80	-1.72	.73	-1.68	.59	.33	77.5	66.8	FMC029
238	29	40	1.27	.37	.80	-1.22	.67	-1.41	.57	.30	75.0	73.2	FMC030
254	30	40	1.41	.38	.80	-1.12	.65	-1.34	.56	.29	80.0	75.2	FMC046
256	27	40	1.01	.36	.80	-1.47	.70	-1.52	.58	.32	72.5	69.3	FMC048
278	29	40	1.27	.37	.80	-1.26	.73	-1.09	.55	.30	80.0	73.2	FMC070
357	26	40	.88	.35	.80	-1.60	.74	-1.46	.57	.33	80.0	68.0	FMC149
548	21	40	.41	.34	.80	-1.93	.77	-1.48	.58	.37	85.0	65.5	FAS093
550	21	40	.41	.34	.80	-1.93	.77	-1.48	.58	.37	85.0	65.5	FAS095
553	21	40	.41	.34	.80	-1.93	.77	-1.48	.58	.37	85.0	65.5	FAS098
765	33	40	1.93	.45	.80	-.74	.56	-.89	.53	.35	85.0	83.0	FE042
791	33	40	1.93	.45	.80	-.74	.56	-.89	.53	.35	85.0	83.0	FE068
860	22	40	.29	.36	.80	-1.47	.72	-1.51	.63	.45	75.0	70.0	FE137
213	28	40	1.14	.36	.79	-1.40	.67	-1.54	.58	.31	67.5	71.2	FMC005
217	25	40	.76	.35	.79	-1.78	.75	-1.52	.58	.33	82.5	66.8	FMC009
260	27	40	1.01	.36	.79	-1.54	.70	-1.56	.59	.32	77.5	69.3	FMC052
473	26	40	1.00	.35	.79	-1.71	.75	-1.12	.56	.34	77.5	68.5	FAS018
486	14	40	-.43	.36	.79	-1.48	.70	-1.62	.62	.38	75.0	70.4	FAS031
503	29	40	1.40	.37	.79	-1.35	.72	-.95	.53	.31	75.0	72.7	FAS048
510	30	40	1.54	.38	.79	-1.16	.68	-1.02	.52	.30	77.5	75.0	FAS055
731	34	40	2.14	.47	.79	-.71	.54	-.82	.52	.33	87.5	85.3	FE008
732	34	40	2.14	.47	.79	-.71	.54	-.82	.52	.33	87.5	85.3	FE009
734	33	40	1.93	.45	.79	-.82	.55	-.94	.54	.35	85.0	83.0	FE011
759	33	40	1.93	.45	.79	-.78	.51	-1.06	.55	.35	85.0	83.0	FE036
762	37	40	3.00	.62	.79	-.32	.38	-.63	.44	.24	92.5	92.5	FE039
774	31	40	1.56	.41	.79	-1.02	.57	-1.15	.57	.38	85.0	78.9	FE051
781	34	40	2.14	.47	.79	-.71	.54	-.82	.52	.33	87.5	85.3	FE058
3	18	40	-.20	.36	.78	-1.67	.68	-1.35	.63	.46	80.0	70.5	FAC003
46	21	40	.18	.36	.78	-1.68	.68	-1.28	.62	.46	77.5	69.6	FAC046
200	27	40	.98	.38	.78	-1.36	.68	-.86	.58	.42	82.5	73.2	FAC200
210	25	40	.76	.35	.78	-1.94	.76	-1.43	.59	.33	82.5	66.8	FMC002
222	25	40	.76	.35	.78	-1.93	.72	-1.76	.61	.33	82.5	66.8	FMC014
223	22	40	.41	.34	.78	-2.13	.75	-1.95	.61	.35	82.5	64.9	FMC015
225	24	40	.64	.34	.78	-2.03	.72	-1.86	.61	.34	82.5	65.9	FMC017
228	27	40	1.01	.36	.78	-1.65	.70	-1.55	.60	.32	77.5	69.3	FMC020
232	27	40	1.01	.36	.78	-1.67	.68	-1.64	.60	.32	77.5	69.3	FMC024
239	23	40	.53	.34	.78	-2.10	.73	-1.94	.61	.34	82.5	65.1	FMC031


Person Fit Statistics of 902 students (continue)

