

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

**Research and Validation
at LanguageCert**

Volume 1

Edited by Peter Falvey and David Coniam

**Language
Cert**

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

Research and Validation at LanguageCert Volume 1

Peter Falvey, The Education University of Hong Kong,
Tai Po, Hong Kong

David Coniam, PeopleCert, London, UK

Publisher details: LanguageCert, London, UK

Date of Publication: February 2022, May 2023

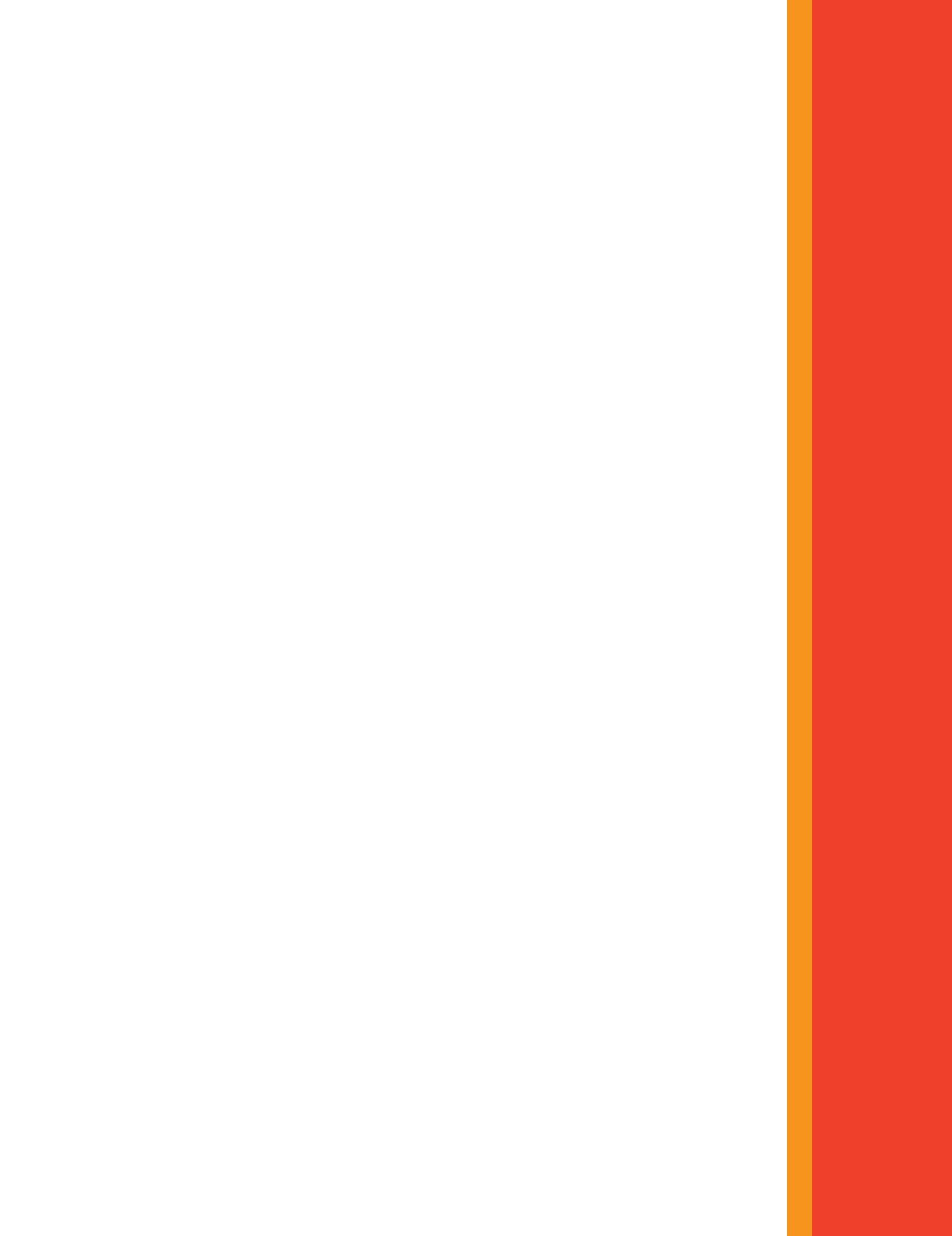
ISBN: 978-9925-34-297-6

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

**Research and Validation
at LanguageCert**

Volume 1

Edited by Peter Falvey and David Coniam



Foreword

Byron Nicolaides / CEO, PeopleCert Group

I am very pleased to be introducing this first volume of research papers produced by our LanguageCert research team. LanguageCert was born in 2016 following PeopleCert's acquisition of the English language tests owned by City and Guilds in 2015. These tests had a long history. They were first developed by Pitman and then redeveloped by City and Guilds. Our acquisition followed a period where PeopleCert was the sole distributor of these exams in Greece. Following the acquisition, we spent some time reviewing the exam content, specifications and delivery methods and once we were happy with the structure and format, submitted everything to the English exams' regulator, Ofqual. The exams and all the processes and systems that underpin them were reviewed by Ofqual's experts and became fully regulated in October 2017 under the awarding body PeopleCert Qualifications.

I have had a long involvement in education. Being born and brought up in Istanbul of Greek parents, both of whom were teachers, education lay at the centre of our family. I moved to Athens as a young man and worked in business including with Merrill Lynch as Vice President International. However, following my work with Merrill Lynch my interest in education led me to assessment. I founded PeopleCert in 2000 and I became involved in the early days of the European Computer Driving License (ECDL), which we distribute in Greece to this day. I have remained involved in ECDL for more than 20 years. I worked with Futurekids, one of the first computer and technology training programs for children. My interest in languages came when PeopleCert acted as the distributor for City and Guilds English exams in Greece.

Until 2008, PeopleCert was predominantly a Greek company operating in Greece but the financial crisis in 2008 hit us hard and led me to the view that we should operate internationally. As a result, I started building relationships with companies outside Greece so that today, in 2023, more than 95% of our business is international. In 2010 we started distributing qualifications for what was to become Axelos, a company owned in part by the UK government and in part by Capita. By 2017 we were the sole distributor globally and in 2021 were able to acquire Axelos making the transition from distributor to IP owner.

In the meantime, our languages business was developing rapidly and in 2019 we became one of only three awarding bodies to be granted rights to offer secure English language tests for visas and immigration to the UK by the Home Office. Our network of centres and, in turn, candidature has grown steadily since this time and as we matured as an English language awarding body, I decided to formally establish a research division in-house. We had commissioned several research studies before setting up our own department but it became clear that monitoring standards, researching test quality, and impact could not be done through external contracts alone.

This volume, therefore, is the first in a series of research volumes that describe how we underpin our exams with relevant research into scale development and validation, test calibration, test bias, the impact of delivery methods on test performance and the role of expert judgement in test development amongst other things. I hope you find the work both interesting and useful and I look forward to introducing Volume 2 in this series.

In closing, my thanks go to the contributors to this volume. Their hard work and commitment to LanguageCert is appreciated. I would also like to thank both the organisation's staff and the greatly valued network of partners around the world. They make the efficient exam delivery possible.

Preface

Marios Molfetas / Chief Languages Officer and Series Editor

LanguageCert is part of the PeopleCert group, a leading international assessment company. An inherent part of PeopleCert's perspective on assessment involves research into its examinations, with a view to showing their robustness and validity, and to make contribution to the academic arena of assessment.

PeopleCert was founded in 2000 and has since spearheaded a revolution in the testing and certification of professional skills, delivering millions of exams across 200 countries through state-of-the-art technology platforms.

LanguageCert was founded in 2015, with language qualifications becoming part of the larger family regarding the certification of professional skills. Since its founding, LanguageCert has been administering its International English for Speakers of Other Languages (IESOL) set of examinations which it acquired from the UK City & Guilds examination body, as well as developing a number of language qualifications – not only for English, but also for other languages. Where PeopleCert's qualifications certify essential professional skills in the workplace, LanguageCert's focus is on languages. Its certificates are internationally recognised, regulated and widely used in elementary and secondary schools as well as in the tertiary sector. They are also used extensively by the Home Office's UK Visa and Immigration (UKVI) programme to certify the English of the UK's visa applicants.

The current volume includes a range of papers produced by LanguageCert's Research Team, formally established in 2020 – see <https://www.languagecert.org/en/about-us/research-and-validation/research-team>. One of LanguageCert's missions is to produce tests of the highest quality and its research agendas are designed to support this mission.

Research into LanguageCert's IESOL examinations and their relationship to the CEFR was carried out in 2018 by the UK's National Recognition Information Centre (UK NARIC), and followed up by the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire in 2019. Since then an in-house research programme has been established, headed by Professor David Coniam. This programme extends and builds on some of this early research but the LanguageCert Research Team's brief extends to topics and issues that not only seek to validate the tests and their examiners but also reveal to stakeholders how the tests maintain consistently high quality, embodying the best principles of high-stakes assessment.



Contents

Foreword - Byron Nicolaides	5
Preface - Marios Molfetas.....	7
Contributors	11
Common Abbreviations	13
Introduction - Peter Falvey	15
SECTION 1: BACKGROUND	21
Chapter 1: LanguageCert – A Multilingual Language Testing System - Michael Milanovic and Angeliki Cheilaris	23
Chapter 2: External Validation of LanguageCert’s English Language Examinations - Yiannis Papargyris and Leda Lampropoulou	27
SECTION 2: EXPLORATIONS INTO TEST QUALITY	35
Chapter 3: LanguageCert IESOL Listening and Reading Test Reliabilities 2018-2020 - David Coniam and Leda Lampropoulou	37
Chapter 4: Examiner Quality and Consistency across LanguageCert Writing Tests - David Coniam and Yiannis Papargyris	41
Chapter 5: Potential Bias in LanguageCert IESOL Items: A Differential Item Functioning Analysis - David Coniam and Tony Lee	51
Chapter 6: Task Equivalence in LanguageCert IESOL Writing Tests - David Coniam and Maria Babatsi.....	61
SECTION 3: CALIBRATION STUDIES	69
Chapter 7: Validating the LanguageCert Test of English Scale: The Paper-based Tests - David Coniam, Tony Lee, Michael Milanovic and Nigel Pike	71
Chapter 8: Validating the LanguageCert Test of English Scale: The Adaptive Test - David Coniam, Tony Lee, Michael Milanovic and Nigel Pike	87
Chapter 9: Externally-Referenced Anchoring: Equating Expert Judgement and Rasch Measurement Values in LanguageCert IESOL English Language Tests - Tony Lee, Michael Milanovic, David Coniam and Nigel Pike	95

SECTION 4: ORIGINAL RESEARCH.....	113
Chapter 10: Identifying Guessing in English Language Tests via Rasch Fit Statistics: An Exploratory Study - David Coniam, Tony Lee and Leda Lampropoulou.....	115
Chapter 11: The Development and Delivery of Online-Proctored Speaking Exams: The Case of LanguageCert International ESOL - Leda Lampropoulou and Yiannis Papargyris.....	125
Chapter 12: Online Proctoring of High-stakes Examinations: A Survey of Past Candidates' Attitudes and Perceptions - David Coniam, Leda Lampropoulou and Angeliki Cheilaris.....	131
Chapter 13: Automated Writing Assessment (AWA) and the Carnegie Speech Writing Assessment System - David Coniam and Tony Lee.....	151
Chapter 14: Towards a Communicative Test of Reading and Language Use for Classical Greek - Polyxeni Poupounaki-Lappa, Tzortzina Peristeri and David Coniam.....	165
Chapter 15: Recapping and Looking Ahead - Peter Falvey.....	181
Glossary: Statistical Techniques Used in The Volume - Peter Falvey.....	185

Contributors

Byron Nicolaides is the founder and CEO of the PeopleCert Group, a global leader in the assessment and certification of professional skills, partnering with multinational organisations and government bodies to develop and deliver market-leading exams worldwide. He is also the president of the Council of European Professional Informatics Societies (CEPIS), where he advances IT trends across the 29-country membership. A pioneer in pushing forward digital skills with groundbreaking technology, he has played a major role in the transformation of the examination and certification industry over the past 30 years. He remains committed to enhancing the lives of others through his daily work and advisory work to a handful of boards. He is fluent in English, French, Greek and Turkish, and holds a BBA from Bosphorus University and an MBA from the University of La Verne.

Marios Molfetas is Chief Languages Officer and Executive Director at LanguageCert, having previously been Business Development Director and Marketing & Communications Manager. He monitors all contracts relevant to activities outsourced to LanguageCert. He is responsible for sales and marketing, as well as for the development and execution of LanguageCert's business development strategy.

Maria Babatsi is an Academic Associate for Marking at LanguageCert. She is responsible for ensuring that all LanguageCert's assessment processes conform to quality standards. She delivers training sessions, liaises with and monitors exam personnel on an ongoing basis. She holds a BA in English and has extensive teaching and assessment experience.

Angeliki Cheilaris is Head of Assessment at LanguageCert. She is responsible for the IESOL exam development. She has an educational background in Linguistics and Language Teaching, is Cambridge Delta qualified (Modules 1, 3) and she holds an MSc in Cultural Organisations Management. She is currently pursuing a master's degree in Teaching English as a Foreign/International Language.

David Coniam is Head of Research at LanguageCert. He has been working and researching in English language teaching, education and assessment for almost 50 years. His main publication and research interests are in language assessment, language teaching methodology and academic writing and publishing.

Peter Falvey is an Honorary Professor at The Education University of Hong Kong. He is a teacher educator and a former Head of Department in the Faculty of Education of the University of Hong Kong. His main publication and research interests are in language assessment, second language writing methodology, and text linguistics.

Leda Lampropoulou is Head of Assessment at LanguageCert, with a focus on the research and validation of speaking exams. In her role, she coordinates the development of speaking tests, ensuring they are fit-for-purpose. She holds a BA in English with Philosophy from London University and an MA in Language Testing from Lancaster University. She is also CELTA qualified and a member of UKALTA.

Tony Lee is Senior Psychometrician at LanguageCert. He has been involved in language assessment statistical analysis work since 1980 in universities in Hong Kong and Australia. His major language assessment work includes the assessment management of the Australian Federal Government's migrant English assessment system ACCESS as well as the Hong Kong Government's English Language Ability scale.

Michael Milanovic, previously CEO of Cambridge Assessment English, has been working extensively with PeopleCert since 2015, and is Chairman of LanguageCert and a member of its Advisory Council. He worked closely with the Council of Europe on its Common European Framework of Reference, has held, and still holds a number of key external roles.

Yiannis Papargyris is an education management professional with over 15 years' experience in the fields of English-medium Higher Education, Qualification Development and Educational Assessment. At PeopleCert, he holds the position of Language Assessment Development Manager and is responsible for the development of the LanguageCert exams portfolio.

Tzortzina Peristeri is Editor of Classical Greek at LanguageCert. She holds a BA in Greek Philology and an MA in Classics and Ancient History. She has considerable experience in teaching Classics and in assessment. She is currently contributing to the editing of Classical Greek assessment materials.

Nigel Pike is highly experienced in assessment, and was Director of Assessment at Cambridge Assessment English, directing the delivery of all Cambridge English examinations. Nigel holds an MBA, and has extensive experience with national and local ministries of education around the globe, delivering consultancy, customised examinations and developing language policy for governments.

Xenia Poupounaki Lappa is a Project Manager at LanguageCert. She coordinates and handles various projects, focusing on the development of new language qualifications. She has an academic background in Linguistics and Language Teaching, as well as an extensive experience in the field of ELT and Educational Assessment. Recently, she has been contributing to the development of a high-quality multilingual exams portfolio in her capacity as Project Manager.

Common Abbreviations used in the Book

3PL	Three-Parameter Logistic model
ALTE	Association of Language Testers in Europe
ANOVA	Analysis of Variance
APA	American Psychological Association
AT	Adaptive Test
AWA	Automated Writing Assessment
C&G	City & Guilds
CEFR	Common European Framework of Reference
CoE	Council of Europe
CRELLA	Centre for Research in English Language Learning and Assessment at the University of Bedfordshire
CS	Carnegie Speech
CTS	Classical Test Statistics
CTT	Classical Test Theory
DIF	Differential Item Functioning
EdUHK	The Education University of Hong Kong
EFL	English as a Foreign Language
ELT	English Language Teaching
ESOL	English to Speakers of Other Languages
FOR	Frame of Reference
GMAT	Graduate Management Admission Test
HKCE	Hong Kong Certificate of Education
IELTS	International English Language Testing System
IESOL	International English for Speakers of Other Languages
IMNSQ	Item Infit Mean Square
IRT	Item Response Theory

LC	LanguageCert
LID	LanguageCert Item Difficulty scale
LTA	Latent Trait Analysis ,
LTCG	LanguageCert Test of Classical Greek
LTE	LanguageCert Test of English
MC	Multiple-Choice
MFRA	Multi-faceted Rasch Analysis .
MOOC	Massive Online Open Course
NARIC	National Recognition Information Centre
NLP	Natural Language Processing
Ofqual	Office of Qualifications and Examinations Regulation
OLP	Online Proctoring
OMNSQ	Item Outfit Mean Square
PB	Paper-Based
PBC	Point Biserial Correlation
PBM	Paper-Based Marking
PDT	Performance Dimension Training
PPM	Pearson Product-Moment Correlation
PTME	Point Measure Correlation
QCA	Qualifications and Curriculum Authority
SD	Standard Deviation
SE	Standard Error
SEM	Standard error of measurement
TCC	Test Characteristic Curve
TEA	Technology Enhanced Assessment
TLU	Target Language Use
TOEFL	Test of English as a Foreign Language
UKVI	UK Visas and Immigration

Introduction

Peter Falvey

This volume is a compilation of research studies conducted in 2020 aimed at supporting LanguageCert's research-led, quality oriented approach to its language assessments. The volume is subdivided into four areas, comprising fifteen separate chapters addressing a variety of assessment topics.

Section 1: Background

This section consists of two chapters, each of which helps to describe the background to assessment research at Language Cert.

Chapter 1 *LanguageCert – A Multilingual Language Testing System* (Milanovic & Cheilaris) describes LanguageCert's background, orientation towards assessment and the CEFR (Common European Framework of Reference), and language as communication. It includes an introduction to research and validation at LanguageCert, and, in this context, discusses the socio-cognitive Weir framework (2005) that underpins the LanguageCert approach to research and validation.

Chapter 2 *External Validation of LanguageCert's English Language Examinations* (Papargyris & Lampropoulou) deals with an issue vital for the creation, development and administration of all tests – validity. In the normal use of the term validity, a simple but profound question is posed: does the test test what it is supposed to test? Chapter 2, however discusses a different but no less important meaning of validity. That question is: has the test been externally validated? External validation occurs when a body or institution, separate and independent from the testing agency itself, examines the ways in which the tests that are created, trialled, administered, analysed and evaluated by the testing body fit four criteria: fairness, integrity, honesty and efficiency. The chapter concludes that Language Cert's tests meet those criteria.

Section 2: Explorations Into Test Quality

This section, containing four chapters, examines how quality is maintained in LanguageCert tests.

Chapter 3 *Gauging Quality in the IESOL LanguageCert Listening and Reading Tests* (Coniam) examines the statistical quality of the Listening and Reading Tests in LanguageCert's (LC) IESOL suite of exams produced from 2018-2020. Classical test statistics (CTS) are the assessment tools described primarily in this paper, that is: test

means, reliabilities, standard deviations and standard errors of measurement. Analysis of these tests reveals a series of well-constructed tests, with high reliability and small standard errors of measurement.

Chapter 4 *Examiner Quality and Consistency across LanguageCert Writing Tests* (Coniam & Papargyris) deals with an equally important issue in assessment – the quality of examiners and their consistency when marking. A test may be well-constructed and valid but if the examiner is a poor judge and /or is inconsistent in awarding marks, the whole process is invalidated. Testing bodies must, therefore, conduct regular checks and training for their examiners to ensure standards of high quality and consistency. This chapter reports on a study of the training and standardisation of examiners who mark LanguageCert's International ESOL (IESOL) suite of English language tests linked to the Common European Framework of Reference (CEFR).

Chapter 5 *Potential Bias in LanguageCert IESOL Items: A Differential Item Functioning Analysis (DIF)* (Coniam & Lee) describes how LanguageCert works to avoid potential bias in their IESOL test items. It explores whether any subgroup of test takers sitting a test or exam is being unfairly disadvantaged or indeed advantaged. The need to avoid test bias is a necessary though complex requirement of leading test providers. Work on test bias and DIF is a comparatively recently analytic technique. The methodology required to carry it out has been identified by researchers and is described in this paper where four variables, namely mother tongue, age, gender and test centre are explored.

Chapter 6 *Task Equivalence in LanguageCert IESOL Writing Tests* (Coniam & Babatsi) describes how LanguageCert ensures that across a large battery of test forms, with every test form using different writing tasks, equivalence is established and maintained (the term "test" is used as a generic form for an individual test or subtest. The term "test form" is used to indicate parallel, or multiple, versions of the same test). This procedure to establish equivalence is crucial so that fairness to candidates is assured. Ensuring task equivalence means that the score achieved by a candidate in the assessment of writing is a function of candidate ability as opposed to task difficulty. The chapter explores the extent to which the difficulty of writing tasks varies across LanguageCert IESOL examinations at levels B2 and C1. The issue is investigated using Multi-faceted Rasch Analysis (MFRA), which confirms a high degree of comparability across test tasks in terms of difficulty.

Section 3: Calibration Studies

Section 3 consists of three chapters each discussing the methods used to calibrate LanguageCert tests. Chapter 7 describes calibration studies for LanguageCert's Tests of English, while Chapter 8 describes how calibration was developed for the LanguageCert Test of English Adaptive Test. Chapter 9 describes the development of an innovative method of ensuring that calibration can take place between language test materials and the alignment of test forms through the use of a methodology that provides external, not internal, anchoring of test items by equating expert judging and Rasch Measurement values in LanguageCert tests.

Chapter 7 *Validating the LanguageCert Test of English Scale: The Paper-based Tests* (Coniam, Lee, Milanovic & Pike) documents the first phase of measurement scale development for the LanguageCert Test of English (LTE). Measurement scales are basic requirements in any reputable set of tests. An appropriately validated measurement scale is a necessary prerequisite for any examination/assessment system. The study describes the validation of the initial LanguageCert Item Difficulty (LID) scale which was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement. The study builds on and develops the original

LanguageCert Item Difficulty scale through the use of Item Response Theory (IRT) and Rasch analysis in addition to expert judgement and CTS. This enhanced LID scale will form the empirical basis for the alignment of all current and future assessment products to the same measurement scale that is itself aligned to the CEFR.

Chapter 8 *Calibrating the LanguageCert Test of English Adaptive Test* (Coniam, Lee, Milanovic & Pike) outlines the calibration of the LanguageCert Test of English (LTE) adaptive test to the previously-calibrated scale developed from the LTE paper-based tests (Coniam et al., 2021). Following an overview of computerised adaptive testing (CAT), a description is then provided regarding the development of CATs, and how CATs function and operate. The LanguageCert adaptive test is then described, along with an overview of the algorithm used by the LTE CAT. Discussion then shifts to a description of the development of the LTE CAT and its related item bank. Through reconciling the different frames of reference of the adaptive and the paper-based tests, a picture is provided of how a coherent and comprehensive LanguageCert LTE scale has been developed, with the LTE scale linked to an item bank that provides both anchoring from individual tests with different frames of reference and individual item-based adaptive tests.

Chapter 9 *Externally-Referenced Anchoring: Equating Expert Judgement and Rasch Measurement Values in LanguageCert IESOL English Language Tests* (Coniam, Lee, Milanovic & Pike) the third chapter in Section 3 on test calibration, reports on the use of externally-referenced anchoring by LanguageCert as a methodology for calibrating language test materials and aligning test forms. This is an important development because it introduces an innovative methodology that provides external, not internal anchoring of test items by equating expert judging and Rasch Measurement values in LanguageCert tests. The datasets used in this paper are taken from tests at each of the six levels of the LanguageCert IESOL suite, all of which have been aligned to the CEFR through expert judgement. The chapter illustrates the extent to which externally-referenced anchoring, using Item Response Theory (IRT) but also based on expert judgement, can be used as an effective, reliable and valid methodology.

The findings of this study indicate that, while the match between the distribution of items in the selected LanguageCert IESOL tests and the LID scale was not perfect, in general, evidence of a relatively close match between the items in the tests and the LanguageCert Item Difficulty (LID) scale was found. As a consequence, they matched the corresponding CEFR level.

Section 4: Original Research

Section 4 consists of five diverse chapters. Each chapter contains descriptions and discussions of a wide range of research, from: the identification of guessing in language tests using Rasch analysis; the development and delivery of online proctoring in speaking examinations; the perceptions of candidates to online proctoring; a study of automated writing assessment, and, finally; a description of the development of a communicative test for Classical Greek.

Chapter 10 *Identifying Guessing in English Language Tests via Rasch Fit Statistics: An Exploratory Study* (Coniam, Lee & Lampropoulou) explores the issue of identifying guessers (candidates who simply guess at the appropriate answer) – with a specific focus on multiple-choice tests. Guessing is a long-standing issue because it compromises validity. A test taker scoring higher than they should through guessing does not provide an authentic picture of their actual ability. A description of issues associated with guessing is first presented, followed by

a discussion of approaches that have been taken to either discourage test takers from guessing or attempt statistically to handle the problem. A novel way of identifying potential guessers: from the post hoc use of Rasch fit statistics is then revealed whereby when using Rasch fit statistics to identify possible guessers, it was possible to identify 80% of the guessers.

Chapter 11 *The Development and Delivery of Online-Proctored Speaking Exams: The Case of LanguageCert International ESOL* (Lampropoulou & Papargyris) describes a study responding to commercial requests for the online delivery of its exams. LanguageCert embarked on an attempt to develop an online-proctored (OLP) equivalent to its established International English for Speakers of Other Languages (IESOL) Speaking exam suite. The study describes the practical aspects of replicating the exam in an online environment, together with the potential need to adjust the content and format of the test, the applicable variants of the exam registration and exam administration processes, security and test integrity issues, as well as any potential impact on the assessment methodology the exam employs. Most importantly, it asserts that the accuracy of assessment outcomes is not compromised in any way by the mode of delivery.

Chapter 12 *Online Proctoring of High-Stakes Examinations: A Survey of Past Candidates' Attitudes and Perceptions* (Coniam, Lampropoulou & Cheilaris) reports reactions by candidates to the use of online proctoring (OLP), 'invigilation', in the delivery of high-stakes English language examinations. The chapter first describes the move from face to face to online modes of delivery. It then explores the challenges and benefits that both modes offer, in terms of accessibility, fairness, security and cheating. Evidence is then presented from a survey exploring the reactions to and perceptions of OLP by candidates who had taken an English language examination via OLP. A strong endorsement of OLP was generally recorded. Feedback revealed that respondents perceived OLP to be a more personal as well as a more efficient way of taking a test. Some pertinent negative comments from a smaller number of respondents could be construed as constructive and are also discussed. The results are indicative of a broad acceptance of OLP, pointing to strong future uptake of the OLP mode of test delivery.

Chapter 13 *Automated Writing Assessment (AWA) and the Carnegie Speech Writing Assessment System* (Coniam & Lee) reports on a study to validate the use by LanguageCert of the Carnegie Speech (CS) Automated Writing Assessment (AWA) system. Following a brief introduction to the computer assessment of writing, the report details a study using a reference dataset from the Hong Kong Year 11 public exam (Coniam, 2009) comprising 300 scripts – largely at CEFR B2 level – where the marker statistics and candidates' overall subject score on the public examination were known background variables. The study explores two key issues. The first is the reliability of the CS AWA system compared with that obtained in the previous study. Second, the study explores a methodology where the output of the CS AWA system (linked to the nine-point IELTS scale) can be equated with LanguageCert's Writing test scale where both scales are also aligned to the CEFR. The analyses performed confirm that the scores produced by Hong Kong (human) markers correlate at a moderate-to-high level with those produced by the CS AWA system. Furthermore, the Rasch methodology used to equate the two scales (the six-point HK scale and the nine-point CS AWA / IELTS scale) illustrates that the scores produced on the two scales might be successfully aligned.