239	23	40	.53	.34	.78	-2.10	.73	-1.94	.61	.34	82.5	65.1	FMC031
250	27	40	1.01	.36	.78	-1.68	.69	-1.63	.60	.32	77.5	69.3	FMC042
257	32	40	1.73	.41	.78	-.97	.57	-1.35	.57	.27	80.0	80.0	FMC049
300	27	40	1.01	.36	.78	-1.65	.68	-1.67	.60	.32	72.5	69.3	FMC092
508	26	40	1.00	.35	.78	-1.76	.70	-1.40	.57	.34	87.5	68.5	FAS053
749	35	40	2.38	.51	.78	-.60	.53	-.69	.49	.30	87.5	87.5	FE026
162	26	40	.84	.37	.77	-1.54	.67	-.99	.60	.43	85.0	72.3	FAC162
230	26	40	.88	.35	.77	-1.87	.69	-1.78	.62	.33	80.0	68.0	FMC022
240	26	40	.88	.35	.77	-1.86	.69	-1.78	.62	.33	75.0	68.0	FMC032
279	23	40	.53	.34	.77	-2.19	.73	-1.95	.62	.34	82.5	65.1	FMC071
535	24	40	.76	.35	.77	-2.08	.70	-1.69	.60	.35	77.5	67.0	FAS080
742	34	40	2.14	.47	.77	-.76	.49	-.96	.53	.33	87.5	85.3	FE019
120	20	40	.05	.36	.76	-1.86	.69	-1.29	.63	.46	85.0	69.8	FAC120
209	22	40	.41	.34	.76	-2.32	.75	-1.88	.62	.35	87.5	64.9	FMC001
242	28	40	1.14	.36	.76	-1.69	.64	-1.70	.62	.31	72.5	71.2	FMC034
512	16	40	-.18	.35	.76	-1.94	.70	-1.93	.64	.38	82.5	68.0	FAS057
751	35	40	2.38	.51	.76	-.70	.45	-.89	.52	.30	87.5	87.5	FE028
801	17	40	-.34	.36	.76	-1.72	.67	-1.73	.66	.45	77.5	70.7	FE078
110	18	40	-.20	.36	.75	-1.85	.66	-1.44	.64	.46	80.0	70.5	FAC110
153	26	40	.84	.37	.75	-1.71	.64	-1.08	.61	.43	80.0	72.3	FAC153
215	24	40	.64	.34	.75	-2.32	.69	-2.10	.65	.34	82.5	65.9	FMC007
236	28	40	1.14	.36	.75	-1.79	.63	-1.76	.64	.31	72.5	71.2	FMC028
33	26	40	.84	.37	.74	-1.80	.61	-1.20	.62	.43	85.0	72.3	FAC033
135	19	40	-.07	.36	.74	-1.98	.66	-1.44	.64	.46	82.5	70.1	FAC135
220	28	40	1.14	.36	.74	-1.80	.63	-1.77	.64	.31	72.5	71.2	FMC012
231	26	40	.88	.35	.74	-2.12	.66	-1.94	.65	.33	80.0	68.0	FMC023
233	26	40	.88	.35	.74	-2.12	.66	-1.94	.65	.33	80.0	68.0	FMC025
542	20	40	.29	.34	.74	-2.45	.71	-1.97	.64	.37	77.5	65.5	FAS087
634	25	40	.88	.35	.74	-2.30	.68	-1.68	.62	.34	82.5	67.7	FAS179
178	25	40	.70	.37	.73	-1.98	.65	-1.16	.63	.44	90.0	71.4	FAC178
221	26	40	.88	.35	.73	-2.24	.65	-2.05	.66	.33	80.0	68.0	FMC013
224	25	40	.76	.35	.73	-2.37	.67	-2.12	.66	.33	82.5	66.8	FMC016
280	25	40	.76	.35	.73	-2.37	.67	-2.12	.66	.33	82.5	66.8	FMC072
533	18	40	.06	.34	.73	-2.47	.67	-2.30	.67	.38	80.0	66.5	FAS078
757	34	40	2.14	.47	.73	-.92	.44	-1.10	.57	.33	87.5	85.3	FE034
36	25	40	.70	.37	.72	-2.04	.61	-1.30	.64	.44	85.0	71.4	FAC036
753	36	40	2.66	.55	.72	-.66	.36	-.93	.53	.28	90.0	90.0	FE030
807	21	40	.17	.36	.71	-2.24	.66	-1.98	.69	.45	87.5	70.1	FE084
870	32	40	1.74	.43	.71	-1.35	.47	-1.35	.62	.36	87.5	80.8	FE147
154	28	40	1.12	.38	.67	-2.09	.52	-1.34	.66	.42	82.5	74.7	FAC154
246	22	40	.41	.34	.67	-3.36	.63	-3.01	.74	.35	87.5	64.9	FMC038
115	25	40	.70	.37	.65	-2.60	.54	-1.61	.68	.44	90.0	71.4	FAC115
142	20	40	.05	.36	.65	-2.90	.56	-1.99	.71	.46	85.0	69.8	FAC142
729	18	40	-.21	.36	.65	-2.68	.58	-2.46	.74	.45	85.0	70.4	FE006
861	13	40	-.88	.38	.62	-2.55	.52	-2.05	.74	.43	90.0	73.8	FE138
MEAN	24.5	40.0	.77	.38	1.00	.0	.99	.0			72.5	72.5	
P. SD	6.3	.0	.91	.08	.15	1.0	.29	1.0			9.8	6.9	