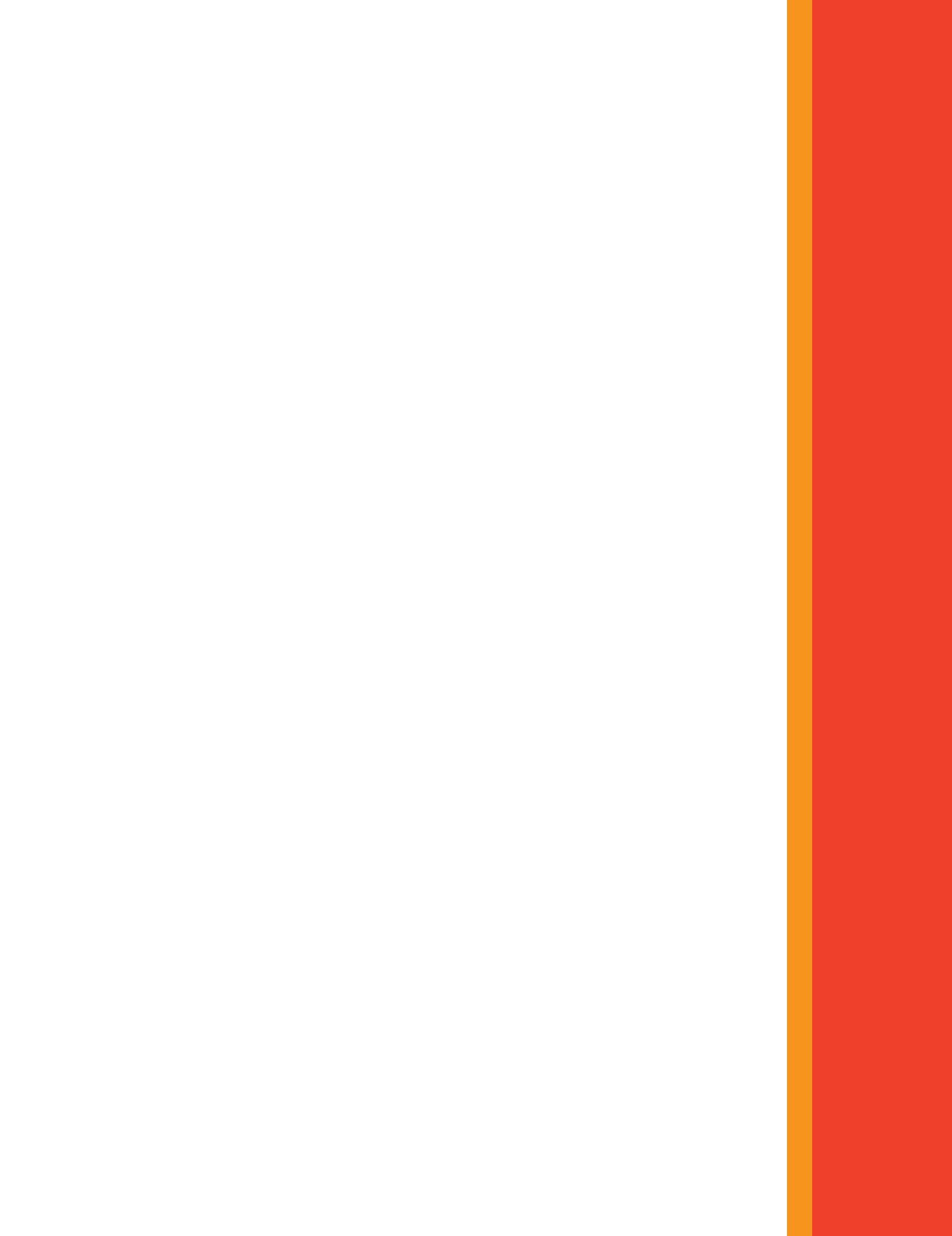
Chapter 14 *Towards a Communicative Test of Reading and Language Use for Classical Greek* (Poupounaki-Lappa, Peristeri & David Coniam) describes the development of a communicative test of Reading and Language Use for Classical Greek, aimed at students at CEFR (Common European Framework of Reference for Languages) levels A1 and A2. A discussion is first provided of traditional pedagogical approaches which have for many decades dominated the teaching of classical languages, followed by suggestions why these may be supplanted with more modern communicative approaches. Focus then moves to assessment, where, it is suggested, methods are equally rooted in traditional, form-focused methods. If teaching is to become more communicative, it is argued, so should assessment. Against this backdrop, the development of a test of Reading and Language Use for students of Classical Greek at CEFR levels A1 and A2 is described.

Chapter 15 Recapping and Looking Ahead pulls all threads in the volume together and looks towards the future.

A **Glossary**, at the end of the book, provides the reader with an overview of the statistical terms and methods used throughout the volume (Falvey).

References

Weir, C.J. (2005). *Language Testing and Validation*. Palgrave MacMillan: London.





SECTION 1: BACKGROUND



Chapter 1: LanguageCert – A Multilingual Language Testing System

Michael Milanovic and Angeliki Cheilari

Abstract

This paper presents an introduction to LanguageCert: its background, orientation towards assessment and the CEFR, and language as communication. An introduction is provided to research and validation at LanguageCert, together with the Weir framework which encompasses the LanguageCert approach to research and validation.

Background

LanguageCert, a part of the *PeopleCert* group, is a leading international assessment company. It administers a suite of language examinations, currently across several languages, that includes examinations for primary/elementary school students, secondary school students, young adults in the workplace or study and adults in the workplace. The *LanguageCert* system has been operating since 2015 and in that time has established a number of examinations across three languages (English, Spanish, Turkish) in the first instance. Central to *LanguageCert*'s mission is a strong focus on quality, validation and research to support the development and use of its examinations.

The initial focus of *LanguageCert*'s research and validation programme is the *LanguageCert International ESOL* (English for Speakers of Other Languages) examination, a set of English Language qualifications, each targeting a different level of the CEFR (Common European Framework of Reference) – from CEFR A1 to C2. These examinations are intended for teenagers and adults or those preparing for entry to higher education or professional employment.

The Common European Framework of Reference (CEFR)

The CEFR emerged as a recognised framework for learning and assessment in Europe in the 1990s. The Framework classifies language learners into three broad levels: Basic Users, Independent Users and Proficient Users, with each level then broken down into two sublevels.

The CEFR aims to encourage and facilitate reflection and communication in language education; and, in the context of examinations, the CEFR is intended to assist assessment providers, publishers, teachers, learners and other relevant stakeholders to articulate both content standards (the nature of the skills being tested) and performance standards (levels of proficiency).

LanguageCert and the CEFR

There are six examinations in the *LanguageCert* International ESOL suite, all aligned to the six CEFR levels: Preliminary (A1), Access (A2), Achiever (B1), Communicator (B2), Expert (C1) and Mastery (C2). The examination specifications reflect the requirements of the CEFR; test materials writers represent the highest international standards and have extensive expertise in, and knowledge and understanding of, the CEFR.

Importance of the CEFR

The CEFR has come to be accepted across Europe (and indeed beyond, with many countries linking their language curricula, syllabuses and examinations to the CEFR) as a specification of common standards across many different European languages. The Framework lays out a set of common standards which permit employers and educational institutions to evaluate the language qualifications of test takers applying for employment or admission to education.

Alignment of *LanguageCert* and the CEFR

One of the most widely applied and helpful uses of the CEFR is to facilitate the comparison of language curricula, syllabuses and examinations. Given that many international language assessment systems claim alignment to the CEFR, this provides a useful way of comparing language examination levels.

Communicative Language Testing

LanguageCert's English language examinations follow a rigorous test development process to ensure that validity – possibly the most crucial aspect in communicative assessment – is achieved. In the context of language tests supporting judgements made with respect to certain *Target Language Use* domains (see e.g., Bachman & Palmer, 2010), what might be expected of test takers in real world language use needs to be considered and defined. In this context, the CEFR has been taken as the driving force determining the constructs underpinning *LanguageCert's* examinations. Its illustrative descriptors across a range of language domains and contexts have been used as a starting point and extensively inform the test development processes employed in the development of the examinations.

Testing Language as Communication

The task types used in the *LanguageCert* examinations have been selected to ensure they have interactional authenticity and can be related to real-world performance. They have been developed to directly sample the cognitive skills, strategies and language knowledge that support judgements made regarding the potential ability of test takers in real-world interactional situations.

For example, in the receptive skills, reading tests are carried out using real-world notices and signs that students are likely to encounter to show their understanding of them. Reading also focuses on aspects of comprehension of texts at the appropriate level using a variety of task types, such as multiple-choice selection, matching, and providing open responses to prompts. The degree of challenge and sub-skills required in the reading tests start with:

1. relatively simple tests of understanding at 'Basic User' A levels
2. text interpretation and understanding writer intention at the 'Independent User' B levels
3. complex interpretation strategies required at the 'Proficient User' C levels that require inferring the meaning of unknown lexis from context and the ability to sift through lengthy reading inputs to identify key information and to show awareness of text cohesion.

The Importance of Testing Communication

Validity may be defined as the extent to which a test measures its intended purpose. As the intended purpose of the *LanguageCert* tests is communicative English language proficiency, the appropriateness and meaningfulness of results are key factors. This is because they demonstrate that, along with reliability, test takers have obtained results that validly reflect their real-world performance. These factors are important in the context, for example, of *LanguageCert* being an approved provider of language proficiency for UKVI (UK Visas and Immigration) purposes. *LanguageCert* is now recognised by government bodies, test takers, teachers and the public at large as trustworthy reliable and valid examinations.

The Place of Research

Although a comparatively new entrant in the language assessment arena, *LanguageCert* considers research – into systems, communication as well as its examinations, tasks and items – a key underpinning factor in its makeup. Since acquiring the suite of examinations from City & Guilds in 2015, research, validation and development have been high on *LanguageCert's* agenda. Internally, there is a research and validation team which investigates the consistency of the materials as well as their setting and the marking.

Validation

Construct validation activities are carried out by *LanguageCert* beginning with test and task design. Experts analyse examination tasks and content in an ongoing manner to ensure they are fair, have interactional authenticity and sample the appropriate language skills for any given level and skill.

Different test forms in the *LanguageCert* International ESOL suite may be considered comparable in terms of content and difficulty due to robust item-banking following the pretesting and trialling before use of all appropriate examination material. In order to ensure quality and the validation of levels, examinations are monitored through ongoing independent external research.

A Framework for Validation

The Weir (2005) Framework provides a useful way of structuring a research and validation programme, where the six categories in the framework allow for a range of characteristics and factors to be taken account of. These include test taker characteristics; contextual characteristics in terms of fairness; cognitive processes required to complete tasks; how far scores may be depended on; the impact of the tests; and external evidence to show that the test is doing what it is intended to do. Weir's framework provides a comprehensive structure by which the different factors and elements associated with the test-making and test-taking process may be coherently presented – with the structure allowing access by different users (test taker, teachers, academics, for example).

LanguageCert Use of the Weir Framework

At a basic level, the Framework has been useful in the setup of the *LanguageCert* website, with the categories driving the ways users navigate through the website. External validation has already been conducted on certain examinations in the IESOL suite (a comprehensive evaluation of the B2 test, with shorter evaluations of other tests); the relevant reports are then located under appropriate Framework headings. The makeup of the Framework also allows for sensible location of internal validation documentation such as marker standardization data and comparability data on different tests forms; similarly, background documentation such as Item Writer Guidelines, Marker Guidelines – which *LanguageCert* makes transparently available – fit cleanly into the Framework.

References

- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
Weir, C.J. (2005). *Language Testing and Validation*. Palgrave MacMillan: London.

Chapter 2: External Validation of LanguageCert's English Language Examinations

Yiannis Papargyris and Leda Lampropoulou

Abstract

LanguageCert began administering its own English language examinations in 2017. Since that time LanguageCert has undergone a series of external validations to provide evidence of the robustness of its examinations, and to provide proof of their validity, reliability and their being fit for purpose. Two key external studies referred to below are those conducted by:

CRELLA – the Centre for Research in English Language Learning and Assessment at the University of Bedfordshire. In 2018, CRELLA was commissioned to investigate LanguageCert's B2 level test and its relationship to the CEFR.

NARIC – the National Recognition Information Centre for the United Kingdom, the Government agency for the recognition and comparison of International qualifications and skills. In 2018, UK NARIC was commissioned to conduct an independent assessment of LanguageCert's ESOL examinations.

The current document first provides a brief outline of key concepts to situate the external studies. Following this, the key issues, major conclusions and recommendations from the external studies are presented.

At a basic level, the Framework has been useful in the setup of the LanguageCert website, with the categories driving the ways users navigate through the website. External validation has already been conducted on certain examinations in the IESOL suite (a comprehensive evaluation of the B2 test, with shorter evaluations of other tests); the relevant reports are then located under appropriate Framework headings. The makeup of the Framework also allows for sensible location of internal validation documentation such as marker standardization data and comparability data on different tests forms; similarly, background documentation such as Item Writer Guidelines, Marker Guidelines – which LanguageCert makes transparently available – fit cleanly into the Framework.

Validity and the LanguageCert IESOL Examinations

The LanguageCert International ESOL tests are designed to ensure fitness for purpose and to deliver assessments which take into account contemporary views of validity. Validity is generally defined as the extent to which a test measures its intended purpose. In the case of LanguageCert, this is communicative language ability (see e.g., Bachman & Palmer, 2010) and the foreign language specifications provided by the Council of Europe in such documents as Waystage (1990) and Threshold (1990). The qualities of validity (and reliability) need to be considered together in order to ensure fairness to candidates and to generate trusted result outcomes that will replicate real-world performance of candidates.

The test development process underpinning LanguageCert's English language exams has been established to ensure validity is achieved. Bachman & Palmer (2010) state that language tests should support inference to some domain of 'Target Language Use' (TLU). That is, in order to judge the validity of test results, what a test-taker is expected to be able to do in real-world language use must be laid out. The Common European Framework of Reference (CEFR) has been utilized to help determine the test construct of the LanguageCert exams for this purpose. Its illustrative descriptors across a range of language domains and contexts have been used as a starting point and extensively inform the test development processes employed.

The task types used in the LanguageCert examinations have been selected to ensure they have interactional authenticity and can be related to real-world performance. They directly sample the cognitive skills, strategies and language knowledge that support inference about the potential ability of a candidate in real-world interactional situations.

In this manner, validity links performance on the tasks in LanguageCert International ESOL tests to an inference about the test taker's ability in a world beyond the test. The tests are designed to elicit a sample of performance which is interpretable and generalizable to the real world. In order to ensure the test results are generalizable CEFR Can-Do statements have been used as the basis for what test-takers need to be able to achieve at each level.

Achieving Reliability

Reliability relates to consistency in test results. This is achieved in the LanguageCert International ESOL tests by ensuring test forms are comparable in terms of content and difficulty, and through robust item-banking techniques, involving the pretesting and trialling of test materials and the placement of all items on the LanguageCert Item Difficulty (LID) scale.

Reliability is crucial for all test stakeholders who need to be sure that different administrations of the test deliver very similar results. This is essential for fairness to test-takers and to ensure that receiving institutions such as universities and employers can be guaranteed that the same ability level is required to pass the same examination at different administrations. The start of the process of ensuring reliability of results is to ensure standardisation of test-taking experience. This begins with test specifications that ensure tests can be replicated over years of administrations, through standardised test-taking conditions and finally through the difficulty of the test materials and the way tests are graded.

Historical LanguageCert-related Validation Study

LanguageCert acquired a range of IESOL test materials from City & Guilds, UK in 2015. Prior to being acquired by LanguageCert, the quality of some of the City & Guilds examinations was put through significant external validation, the most prominent study being:

O'Sullivan, B. (2009). City & Guilds Communicator Level IESOL Examination (B2) CEFR Linking Project Case Study Report. Roehampton University, UK.

This study was thorough and extensive, and reported positive outcomes regarding the makeup of the City & Guilds B2 level test, supporting the claims about the test's links to the CEFR.

The Executive Summary may be found in Appendix 1. The full report may be accessed on the LanguageCert website.

Recent External Validation Studies of LanguageCert Tests

Since administering its own English language examinations in 2018, two large-scale validation studies of LanguageCert examinations have been conducted. These have been:

Validation Study 1	Green, A. 2019. Relating LanguageCert Communicator to the CEFR. Centre for Research in English Language Learning and Assessment: University of Bedfordshire, UK.
---------------------------	---

This thorough and extensive examination of the B2 test was conducted by CRELLA, the Centre for Research in English Language Learning and Assessment at the University of Bedfordshire.

While there were some recommendations for LanguageCert to consider, the findings from the study strongly support the claim that material throughout the Spoken and Written Exams closely reflected the B2 level.

The Executive Summary may be found in Appendix 2. The full report may be accessed on the LanguageCert website.

Validation Study 2	National Recognition Information Centre for the United Kingdom. 2019. LanguageCert ESOL International Qualifications: Independent CEFR Referencing—Summary Report. UK NARIC: Cheltenham, UK.
---------------------------	---

The National Recognition Information Centre for the United Kingdom (NARIC), the UK Government agency for the recognition and comparison of International qualifications and skills was commissioned in 2018 to conduct an independent assessment of LanguageCert's ESOL examinations. The independent assessment was

a mandatory pre-requisite, for organisations wishing to be eligible to participate in a UK government procurement for English language testing.

In its evaluation, UK NARIC recognised that LanguageCert's IESOL qualifications had been developed with CEFR as the source document. The extensive evaluation deemed that the item writing process underwent clear technical and content checks. Following successive investigations into standard setting, vetting, statistical analysis, modification, proofreading and finalisation, UK NARIC determined that CEFR alignment was evident at all stages of test development and delivery.

The Executive Summary may be found in Appendix 3.

References

- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
Van Ek, J. A., & Trim, J. L. M. (1991). *Waystage 1990*. Manhattan Publishing Company.
Van Ek, J. A., & Trim, J. L. M. (1991). *Threshold level 1990*. Council of Europe.).

Appendix 1

O'Sullivan, B. 2009. City & Guilds Communicator Level IESOL Examination (B2) CEFR Linking Project Case Study Report. Roehampton University, UK.

Executive Summary

Background

This project was a joint undertaking by City & Guilds and the Centre of Language Assessment Research (CLARe) at Roehampton University. The object of the project was to provide evidence of the validity of City & Guilds' Communicator examination, particularly in relation to the central claim that it is aimed at Level B2 in the Common European Framework of Reference for Languages (commonly referred to as the CEFR). In doing this, it was planned that the project would act as a formal review of the existing examination, and it was planned that any areas of concern within the papers would be identified and brought into line with best practice in the area.

The Communicator (and the other examinations in the suite) was developed using the CEFR (Council of Europe 2001) as source document to inform the assessment tasks, specifications and assessment criteria. During the development phase, however, the Draft Manual (2003) for relating examinations to the framework was not in existence, so the organisation embarked on a series of internal activities to ensure alignment to the external standards. However, with the publication of the Manual the logical step for the organisation was to register as a case study for operationalising the concepts and processes encapsulated there.

A secondary aim of the project was to provide feedback to the Council of Europe on their Draft Manual (2003) which was used as a basis for the methodology.

Methodology

As mentioned earlier, the methodology used in the project was based on the procedures recommended by the Council of Europe in their Draft Manual of 2003. However, as the project progressed a number of changes were made to facilitate the operationalisation of the process. The project adapted the four-stage approach suggested in the Draft Manual:

1. Familiarisation
2. Specification
3. Standardisation
4. Validation

In terms of the methodology used, a number of important recommendations were made, these related to the nature of the process (which we suggest is iterative rather than linear as implied in the Draft Manual) and the notion of embedding the process in the institution's test development cycle.

Summary of the Main Findings

The main findings of the project can be summarised as follows:

1. It was found that in order to claim a link to the CEFR at Level B2 the cut score for a passing grade for the Communicator Reading paper should be set at 15 (from a maximum of 30). The same cut score was recommended for the Communicator Listening paper. This is actually in line with current practice for Communicator.
2. Passing levels for the Communicator Writing paper were found to be in line with the Council of Europe recommended tasks for CEFR Level B2. The recommendation is that the cut level for this decision should not be altered at this point in time.
3. The linking process is long and demanding, both at the individual and institutional level. The complexity of the design means that it is expensive for any institution to undertake, certainly to the extent undertaken by City & Guilds in this project. While this perhaps explains the reluctance of many examination boards to undertake a full linking project, we nevertheless recommend that the process be extended to as many of the other examinations in the ESOL suite as feasible.
4. Unless the test which is the focus of the linking project is shown to be robust in terms of quality and level, there is no point in even starting a linking project, as the process is unlikely to succeed beyond the standardisation stage without serious issues emerging. In fact, we feel that with a more demanding specification phase, issues should emerge more clearly at this early stage.
5. Limiting the validation evidence to estimates of internal and external validity is far too simplistic a view of validation. The CEFR should be demonstrated to impact on all aspects of the test, from the test taker to the task to the psychometric qualities and relative meaning or value of the test score.

Based on this project, it is the belief of the project team that the evidence presented here supports the claim that the Communicator tests English ability at CEFR Level B2.

We feel that the process of linking the Communicator examination to the CEFR, has resulted in systematic and sustainable improvements to the test and to the system that supports the test.

It is clear to us that the process has resulted in a test that is more clearly at level, is sound from an internal psychometric perspective and is more replicable and of a high quality. However, that is not all. The systems that support the examination have also been systematically improved and more explicitly linked to the CEFR. The item writers' guidelines are, we believe, up-to-date and more robust than in the past. The specifications are now more likely to result in accurate replication of tests on level – one criticism of the old specification was the lack of detail and exemplification, this appears to have led to a tendency to drift away from the level. This is a warning for other test developers, who take time to specify their tests but do not routinely review these specifications (and their use) to ensure that there is no level or construct drift.

We now feel that we are in a position to consider suggesting a number of Communicator tasks to the Council of Europe for use as recommended level indicators in future linking projects.

Appendix 2

Green, A. 2019. Relating LanguageCert Communicator to the CEFR. Centre for Research in English Language Learning and Assessment: University of Bedfordshire, UK.

Abstract

This study was undertaken to relate the LanguageCert Communicator Exam to the Common European Framework of Reference (CEFR; Council of Europe 2001). It includes both the Spoken (Speaking) and Written (Listening, Reading, Writing) Exams for which separate certificates are awarded. The study employed the staged approach recommended by the Council of Europe (2009) which includes Familiarisation, Specification, Standardisation, Benchmarking/ Standard setting and Validation.

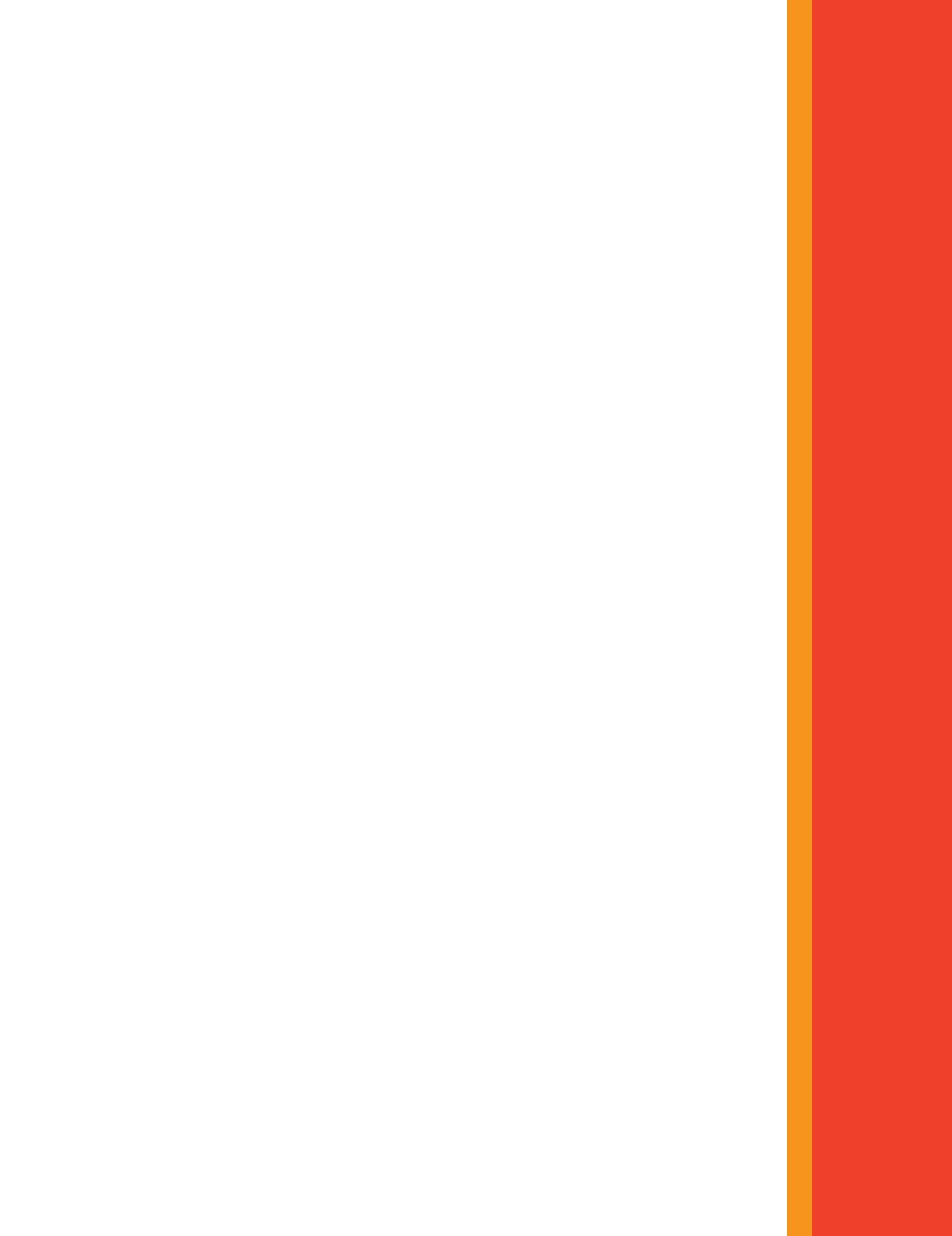
Following Familiarisation, which involves building and confirming understanding of the CEFR, Specification was carried out by LanguageCert staff in collaboration with the researchers. This made use of a standard text template developed from the forms used in the Council of Europe (2009) Manual, but designed to better convey the outcomes to test users and other stakeholders.

Benchmarking and Standard setting combined a qualitative perspective based on the analysis of test materials and rating scales with the 'Benchmarking with FACETS' approach suggested by North and Jones (2009) which makes use of calibrated performance samples and cut scores for the CEFR level descriptors. The twin-panel approach involved two-day meetings in Greece and the UK between a total of 16 expert panellists (nine meeting in Athens and seven in Luton). The panellists reviewed test material and sample performances and related these to the CEFR. The review of material confirmed that all four papers (Listening, Reading, Writing and Speaking) reflected the B2 level of the CEFR in the targeted Communicative Activities.

Findings from the Benchmarking of performance samples and Standard setting panels broadly supported the current interpretation that passing scores on the four Communicator subtests (Listening, Reading, Writing and Speaking) represent B2 on the CEFR in the areas tested, but results from both panels suggested that the current passing scores for B2 should be raised across all four papers.

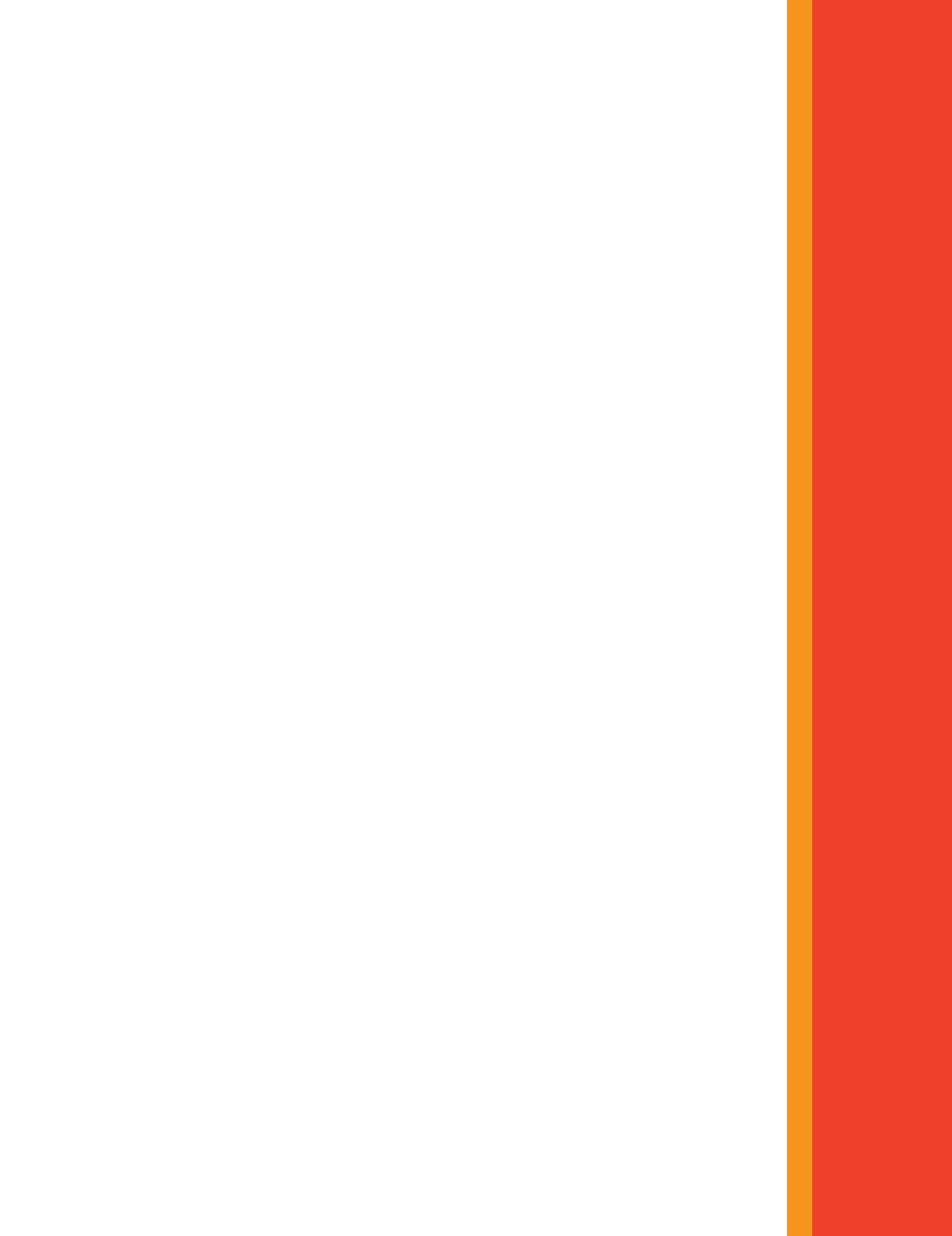
Appendix 3

National Recognition Information Centre for the United Kingdom. 2019. LanguageCert ESOL International Qualifications: Independent CEFR Referencing—Summary Report. UK NARIC: Cheltenham, UK.





SECTION 2: EXPLORATIONS INTO TEST QUALITY



Chapter 3: LanguageCert IESOL Listening and Reading Test Reliabilities 2018-2020

David Coniam and Leda Lampropoulou

Introduction

This paper reports briefly on the analysis of Listening and Reading Tests produced by LanguageCert in its IESOL suite over the period 2018-2020.

All test forms were analysed using classical test statistics, namely: reliability; standard deviation; and standard error of measurement. Intercorrelations between Listening and Reading subtests area also reported.

In summary, all tests have high reliability estimates. Standard deviations show an appropriately broad spread of candidate ability and standard errors of measurement are in a satisfactory narrow range, generally around 5%.