APPENDIX H

LRN COPYRIGHT PERMISSION

1/00/2021 Gmail - Requesting to facilitate to gain the written permission for copyrighted materials

 Gmail Fouzul Kareema Ismail <mfikareema@gmail.com>

Requesting to facilitate to gain the written permission for copyrighted materials
4 messages

Fouzul Kareema Ismail <mfikareema@gmail.com> Mon, Aug 10, 2020 at 4:57 PM
To: Language Testing Research and Practice <LTEST-L@lists.psu.edu>

Dear Sirs and Madams of LTEST-L subscribers

It gives me pleasure to inform you that I am a Ph.D. student from International Islamic University Malaysia. I am researching under the topic of VALIDATION OF MULTI-LEVEL ENGLISH READING TEST ALIGNED WITH COMMON EUROPEAN FRAMEWORK OF REFERENCE USING RASCH MODEL.

I have selected around ten reading passages aligned with CEFR levels from the public domains of Cambridge International Examinations, British Council, Council of Europe, and Euroexam international websites. I need to get written permission from the aforesaid institutions to utilize their materials for the adaptation of reading tests for my own research purposes.

Could any of you please advise me on how I can get the proper written permission from the copyright holders of these institutions? Please provide me the email address to whom or to which department I have to write to get the written permission to use the copyrighted materials.

I am very much obliged to you if you would assist my request.

Thank you so much.

With kind regards,

M.F.Kareema
Senior Lecturer in English
DELTA
SEUSL
Sri Lanka

Zohaib Tariq <zohaibtariq@hotmail.com> Mon, Aug 10, 2020 at 5:34 PM
To: Fouzul Kareema Ismail <mfikareema@gmail.com>

Dear Fouzul,

We are recognised exam board in the UK similar like Cambridge or Pearson we can provide you reading passages for research.

Kind regards,
Muhammad Tariq

From: Language Testing Research and Practice <LTEST-L@lists.psu.edu> on behalf of Fouzul Kareema Ismail <mfikareema@GMAIL.COM>
Sent: 10 August 2020 09:57
To: LTEST-L@LISTS.PSU.EDU <LTEST-L@LISTS.PSU.EDU>
Subject: [LTEST-L] Requesting to facilitate to gain the written permission for copyrighted materials

[Quoted text hidden]

You can delete yourself from the list by simply sending mail to

LTest-L-unsubscribe-request@lists.psu.edu.

No subject or message text is required.
To change your settings or look through the archives, go to

<http://lists.psu.edu/archives/test-l.html>.

<https://mail.google.com/mail/u/0/?ui=01d37be02961vw=pt&search=all&permhid=thwad-a%3A%5045066633167923574&siml=map4%3A-20002...> 1/2

Fouzul Kareema Ismail <mfkareema@gmail.com>
To: Zohaib Tariq <zohaibtariq@hotmail.com>

Mon, Aug 10, 2020 at 5:43 PM

Thank you so much sir for your willingness to provide me materials. I proposed to develop and validate four reading tests including three CEFR aligned B1, B2, and C1 passages in each test. Could you please let me know whether your exam materials are CEFR aligned?

Thank you so much for your help.

With kind regards,

M.F.Kareema
Senior Lecturer in English
DELT
SEUSL

[Quoted text hidden]

Zohaib Tariq <zohaibtariq@hotmail.com>
To: Fouzul Kareema Ismail <mfkareema@gmail.com>

Mon, Aug 10, 2020 at 5:51 PM

Hello

Yes, our exams are aligned to the CEFR, have a look the test website <http://www.jelca.org/>

Please see the research section for the recent research.

If you require any documents or material please feel free to ask me.

Kind regards,
Muhammad

From: Fouzul Kareema Ismail <mfkareema@gmail.com>
Sent: 10 August 2020 10:43
To: Zohaib Tariq <zohaibtariq@hotmail.com>
Subject: Re: [LTEST-L] Requesting to facilitate to gain the written permission for copyrighted materials

[Quoted text hidden]