The Data

Each Listening and Reading subtest comprises 26 items, resulting total of 52 items. All test items are either multiple-choice or limited response, the latter requiring test takers to produce a short, written response of between one to eight words depending on the level being attempted. All items are scored either one or zero and marked objectively.

Classical Test Statistics

Table 1 presents a composite picture for the six levels of the key classical test statistics: comprising test means, the median; the range of reliability, the range of standard deviation (as a percentage of the maximum score) and the range of the standard error of measurement (as a percentage of the maximum score).

Table 1. Summary of key test statistics for IESOL A1-C2 tests

	KR20	SEM	SD
A1 Overall	.88-.93	4-5%	10-18%
A1 Listening	.75-.85		9-15%
A1 Reading	.82-.91		12-22%
A2 Overall	.87-.94	4-7%	13-21%
A2 Listening	.75-.89		13-22%
A2 Reading	.83-.90		15-23%
B1 Overall	.86-.94	5-8%	15-22%
B1 Listening	.76-.89		14-23%
B1 Reading	.80-.90		15-23%
B2 Overall	.89-.93	5-6%	15-20%
B2 Listening	.74-.87		16-22%
B2 Reading	.82-.89		16-23%
C1 Overall	.89-.94	4-7%	15-24%
C1 Listening	.78-.90		16-26%
C1 Reading	.81-.90		17-24%
C2 Overall	.87-.90	5-6%	15-18%
C2 Listening	.70-.83		15-19%
C2 Reading	.81-.85		16-20%

Correlations between Listening and Reading Tests

In order to indicate the nature of the relationship between the Listening and Reading components Pearson correlations were calculated. Table 2 presents a summary of the results for each CEFR level. All correlations are significant at the 1% level.

Table 2: Correlations between IESOL Listening and Reading tests

Test level	Correlation Set 1	Correlation Set 2	Mean Correlation
A1	0.64	0.78	0.72
A2	0.74	0.76	0.75
B1	0.72	0.69	0.71
B2	0.68	0.65	0.66
C1	0.79	0.74	0.77
C2	0.65	0.61	0.63

Conclusion

This paper has presented a picture of 62 LanguageCert IESOL tests from A1 to C2 level, developed and administered over the period 2018 to 2020. What has emerged is a picture of a series of well-constructed tests, with high reliability: all tests have reliability figures above 0.80, which is high for tests consisting of 52 items (Ebel, 1965) and 34 have a reliability above 0.9. Standard deviations show broad spread of abilities among the different levels of test, although this is to be expected with proficiency tests, which are open to the public, and able to be taken by any applicant. Standard errors of measurement are in a narrow range, generally around 5%, and indicative of test takers' "true scores" occurring within a range of plus or minus 2 points (5% of 52). With the Pass set at 50% of the total (26/52), the comparatively narrow range of 24-28 may then be confidently taken as the 'boundary'. Correlations between the Listening and Reading tests have emerged as either moderate-to-strong or strong.

References

- Ebel, R. L. 1965. *Measuring educational achievement*. Prentice-Hall, NJ: Englewood Cliffs.



Chapter 4: Examiner Quality and Consistency across LanguageCert Writing Tests

David Coniam and Yiannis Papargyris

Abstract

This paper reports on a study of the training and standardisation of examiners who mark LanguageCert's International ESOL (IESOL) suite of English language tests linked to the Common European Framework of Reference (CEFR). Subjects in the study were a set of examiners ($N=27$) who had been marking LanguageCert's IESOL Writing tests across the six CEFR levels. The focus of the study was on the consistency of marking in terms of severity within and across the six tests that the examiners mark.

Correlations between examiner person measures across all six tests indicated that examiners were broadly consistent across tests, with examiner person measures generally correlating highly with their 'partner' test: A1 with A2, C1 with C2, and B1 with B2 tests. LanguageCert examiners – who undergo careful training and standardisation – may therefore be seen to mark consistently and accurately across a range of ability levels.

Introduction

One of the maxims of assessment is that tests be valid and provide accurate assessments of candidates' abilities, in particular in the context of how far a given test score may be interpreted as an indicator of the abilities or constructs to be measured (Bachman & Palmer, 1996; Messick, 1989). Under such a precondition, the marking of candidates' assessment therefore needs to be accurate if reliable assessments are to emerge. However, such accurate marking in performance assessment involving examiner judgment is an enduring challenge because scores assigned to candidate performance are mediated, interpreted and applied by examiners who are a potential source of error (Engelhard, 2002). From this, it naturally follows that all examiners need to be properly trained and standardised – in particular with performance tests such as Speaking and Writing subjectively-marked.

This paper reports on a study of the training and standardisation of examiners who mark LanguageCert's International ESOL (IESOL) suite of English language tests linked to the Common European Framework of Reference (CEFR). Subjects in the study were a set of examiners ($N=27$) who had been marking LanguageCert's IESOL Writing tests across the six CEFR levels.

The focus of the study was on the consistency of marking in terms of severity within and across the six tests that the examiners mark.

Background to Tests, Examiners and Scripts

The data in the study were drawn from six examinations which comprise LanguageCert's International ESOL suite of English language tests. In the *LanguageCert Writing* tests, candidates complete two writing tasks which elicit a range of writing skills. Responses are marked using an analytic mark scheme which reflects the CEFR descriptors. Separate marks are awarded by marking examiners for different aspects of writing ability – Task fulfilment, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary and Organisation of the text. This set of criteria ensures that a wide range of writing skills are considered, thus enhancing the reliability and representativeness of test scores.

The format of the tests and the nature of the assessment criteria reflect the broad multi-faceted construct underlying these examinations. Communicative ability is the primary concern, while accuracy and range are increasingly important as the CEFR level of the test increases.

Examiner Training

The importance of examiner training in any English language examination is an issue which has long been accepted as an essential factor in determining the reliability of a test (see e.g., Webb et al., 1990). Although empirical studies on examiner training have generated mixed results, a general consensus is that examiner training, if well designed, can improve the reliability and validity of examiner-mediated assessment (Kang et al., 2019). Studies have shown trained examiners to be more reliable (Saito, 2008) as well as more self-consistent (Davis, 2016) than untrained examiners.

In the case of performance-based assessment, it is important to attempt to ensure reliability through extensive examiner training and standardisation, including even sanctioning inconsistent examiners (see Elder et al., 2007).

Webb et al. (1990) discuss the problems associated with examiner stringency, leniency and inconsistency. They state that problems with examiner stringency and leniency can be handled by statistical adjustment. They make it clear nonetheless that examiner training is essential for other problems – specifically, examiner inconsistency. As Weigle (1998) notes, examiner training was more effective in enhancing intra-examiner reliability than inter-examiner reliability. Lumley & McNamara (1995), in discussing inconsistency in examiners report that training and standardisation are not only essential, but also that further moderation is required shortly before the administration of Writing or Speaking Tests because a time gap between the training and the assessment event reveals that inconsistencies re-emerge.

In order to address the issue of consistency, severity and leniency amongst the group of LanguageCert examiners, Multi-Faceted Rasch Analysis (MFRA), via the computer program FACETS (Linacre, 2020) has been utilised. The reader is referred to the outline of the Rasch measurement model and MFRA provided in the Glossary of statistical terms at the end of the volume.

Principles and Procedures in Training Examiners

As stated earlier, in any examination of direct performance it is important to attend to the question of examiner reliability. Although there is no agreement regarding the most effective training and standardisation methods (Kogan et al., 2015), in assessments of performance which rely wholly on examiner applications of the criteria established for the assessment, reliability can be established through a process of:

- agreement on the validity of assessment constructs
- creation of detailed specifications
- creation of valid, detailed and usable descriptors
- provision of credible and regular examiner training and standardisation

See also Feldman et al. (2012), where a cogent summary of different modes of examiner training is provided.

The purpose of standardising examiners is to ensure that strong measures of agreement occur whenever a number of examiners apply grade descriptors to a criterion-referenced assessment instrument. This is the case with the LanguageCert Writing tests. In criterion-referenced assessment, which depends on the application of examiners' judgements to the criteria described in the descriptors, it is important that two principles are adhered to:

- Judgements by one examiner over time with a number of candidates need to be consistent.
- Different examiners judging an individual candidate should provide assessments that are in close agreement.

There are a number of well-established standard procedures that can be used to train and standardise language examiners (see e.g., Coniam & Falvey, 2018). These procedures were applied in the specific training procedures used with trainee examiners for the IESOL Writing Tests and are described below.

Participants

All writing examiners must meet minimum requirements in terms of professional qualifications and experience in order to be eligible for consideration as an examiner. Prospective examiners go through a training process before they are approved and allowed to mark. The training process includes marking sample scripts. Candidates for the examiner role must show they can mark accurately and consistently before they are certificated as examiners. During live marking, if an examiner is found to be marking inaccurately and/or inconsistently, they may be removed from the marking session and/or retrained or dismissed as an examiner. Examiners are then monitored on an ongoing basis and required to attend standardisation meetings on a regular basis.

Participants involved 27 examiners who have been marking LanguageCert's IESOL suite of examinations for a considerable period of time. All 27 examiners marked the A1 and A2 scripts; however, only 24 examiners were available for the other four tests, i.e., B1 to C2.

Standardisation

Examiners were familiar with the rating scales, since they have been using them for five years. The standardisation session described in this paper took place in 2018 and is a regular feature of re-training and standardising undergone by LanguageCert assessment personnel. The process was led by the Chief Examiner, who has marked examinations linked to the CEFR for over 20 years.

Examiners were first given the rating scales and LanguageCert's *Guide for Examiners* and asked to familiarise themselves with the constructs and levels in the scales. Some brief discussion was then followed by two stages of training, Induction and Training, each consisting of the assessment of 36 benchmarked scripts – six per CEFR level – and subsequent discussion of queries, potential discrepancies between raters, the applicability of descriptors, etc. The sample scripts shared with examiners during the Induction and Training stages exemplified the four criteria along with the performance descriptors which constitute the marking scheme.

Over a period of a day and a half, examiners then marked, one test at a time, six scripts from each of the six tests in the LanguageCert IESOL suite (i.e., from A1 to C2). The marking began with the six A1 tests, progressing upwards. After each set of marking and after all examiners had submitted their awarded marks, the Chief Examiner revealed the scores he had awarded and led some discussion about the merits of different scripts.

LanguageCert training and standardisation procedures and practices may be seen therefore to equate with those employed primarily under a performance dimension training (PDT) – see Kogan et al. (2015) – as all three training stages (Induction, Training, Standardisation) are based on the assessment of a series of sample scripts (performances), selected and/or adapted to demonstrate certain issues in candidate performance. To account for potential discrepancies in marking as a result of raters' idiosyncratic tendencies (e.g., leniency), elements of a frame of reference training (FoRT) methodology (see Pam, 2013) were employed so that the role of subjectivity in the application of the marking criteria was minimised. Frame of reference (FOR) training is intended **to get all raters on the same page and reduce idiosyncrasies**

The IESOL Writing Test

The IESOL Writing tests comprise two tasks, as laid out in Figure 1.

Figure 1: IESOL Writing Test Tasks and Scales

Level	Part 1: Candidates produce	Word length	Part 2 : Candidates produce	Word length
A1	four sentences on a specified topic	30	a simple text for a specified reader	20-30
A2	an informal response to an informal text	30-50	a neutral response to a specified public reader	30-50
B1	a neutral or formal text for a public audience	70-100	a letter using informal language	100-120
B2	a neutral or formal text for a public audience	100-150	a text using informal language	150-200
C1	a neutral or formal text for a public audience	150-200	a text using informal language	250-300
C2	a neutral or formal text for a public audience	200-250	a text using informal language	250-300

Concerning marking, all tasks conform to CEFR 'can do' statements for writing and are assessed on a four-point scale on four domains. Figure 2 illustrates.

Figure 2. Rating scale domains

Task Fulfilment
Accuracy and range of grammar
Accuracy and range of vocabulary
Organisation

Method

The key research question for this study is whether examiner severity will be comparable within each test and across tests at the six CEFR levels; i.e., whether examiners will apply the marking descriptors accurately and be consistently lenient / severe on tests within a level and across levels. Two indicators of examiner severity and consistency were examined to address the research question.

The first indicator generated from the Rasch analysis is the person fit statistic. This statistic is not a direct indicator but a pre-requisite of examiner consistency. Examiner performance has to satisfy Rasch measurement requirements (i.e., the fit to the Rasch model) before any meaningful discussions on severity estimates may be made. The computer program FACETS (Linacre, 2020) provides a number of statistics which give an indication as to how well the data fits the model. One of these is the mean square statistic. For person fit statistics (examiners, in this case), acceptable practical limits of fit have been proposed as 0.5 for the lower limit and 1.5 for the upper limit (Lunz & Stahl, 1990).

The second indicator relates to examiner invariance across tests. While MFRA provides a framework for obtaining fair measurements of examinee ability that can be statistically invariant over examiners, tasks, and other aspects of performance assessment procedures, this only applies across one test. In the current study,

examiner invariance across the six tests is examined via the Spearman's rho, which reports rank order correlations between tests. A high correlation indicates consistency of rank order of examiner severity estimates.

Results and Discussion

Examiner Fit to the Rasch Model

As the cornerstone of good rating is fit to the Rasch model, results are first presented below for the examiners on each of the six tests. Tables 2a and 2b present the results for the 27 examiners who participated in the standardisation exercise. As mentioned, 24 examiners marked all six tests, with the whole cohort of 27 examiners marking tests A1 and A2. In the tables below, Infit is reported. Infit shows the 'big picture' in that it scrutinises the internal structure of a facet (examiners, in this case). Generally speaking, high infit (above 1.5) values would suggest an examiner's ratings were rather 'scattered', providing a confused picture about the placement of the examiner's ratings. Very small (below 0.5) infit values indicate only very small variation in the data, thereby providing little information to articulate clear and meaningful judgments about the examiner – and their ratings.

Infit figures above 1.5 are highlighted in yellow, while Infit figures below 0.5 are highlighted in green. In the data and discussion below, all examiner names have been anonymised.

Table 2a: Examiner measures for tests A1 and A2 (N=27)

Examiners	Nu	A1-Measure	A1-S.E.	A1-Infit	A2-Measure	A2-S.E.	A2-Infit
Andy	1	-0.1	0.46	0.64	0.69	0.44	1.14
Brian	2	0.5	0.44	0.85	1.47	0.45	0.87
Cathy	3	-0.78	0.5	0.68	-0.92	0.47	1
Dot	4	0.31	0.44	0.52	-0.09	0.45	1.29
Ellen	5	0.69	0.43	0.65	0.3	0.44	0.71
Fred	6	0.31	0.44	0.81	-0.09	0.45	0.72
Gary	7	0.11	0.45	1.7	-1.15	0.48	1.06
Terri	8	-1.61	0.56	0.61	-0.92	0.47	0.86
Iris	9	0.11	0.45	0.93	-0.09	0.45	0.88
Jack	10	1.92	0.41	1.01	0.69	0.44	0.76
Katie	11	0.88	0.43	1.43	0.11	0.45	1
Lenny	12	0.31	0.44	1.11	-0.92	0.47	1.05
Martha	13	-1.61	0.56	0.61	-0.92	0.47	0.86
Nonie	14	-0.54	0.48	0.53	0.11	0.45	0.61
Oliver	15	0.31	0.44	0.94	-1.62	0.5	0.84
Perry	16	0.5	0.44	0.99	1.08	0.44	1.03
Queenie	17	-1.04	0.52	2.46	-0.09	0.45	0.83
Robert	18	0.31	0.44	1.17	0.69	0.44	1.7

Table 2a: Examiner measures for tests A1 and A2 (N=27) (continued)

Examiners	Nu	A1-Measure	A1-S.E.	A1-Infit	A2-Measure	A2-S.E.	A2-Infit
Susan	19	-0.54	0.48	1.21	-0.5	0.46	1
Terri	20	-1.31	0.54	0.76	-0.29	0.45	0.83
Ursula	21	-0.1	0.46	0.77	-0.5	0.46	0.78
Vanesa	22	0.11	0.45	1.53	1.08	0.44	1.39
Windy	23	-0.1	0.46	1.27	0.5	0.44	1.54
Xerxes	24	0.31	0.44	1.01	0.69	0.44	0.79
Yana	25	1.24	0.42	0.68	1.28	0.44	0.61
Zoe	26	-0.1	0.46	1.2	-0.71	0.46	0.98
Albert	27	-0.1	0.46	0.9	0.11	0.45	0.96

Table 2b: Examiner measures for tests B1, B2, C1 and C2 (N=24)

Examiners	Nu	B1-Measure	B1-S.E.	B1-Infit	B2-Measure	B2-S.E.	B2-Infit	C1-Measure	C1-S.E.	C1-Infit	C2-Measure	C2-S.E.	C2-Infit
Andy	1	0.88	0.41	0.67	1.82	0.46	0.81	0.63	0.36	1.42	0.75	0.43	0.88
Brian	2	0.54	0.41	0.96	1.19	0.46	0.68	0.24	0.36	0.64	1.29	0.43	0.84
Cathy	3												
Dot	4	0.54	0.41	1.16	-0.23	0.45	1.11	0.63	0.36	1.3	0	0.44	1.04
Ellen	5	-1.41	0.49	0.81	-0.23	0.45	0.88	0.63	0.36	0.71	0.75	0.43	0.67
Fred	6	-0.96	0.47	1.11	-0.64	0.45	0.81	0.5	0.36	0.91	1.29	0.43	1.27
Gary	7	-1.9	0.51	1.59	-0.84	0.46	1.04	-0.98	0.38	0.97	-0.38	0.44	1.47
Terri	8												
Iris	9	-1.65	0.5	1.77	-0.43	0.45	0.87	0.11	0.36	1.11	-0.57	0.44	0.79
Jack	10	0.71	0.41	1.07	1.61	0.46	0.94	0.5	0.36	0.42	0.93	0.43	0.58
Katie	11	-0.54	0.45	0.67	-0.03	0.45	0.88	0.24	0.36	0.57	0.56	0.43	0.46
Lenny	12	-0.16	0.43	1.23	-0.03	0.45	1.04	-0.98	0.38	1.27	-1.77	0.46	0.44
Martha	13												
Nonie	14												
Oliver	15	-1.18	0.48	1.44	-1.05	0.46	0.63	-0.7	0.37	0.78	-0.96	0.45	0.56
Perry	16	0.2	0.42	0.97	-0.43	0.45	1.27	-0.98	0.38	1.3	-1.15	0.45	0.72
Queenie	17	-0.75	0.46	0.7	0.18	0.45	0.66	-0.43	0.37	1.05	-1.15	0.45	0.95
Robert	18	0.54	0.41	1.18	0.58	0.45	1.15	0.11	0.36	1.59	0.93	0.43	1.22
Susan	19	0.2	0.42	0.83	-0.84	0.46	0.93	-0.7	0.37	1.28	-1.36	0.45	1.4
Terri	20	0.71	0.41	0.6	0.38	0.45	0.72	1.53	0.36	1.02	1.47	0.42	0.36
Ursula	21	0.71	0.41	0.6	-0.43	0.45	2.09	-0.43	0.37	1.02	0.38	0.43	1.14
Vanesa	22	1.04	0.41	0.84	0.99	0.45	1.02	-0.29	0.37	1.12	0.38	0.43	1.67
Windy	23	0.02	0.43	1.14	-0.03	0.45	0.77	-0.29	0.37	1	0.56	0.43	2.03
Xerxes	24	1.85	0.4	0.66	-0.84	0.46	0.33	0.24	0.36	0.59	0.38	0.43	0.54
Yana	25	1.04	0.41	0.77	-0.23	0.45	1.03	0.5	0.36	0.89	-0.57	0.44	0.74
Zoe	26	-1.18	0.48	1.48	-0.64	0.45	1.63	-0.43	0.37	0.85	-1.77	0.46	0.71
Albert	27	0.71	0.41	0.76	0.18	0.45	1.09	0.37	0.36	0.66	0	0.44	1.48

As can be seen from the data in the above table, examiner fit to the model was generally good; there were only one or two examiners who showed underfit (i.e., with a mean square of over 1.5) on each test.

Examiner Consistency across Tests

Having established that examiners broadly fit the model acceptably, the next step involves examining examiner consistency across tests. Table 3 presents the results of rank order correlations (via Spearman's rho) conducted against examiner person measures across the 6 tests. Correlations significant at the 0.01 level are highlighted in yellow, and those significant at the 0.05 level in green.

Table 3: Examiner measure rank order correlations across the six tests

		A1-Measure	A2-Measure	B1-Measure	B2-Measure	C1-Measure	C2-Measure
A1-Measure	Correlation Coefficient	–	.531**	.014	.035	.183	.187
	Sig. (2-tailed)	.	.004	.951	.876	.404	.393
	N	27	27	23	23	23	23
A2-Measure	Correlation Coefficient	.531**	–	.582**	.551**	.395	.425*
	Sig. (2-tailed)	.004	.	.004	.006	.062	.043
	N	27	27	23	23	23	23
B1-Measure	Correlation Coefficient	.014	.582**	–	.459*	.388	.302
	Sig. (2-tailed)	.951	.004	.	.028	.067	.162
	N	23	23	23	23	23	23
B2-Measure	Correlation Coefficient	.035	.551**	.459*	–	.447*	.514*
	Sig. (2-tailed)	.876	.006	.028	.	.033	.012
	N	23	23	23	23	23	23
C1-Measure	Correlation Coefficient	.183	.395	.388	.447*	–	.696**
	Sig. (2-tailed)	.404	.062	.067	.033	.	.000
	N	23	23	23	23	23	23
C2-Measure	Correlation Coefficient	.187	.425*	.302	.514*	.696**	–
	Sig.	.393	.043	.162	.012	.000	.
	N	23	23	23	23	23	23

**. Correlation significant at the 0.01 level; *. Correlation significant at the 0.05 level

As may be seen from Table 3, in general, tests (that is, via examiner person measures) correlate highly with their 'partner': hence the A1 and A2 tests correlate highly (at the $p<.01$ level), as do the C1 and C2 tests; and the B1 and B2 tests correlate quite highly (at the $p<.05$ level). While the A2 test appears to correlate with almost all tests, all tests correlate quite highly with at least two or more different tests. The implication of these correlations is that the rank order of the examiners is broadly consistent across tests: if an examiner is going to be strict on one test, it is quite likely that they will be strict on other tests.

Conclusion

This study has examined the issue of examiner severity and invariance across LanguageCert's six CEFR-linked IESOL Writing tests. The research question was whether examiner severity would be comparable within each test and across the six tests; i.e., examiners would be consistently severe on each test. If examiners are seen to be erratic in their severity at some levels but not at others, this may impact on fairness in terms of grades awarded to candidates.

An examination of 27 examiners standardised to mark LanguageCert's six CEFR-linked IESOL Writing tests, illustrated that examiner fit to the Rasch model was generally good – a key background consideration.

From correlations run among the examiner person measures across all six tests, a rank order emerged indicating that examiners were broadly consistent across tests. Examiner person measures generally correlated highly with their 'partner' test: A1 with A2, C1 with C2, and B1 with B2 tests. While the A2 test correlated with almost all tests, all tests correlated quite highly with at least two or more different tests.

A major implication which arises regarding consistency is the following: if an examiner is going to be strict at one level, they will quite likely be strict at other levels – and strictness can be compensated for. Given that LanguageCert examiners undergo careful training and standardisation, what the current study illustrates is that LanguageCert examiners may be seen to mark consistently and accurately across a range of ability levels.

References

- Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). New York: Routledge.
- Coniam, D., & Falvey, P. (2018). *High-stakes testing: The impact of the LPATE on English language teachers in Hong Kong*. Springer Nature: Singapore.
- Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33, 117-135.
- Elder, C., Barkhuizen, G., Knoch, U., & Radow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37-64.
- Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal, & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Erlbaum.
- Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, 32(4), 279-286.
- Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, 36(4), 481-504.
- Kogan, J. R., Conforti, L. N., Bernabeo, E., lobst, W., & Holmboe, E. (2015). How faculty members experience workplace-based assessment rater training: A qualitative study. *Medical Education*, 49(7), 692-708.
- Linacre, J. M. (2020) *Facets computer program for many-facet Rasch measurement*. Beaverton, Oregon: Winsteps.com.
- Lumley, T., & T. McNamara. (1995). Examiner characteristics and examiner bias: Implications for training. *Language Testing*, 12, 1, 54-71.
- Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession*, 13, 425-444.
- McNamara, T. (1996). *Measuring second language performance*. New York: Longman.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement test*. Copenhagen, Denmark. Danish Institute for Educational Research. Expanded ed. (1980). Chicago, IL: The University of Chicago Press.
- Pam, N. 2013. Frame-of-reference training. *sychologyDictionary.org*, May 11, 2013, <https://psychologydictionary.org/frame-of-reference-training/> (accessed July 28, 2021).
- Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25(4), 553-581.
- Webb, L., Raymond, M. & Houston, W. (1990). *Examiner stringency and consistency in performance assessment*. Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
- Weigle, S. (1998). Using FACETS to model examiner training effects. *Language Testing*, 15, 2, 263-287.
- Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, 16(4), 33-45.

Chapter 5: Potential Bias in LanguageCert IESOL Items: A Differential Item Functioning Analysis

David Coniam and Tony Lee

Abstract

Differential Item Functioning (DIF) analysis is a statistical procedure undertaken to explore whether any subgroup of test takers sitting a test or exam is being unfairly disadvantaged or indeed advantaged. Investigating DIF is key to understanding and dealing with test bias, a necessary though complex requirement of leading test providers. To date, PeopleCert has not been in a position to address this issue in any depth. This has the potential to diminish the organization's standing in the international assessment community. However, preliminary work on test bias and DIF has now begun and the methodology required to carry it out has been identified and is described in this paper.

The current study reports on a DIF analysis of the six IESOL exams aligned to the CEFR and delivered by LanguageCert between 2018-2020. For each CEFR level, four variables, namely mother tongue, age, gender and test centre were explored. DIF analysis was conducted using the computer program Winsteps, with DIF strength reported in line with Zwick (1999).

Some moderate-to-large DIF was reported for *mother tongue* and *decade of birth* (a recording of *age*). This, however, may well be due to the fact that these two categories are very diverse with only very few entries.

For *gender* – typically a key variable in the exploration of DIF – a very low incidence of 3% DIF was reported. For *centre* (i.e., a comparison of OLP vs non-OLP delivery), zero and moderate-to-large DIF was observed. An examination of reading or listening items indicated that there was no predominance of DIF in either skill.

These are encouraging preliminary findings and confirm that the six *LanguageCert* tests analysed here are showing relatively low levels of bias. Mechanisms will need to be put in place to monitor DIF on all *LanguageCert*

products going forward. This will include gathering more information about candidates in a systematic and comprehensive manner so that important potential sources of bias can be investigated.

Background

Differential Item Functioning (DIF) analysis is a statistical procedure undertaken to explore whether any subgroup of test takers sitting a test is being unfairly disadvantaged. The exploration of potential sources of bias among subgroup types typically investigate variables such as gender, cultural affiliation, age etc. Indeed, in many DIF studies, and often for political reasons, key variables studied have tended to be gender and ethnicity (Ferne & Rupp, 2007).

Evidence of DIF may be apparent in a test item (or indeed on an entire test) when the responses of two groups of test takers who should have equal “latent trait ability” show different probabilities in terms of correctly answering a test item (Swaminathan & Rogers, 1990). While DIF analysis has been used for a considerable time by the general educational community, it is only in the past three decades that the use of DIF has become more prevalent in the language assessment field (Aryadoust et al., 2011; Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000). Ferne & Rupp (2007) provide a cogent synthesis of 15 years of research on DIF in language testing.

While DIF was initially conducted using classical test statistics, more recently Rasch-based methods (see, e.g., Roznowski & Reith, 1999) have come to be the preferred statistical mode of analysis. A useful extension of DIF may be seen in ‘bundling’, that is grouping items into sets that share the same latent trait (e.g., Gierl et al., 2001). ‘Bundling’ in Rasch analysis (Linacre, 2012) is referred to as Differential Group Functioning (DGF), and in the case of test development purposes, DGF may be seen to be procedurally more informative than DIF (see Linacre, 2012). For ease of reference, however, given the general acceptance of the term “DIF”, it is “DIF” that will be referred to in the current study to maintain consistency.

Depending on the type and level of test, it is not unlikely that some DIF will be found among certain background variables. While some studies investigating DIF have reported zero DIF (Chen & Henning, 1985), other studies have reported quite high incidences: Abbott (2004), for example, reports DIF of 62%. And even in studies where DIF has been found, a deeper exploration of DIF does not necessarily indicate any actual difference in performance between different DIF groups (Prieto & Nieto, 2014).

Data in the Current Study

DIF reported on in the current study was conducted on the objectively-marked Listening and Reading components of tests delivered by LanguageCert between 2018-2020. LanguageCert produce and administer a suite of exams – the International ESOL suite – which are aligned to the six CEFR levels: Preliminary (A1), Access (A2), Achiever (B1), Communicator (B2), Expert (C1) and Mastery (C2). The examination specifications reflect the six levels of the CEFR with regard to language attributes such as grammar, functions, vocabulary and discourse, function and how these relate to communication. Each exam has 52 items, of which 26 focus on reading and 26 on listening.

Since DIF optimally requires large sample sizes (Linacre, 2012), an exam with large sample sizes was identified for each CEFR level. Four background demographic variables have been used in the current study to explore DIF. Three of these – *gender, age, mother tongue* – are data supplied (optionally) by candidates upon registering for an exam. Given that many LanguageCert exams are conducted remotely through online proctoring (OLP), *test centre* is also taken as a variable. To make the analysis tractable, some recoding has been necessary, as laid out below.

- Gender: coded male / female.
- Age: recoded into decade of birth
- Mother tongue: only analysed where the sample size is greater than 10 incidences.
- Test centre:
 - (1) analysed as is
 - (2) recoded into either test taken face to face at a centre / test conducted via OLP

In the discussion and analysis below, DIF results are only presented for variables for which data exist; blank categories – i.e., where candidates did not report – have not been included. Further, only exam levels where DIF was observed are presented in the analysis. If a particular level does not appear in a table, that is because there was no DIF recorded for that level.

DIF investigations can operate at several levels. The aim of the present paper is to investigate initial DIF, i.e., the DIF of critical background variables. Such investigations provide evidence as to the overall quality and degree of bias in LanguageCert tests.

Results

In this section, Differential Group Functioning (DGF) using the computer program Winsteps (Linacre, 2010) is applied. Zwick (1999) provides an interpretation of significance (see also Linacre, 2010), where DIF strengths are graded into three categories, as in Table 1.

Table 1: DIF Strengths (after Zwick, 1999)

DIF Category	Strength	Logit size	Significance value
A	Negligible		
B	Slight to moderate	> 0.43 logits	p < 0.05
C	Moderate to large	> 0.64 logits	p < 0.05

In the discussion below, the focus will therefore be on Category C, moderate-to-large DIF as this is the DIF category that would be the most worrying if large DIF trends emerged. For brevity's sake, only overall summaries are presented for each test level.

Gender

Table 2 presents the results for gender. The reader's attention is drawn, as mentioned, to Category C – moderate-to-large DIF.

Table 2: DIF by Gender

Level		A	B	C	Total
A1	No.	20	3	1	24
	% within row	83.33 %	12.50 %	4.17 %	100.00 %
A2	No.	22	2	0	24
	% within row	91.67 %	8.33 %	0.00 %	100.00 %
B1	No.	11	3	1	15
	% within row	73.33 %	20.00 %	6.67 %	100.00 %
B2	No.	23	0	1	24
	% within row	95.83 %	0.00 %	4.17 %	100.00 %
C1	No.	22	1	1	24
	% within row	91.67 %	4.17 %	4.17 %	100.00 %
C2	No.	5	1	0	6
	% within row	83.33 %	16.67 %	0.00 %	100.00 %
Total	No.	103	10	4	117
	% within row	88.03 %	8.55 %	3.42 %	100.00 %

As can be seen, there are very few instances of DIF in Category C – 3.4% of the total.

Mother Tongue

LanguageCert has a list of over 100 mother tongues. The majority of these categories in the current dataset were either empty or had only one or two entries. Analysis has therefore only been conducted, as mentioned, where the sample size was greater than 10.

Table 3 below reports on the incidence of DIF totals for the 6 exam levels.

Table 3: DIF by Mother tongue

Level		A	B	C	Total
A1	No.	36	3	9	48
	% within row	75.00 %	6.25 %	18.75 %	100.00 %
A2	No.	67	12	9	88
	% within row	76.14 %	13.64 %	10.23 %	100.00 %
B1	No.	20	8	12	40
	% within row	50.00 %	20.00 %	30.00 %	100.00 %
B2	No.	72	13	19	104
	% within row	69.23 %	12.50 %	18.27 %	100.00 %
C1	No.	61	8	3	72
	% within row	84.72 %	11.11 %	4.17 %	100.00 %
C2	No.	8	0	0	8
	% within row	100.00 %	0.00 %	0.00 %	100.00 %
Total	No.	264	44	52	360
	% within row	73.33 %	12.22 %	14.44 %	100.00 %

There is some incidence of DIF, with 14.4% of DIF reported for the C grade. In part this may be attributed to the wide scattering of different first languages and low sample sizes.

Decade of Birth

Year of birth may be seen to be an even more multifaceted variable than mother tongue. To this end, year of birth has recoded into decade of birth: 1960, 1970, 1980 etc. Table 4 presents the results.

Table 4: DIF by decade of birth

Level		A	B	C	Total
A1	No.	28	3	9	40
	% within row	70.00 %	7.50 %	22.50 %	100.00 %
A2	No.	42	7	7	56
	% within row	75.00 %	12.50 %	12.50 %	100.00 %
B1	No.	150	25	35	210
	% within row	71.43 %	11.90 %	16.67 %	100.00 %
B2	No.	43	3	2	48
	% within row	89.58 %	6.25 %	4.17 %	100.00 %
C1	No.	37	3	0	40
	% within row	92.50 %	7.50 %	0.00 %	100.00 %
C2	No.	10	0	0	10
	% within row	100.00 %	0.00 %	0.00 %	100.00 %
Total	No.	310	41	53	404
	% within row	76.73 %	10.15 %	13.12 %	100.00 %

The incidence of DIF is 13.1%. From an examination of the data, there is no clear pattern of age or level.

Centre: Face to face vs. OLP

Over 200 centres around the world conduct tests in face-to-face mode. However, many of these conduct a very few tests.

Table 5: DIF by Centre

Level		A	B	C	Total
A1	No.	62	19	31	112
	% within row	55.36%	16.96%	27.68%	100.00%
A2	No.	99	18	35	152
	% within row	65.13%	11.84%	23.03%	100.00%
B1	No.	128	13	44	185
	% within row	69.19%	7.03%	23.78%	100.00%
B2	No.	16	0	0	16
	% within row	100.00%	0.00%	0.00%	100.00%
C1	No.	16	0	0	16
	% within row	100.00%	0.00%	0.00%	100.00%
C2	No.	4	0	0	4
	% within row	100.00%	0.00%	0.00%	100.00%
Total	No.	325	50	110	485
	% within row	67.01%	10.31%	22.68%	100.00%

The incidence of C grade DIF is 22.7%. In part, this may again be attributed to the large number of centres, with some administering tests to a very small number of candidates.

It is difficult to comment objectively on DIF across centres since there are over 200 LanguageCert centres around the world. All centres are face-to-face institutions, with the exception of the LanguageCert centre which operates out of Athens, and which conducts exams remotely via OLP with candidates who are potentially of B2-C2 level. To shed some more light on the centre issue – and to the remote delivery of English language tests – a further focused analysis is now presented.

Centre: Face to face vs. OLP

LanguageCert is becoming a key player in delivering tests remotely through online proctoring (OLP). OLP is used to administer approximately 50% of LanguageCert's English language tests from the Athens centre. Against this backdrop and the multiplicity of centres, many of which have a very few candidates, *centre* has been recoded into OLP / face to face. Table 6 presents the DIF results for this analysis.

Table 6: Centre – OLP vs. face to face exam delivery

Mode	Level		A	B	C	Totals
OLP	B2	No.	8	0	0	8
		%	100%	0%	0%	100%
OLP	C1	No.	8	0	0	8
		%	100%	0%	0%	100%
OLP	C2	No.	2	0	0	2
		%	100%	0%	0%	100%
OLP	Total	No.	18	0	0	18
		%	100%	0%	0%	100%
Face to face	B2	No.	8	0	0	8
		%	100%	0%	0%	100%
Face to face	C1	No.	8	0	0	8
		%	100%	0%	0%	100%
Face to face	C2	No.	2	0	0	2
		%	100%	0%	0%	100%
Face to face	Total	No.	18	0	0	18
		%	100%	0%	0%	100%

From the aggregated analysis, no instances of C (nor of the less severe B) grade DIF were reported, either face to face at a physical centre or via OLP. Given the importance that LanguageCert attaches to its online proctoring operation, it is crucial that no bias be attached to this mode of delivery. The results in Table 6 above would appear to support this contention. ANOVA was used to investigate further and no significance bias was observed.

Reading or Listening

The IESOL examinations each comprise an equal number (26) of reading and listening items. From an investigation of DIF across both skills, it was concluded that there was no significant DIF in either reading or listening.

Conclusion

This study has investigated the incidence of DIF, or DGF (i.e., bundled DIF) across four of the background and location variables related to LanguageCert's suite of IESOL exams. The key focus of analysis has been on Zwick's moderate-to-large defining of DIF of 0.64 of a logit or greater. The overall preliminary finding based on the six exams analysed is that DIF is predominantly in Category A (negligible). A summary of the analysis of the four key variables explored is presented below.

- For *mother tongue*, a diverse category comprising over 100 languages, 15% moderate-to-large DIF was reported. Much of this may be attributable to the fact that many categories have only a very few entries.

- For *decade of birth* (recoded from year of birth), 13% moderate-to-large DIF was reported. Although decade of birth is a rather crude measure, it was used due to the small sample sizes. When larger sample sizes are available for analysis, it will be interesting to investigate in more depth.
- For *gender* – typically a key variable in the exploration of DIF – a very low incidence of 3% DIF was reported.
- Given LanguageCert's strong presence in remote delivery of tests, *centre* – comparing OLP vs non-OLP centres) – was also considered a variable of importance. Zero and moderate-to-large DIF was recorded for this variable. This is an encouraging finding on an important issue.

In closing, it is worth comparing the incidence of DIF revealed in the current study with Ferne & Rupp's (2007) meta-analysis of DIF studies. The studies reported by Ferne & Rupp were essentially all tightly focused, that is, usually a single test with the focus on a single variable examining two clearly contrastive groups. A wide range of DIF across different studies was reported, as mentioned above.

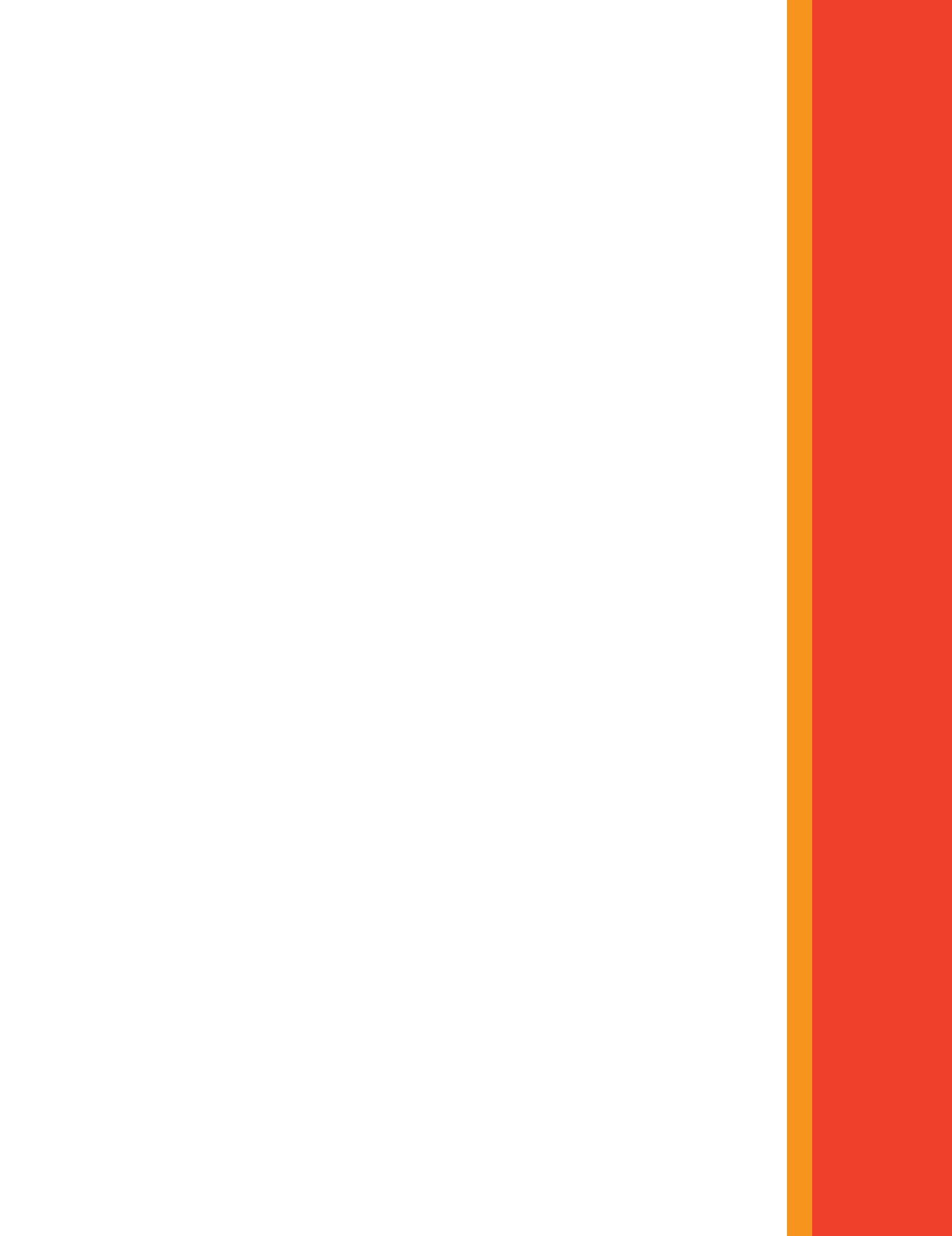
The current study has involved the investigation of DIF over six ability levels (as per the CEFR), in the context of four background variables, two of which have 100 or more sub-categories. In this light, it may not be surprising that a degree of DIF was observed. However, while a degree of DIF was observed, it was encouraging to note that DIF related to the categories of gender and OLP versus centre delivery, was negligible.

In light of Prieto & Nieto's (2014) claims that DIF does not necessarily impact on overall candidate performance, our preliminary conclusion is that LanguageCert exams in this study reflect the fact that relatively, they are carefully and professionally developed. Equally reassuring is the finding that there was minimal observable DIF in reading or listening components. Results generated from the six LanguageCert IESOL exams in this study may be seen as fair in relation to gender and delivery mode. Findings related to mother tongue and age are less compelling, but this is largely due to sample size.

References

- Abbott, M. (2004). The identification and interpretation of group differences on the Canadian Language Benchmarks Assessment Reading Items. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Aryadoust, V., Goh, C.C.M., & Kim, L.O. (2011). An investigation of differential item functioning in the MELAB Listening Test. *Language Assessment Quarterly*, 8(4), 361-385.
- Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2, 155–163.
- Dunson, D. B. (2001). Commentary: practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, 153(12), 1222–1226.
- Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4, 113-148.
- Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement Issues and Practice*, 20(2), 26-36.

- Linacre, J. M. (2010). WINSTEPS, Version 3.69. Chicago, IL: Winsteps.com.
- Linacre, J. M. (2012). A user's guide to WINSTEPS. Chicago, IL: Winsteps.com.
- Prieto, G., & Nieto, E. (2014). Influence of DIF on differences in performance of Italian and Asian individuals on a reading comprehension test of Spanish as a foreign language. *Journal of Applied Measurement*, 15(2), 176-188.
- Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, 59, 248-269.
- Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, 9, 12-29.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, 17, 323-340.
- Zwick, R., Thayer, D. T., Lewis, C. 1999. An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36(1), 1-28.



Chapter 6: Task Equivalence in LanguageCert IESOL Writing Tests

David Coniam and Maria Babatsi

Abstract

Each *LanguageCert IESOL* level has at its disposal a large battery of test forms, with every test form using different writing tasks. In order to ensure fairness to candidates, it is important that the input provided by these writing tasks be as equivalent as possible. Demonstrating such equivalence of input ensures that one potential source of measurement error is managed effectively with tasks posing a comparable challenge to candidates. This in turn means that the score achieved by a candidate in the assessment of writing is a function of candidate ability as opposed to task difficulty. This paper aims to explore the extent to which, the difficulty of writing tasks varies across *LanguageCert IESOL* examinations at levels B2 and C1.

Writing tests at *LanguageCert IESOL* levels B2 and C1 comprise two tasks, the first quite short, the second requiring rather longer output. Test constructors aim to make the input as comparable as possible. This objective is investigated using Multi-faceted Rasch Analysis (MFRA), which confirmed a high degree of comparability across test tasks in terms of difficulty. In addition, the difficulty range at both levels (B2 and C1) was found to be under a logit for the B2 tests and 1.5 logits for the C1 tests. Such comparable difficulty of input is the starting point for candidate output and helps examiners to rate more effectively without the distraction of tasks displaying a significantly different challenge to candidates.

Introduction

Research into the performance skills (writing and speaking) has highlighted a range of factors that potentially impact on the assignment of scores to candidates. While rater severity is generally identified as the significant factor, three other factors are relevant. These are: task comparability; the equivalence of rating scales; and test taker demographics. These three factors are potentially relevant because they may contribute to measurement variance, and in turn impact on a candidate's score (Carrell, 1995).

Research by Barkaoui & Knouzi (2012) describes how task variability in terms of factors such as wording, content, audience, purpose, complexity, genre may impact in different ways on test scores. They comment on how writing tasks – if not well specified – may produce very different outputs, with a concomitant effect on scores awarded.

Such task effects have been observed by numerous researchers. Weigle (1999) reported novice raters assigning lower scores to particular task types. Cumming et al. (2002) found that task type affected rater behaviour and the writing features attended to by raters. Suh & Bae (2016) illustrated how prompts in a creative writing task were not equal in terms of difficulty.

The genre of the task has also been researched. Coniam (1991), and Hamp-Lyons and Mathias (1994), found that tasks judged to be difficult (argumentative impersonal topics) resulted in higher mean essay scores than tasks judged to be easy (expository personal topics). Koda's (1993) investigation of task difficulty with American students studying Japanese revealed that descriptive tasks placed fewer linguistic and cognitive demands on students than narrative tasks.

In English language assessment, the statistical procedure most widely used in the analysis of performance test output is Multi-faceted Rasch Analysis (MFRA). Bachman et al. (1995), for example, used MFRA to investigate the degree of variability in spoken language tasks. Bonk & Ockey (2003) used MFRA analysis to explore the effect of prompt in peer group discussion tasks with Japanese English-major university students. The reader is referred to the outline of the Rasch measurement model and MFRA provided in the Glossary of statistical terms at the end of the volume.

The IESOL Writing Test

The IESOL Writing tests comprise two tasks, as laid out in Figure 1.

Figure 1: IESOL Writing Test tasks and scales

Level	Part 1: Candidates produce	Word length	Part 2: Candidates produce	Word length
A1	four sentences on a specified topic	30	a simple text for a specified reader	20-30
A2	an informal response to an informal text	30-50	a neutral response to a specified public reader	30-50
B1	a neutral or formal text for a public audience	70-100	a letter using informal language	100-120
B2	a neutral or formal text for a public audience	100-150	a text using informal language	150-200
C1	a neutral or formal text for a public audience	150-200	a text using informal language	250-300
C2	a neutral or formal text for a public audience	200-250	a text using informal language	250-300

All tasks conform to CEFR ‘can do’ statements for writing and are assessed on a four-point scale on four subscales as illustrated in Figure 2.

Figure 2. Rating subscales

Full name	Short form
Task Fulfilment	TF
Accuracy and range of grammar	ARG
Accuracy and range of vocabulary	ARV
Organisation	IO

The rating scale for each subscale extends from 0 to 3, where, for a given CEFR level, level 2 of the subscale is interpreted as the ‘canonical’ level. Consider CEFR B2. A candidate being awarded a level 2 would be considered as being exactly at level B2. A candidate at level 1 would therefore be seen as not quite reaching the B2 threshold, while a candidate scoring a 3 would be seen as a high B2. For examiners to make such judgements, it is therefore critical that tasks offer sufficient direction and guidance but are neither too demanding nor too easy.

Method

The key research question for this study is whether task severity is comparable across the range of test forms at the different CEFR levels. As mentioned above, since the *LanguageCert* examinations with the largest candidate cohorts were B2 and C1, these two examinations are investigated in the current study. To avoid overwhelming the reader, detail is only provided for the task in Part 2 of the examination, since this requires a slightly longer response from candidates.

Table 1 details the makeup of the two tests.

Table 1. Makeup of the two tests

	B2	C1
No. of tests	16	17
No. of candidates	6,656	4,863
No. of tasks analysed	16	17

Tasks were analysed from 16 B2 examinations and 17 C1 examinations with over 6,000 candidates taking the B2 exam and nearly 5,000 the C1 exam.

Multi-faceted Rasch Analysis (MFRA) is the statistical procedure used – via the computer program FACETS (Linacre, 2020), which provides a number of statistics which give an indication as to how well the data fits the model. In MFRA, the key indicators generally scrutinised are the fit statistics, with the principle fit statistic being the mean square statistic. For fit statistics, acceptable practical limits of fit have been proposed as 0.5 for the lower limit and 1.5 for the upper limit (Lunz & Stahl, 1990).

While this statistic may not be a direct indicator of consistency, it is a necessary pre-requisite. Performance has to satisfy Rasch measurement requirements (i.e., the fit to the Rasch model) before any meaningful discussions on severity estimates may be made.

The standardised Z-score (ZSTD) is an extension to the interpretation of the Infit mean square values. It is a t-test exploring how well the data fit the model; figures above 2.0 indicate distortion in the measurement system (Linacre, 2003.).

The point measure correlation (PTME) in the Rasch model is comparable to the conventional point biserial correlation. Negative PTME values indicate a lack of model fit.

Results and Discussion

To give an overview of the measurement, the vertical ruler (the 'facet map') produced in the output is first presented below. This is a visual representation of where facets (candidates, tasks etc.) are located on the scale.

Following this, a table containing Infit mean square data is provided.

Figure 3. B2 tests facet map

Measr	+Candidates	-Tasks	-Subscales	
10 +	+		+	+
9 +	+		+	+
8 +	+		+	+
7 + **.	+		+	+
*.				
6 + *****.	+		+	+
***.				
5 + ****.	+		+	+
****.				
4 + ***.	+		+	+
***.				
3 + ****.	+		+	+
***.				
2 + ****.	+		+	+
***.				
1 + ***.	+		+	+
*.	061-T2 251-T2 511-T2 571-T2 921-T2			
* 0 * *****.	* 191-T2 391-T2 471-T2 521-T2 631-T2 681-T2 691-T2 821-T2	* ARV IO *		
**.	181-T2 311-T2			
-1 + **.	+ 811-T2		+ IF	+
*.				
-2 + *.	+		+	+
.				
-3 + .	+		+	+
.				
-4 + .	+		+	+
.				
-5 +	+		+	+
.				
-6 +	+		+	
.				
-7 +	+		+	+
.				
-8 +	+		+	+
.				
-9 +	+		+	+
Measr	* = 88	-Tasks	-Subscales	

As may be seen from Figure 3, candidates demonstrate a nine-logit spread across the ability spectrum – unsurprising with a cohort of over 6,600 candidates. By contrast, both the tasks and rating scales are within much narrower ranges. This is a reassuring finding given that it suggests continuity of input.

Table 2 below presents the key statistics for the B2 tests. Potentially problematic statistics are presented in bold font.

Table 2. Key MRFA statistics, Test B2 (N=6,656)

Total count	Measure	Model S.E.	Infit		PTME	Tasks
			MnSq	ZStd		
28	0.41	0.47	1.10	0.4	0.41	921-T2
1272	0.40	0.07	0.98	-0.4	0.39	571-T2
1260	0.39	0.07	0.97	-0.7	0.39	251-T2
1148	0.35	0.07	1.07	1.6	0.39	061-T2
720	0.28	0.09	1.08	1.6	0.40	511-T2
4476	0.19	0.04	0.97	-1.4	0.48	191-T2
4368	0.16	0.04	1.13	5.4	0.47	821-T2
1524	0.05	0.07	1.05	1.3	0.47	631-T2
652	-0.01	0.10	0.92	-1.4	0.45	391-T2
4108	-0.04	0.04	0.94	-2.5	0.48	521-T2
696	-0.05	0.10	0.98	-0.3	0.39	681-T2
2312	-0.07	0.05	0.97	-1.0	0.49	471-T2
944	-0.11	0.08	0.76	-5.8	0.40	691-T2
4016	-0.25	0.04	0.88	-5.4	0.48	311-T2
88	-0.46	0.27	2.01	5.0	0.32	181-T2
28	-1.23	1.17	0.51	-0.7	0.44	811-T2
1727.5	0.00	0.17	1.02	-0.3	Mean	
1560.9	0.40	0.28	0.29	2.9	S.D.	

Model, Sample: RMSE .33 Adj (True) S.D. .25 Separation .77 Strata 1.36 Reliability .37

We can see that the infit statistics are good. Task 181-T2 falls outside the 0.5 – 1.5 accepted limits of fit, while task 821-T2 has a high standardised t-test score, suggesting some possible distortion in the data. All point measure correlations are, however, high indicating good model fit.

Two tasks (811-T2 and 921-T2 – in bold font) have high standard errors. These three tasks have, however, only been taken by a very small number of candidates. Since standard error is directly linked to sample size, the fact that these tasks have been administered to very small numbers of candidates in large part accounts for the large error size.

Despite being taken by over 6,600 candidates, the 16 tests exhibit a 1.5 logit range: extending from 0.41 to -1.23 logits. The easiest task (871-T2) was, however, only administered to a very small number of candidates.

If this task is disregarded, along with the two tasks mentioned above with high standard errors, we see that, essentially, all tasks fall within two thirds of a logit range (+0.40 to -0.25) and are statistically robust.

To complement the picture, Table 3 presents the key statistics for the C1 tests.

Table 3. Key MRFA statistics, Test C1 (N=4,863)

Total count	Measure	Model S.E.	Infit		PTME	Tasks
			MnSq	ZStd		
12	1.15	0.68	1.64	1.6	0.23	172-T2
20	0.90	0.64	0.53	-1.1	0.60	912-T2
472	0.79	0.11	0.99	-0.1	0.41	702-T2
16	0.68	0.68	0.76	-0.6	0.65	822-T2
360	0.65	0.13	0.84	-2.2	0.48	692-T2
660	0.62	0.10	0.93	-1.3	0.39	072-T2
504	0.32	0.11	0.95	-0.8	0.47	582-T2
676	0.29	0.09	1.08	1.4	0.42	262-T2
3584	0.05	0.04	1.04	1.6	0.53	202-T2
396	-0.06	0.12	1.05	0.7	0.43	402-T2
3456	-0.13	0.04	0.91	-3.9	0.53	532-T2
3924	-0.23	0.04	0.98	-0.7	0.52	832-T2
3404	-0.27	0.04	1.04	1.7	0.52	322-T2
696	-0.29	0.10	0.99	-0.2	0.51	482-T2
136	-0.47	0.41	0.87	-0.6	0.54	902-T2
600	-0.60	0.09	0.89	-2.3	0.48	522-T2
44	-3.42	1.22	1.36	0.7	-0.05	942-T2
1115.3	0.00	0.27	0.99	-0.4	Mean	
1400.0	0.99	0.33	0.23	1.5	S.D.	

Table 3 shows fit to the model to be generally good. Infit mean square figures are good, being within 0.5-1.5; no high standardised t-test scores are above 2.0, and all point measure correlations are high indicating good model fit.

One task (172-T2) has an unacceptable infit mean square figure as well as a high standard error; this task has, however, been taken by a very small number of candidates. Three other tasks, with high standard errors (and in bold font) have also been taken by a very small number of candidates. If these four tasks are removed from the analysis, a logit range of 1.5 logits (+0.79 to -0.60) is observed. This range is slightly larger than the logit range of the B2 tests but it is nonetheless indicative of a set of tasks with good statistics that present candidates with input of comparative difficulty.

Conclusion

This study has explored the issue of task difficulty across two of LanguageCert's CEFR-linked IESOL Writing tests -- the B2 and C1 level tests. The research question focused on the extent to which task difficulty is comparable across tests at the same level. Stability of input is potentially important because if tasks are of significantly different levels of difficulty, it is likely that candidates will produce similarly skewed output thus placing much greater pressure on examiners.

An examination of the tests illustrated that task statistics were generally good. Omitting the small number of tasks with very low numbers of candidates, tasks displayed good fit to the Rasch model – a key background consideration.

While candidates represented a relatively wide ability range (as illustrated by a wide logit range), task difficulty range was constrained to a range of less than one logit for the 13 B2 tests and 1.5 logits for the 13 C1 tests.

More *LanguageCert* examinations will need to be explored but the current study suggests that the tasks analysed from the two LanguageCert IESOL Writing tests may be seen to be comparable in terms of difficulty. Such comparative difficulty of input is the starting point for output produced by candidates against which fair comparisons may be made by examiners.

References

- Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238-257.
- Bae, J., & Lee, Y. S. (2011). The validation of parallel test forms: 'Mountain' and 'beach' picture series for assessment of language skills. *Language Testing*, 28(2), 155-177.
- Barkaoui, K., & Knouzi, I. (2012). Combining score and text analyses to examine task equivalence in writing assessments. In *Measuring Writing: Recent Insights into Theory, Methodology and Practice* (pp. 83-115). Brill.
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110.
- Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing*, 2(2), 153-190.
- Coniam, David. 1992. The effect of choice of question on grade in an essay paper. In Bird, Norman & Harris, John (eds.) *Quill and Quill: achieving and maintaining quality in language teaching and learning*, pp. 442-457. Hong Kong: Education Department.
- Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3(1), 49-68.
- Jung, J., & Bae, J. (2013). The influence of picture prompt variation on writing performance: 'Series' vs. 'Imagine Before and After.' *English Language Teaching*, 25(2), 27-46.
- Koda, K. (1993). Task-induced variability in FL composition: Language-specific perspectives. *Foreign Language Annals*, 26, 332-346

- Linacre, J. M. (1997). Communicating examinee measures as expected ratings. *Rasch Measurement Transactions*, 11(1), 550-551.
- Linacre, J. (2003). Rasch power analysis: Size vs. Significance: infit and outfit mean-square and standardized chi-square fit statistic. Durham, NC: Institute for Objective Measurement.
- Wang, D. (2010). Chinese students' choice of writing topics: A comparison between their self-selected topics and writing prompts in large-scale tests. *Journal of Asia TEFL*, 7(3).
- Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal*, 84(2), 171-184.
- Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing writing*, 6(2), 145-178.



SECTION 3: CALIBRATION STUDIES



Chapter 7: Validating the LanguageCert Test of English Scale: The Paper-based Tests

David Coniam, Tony Lee, Michael Milanovic and Nigel Pike

Abstract

An appropriately validated measurement scale is a necessary prerequisite for any examination/assessment system. Such scales can be developed on the basis of: expert judgement; through the use of statistical techniques; or through a combination of both approaches. However, the most common method of effective scale construction is the use of a combination of expert judgement and statistical techniques.

This report documents the first phase of measurement scale development for the LanguageCert Test of English (LTE). The study describes the validation of the initial LanguageCert Item Difficulty (LID) scale which was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement. The study then builds on the original LanguageCert Item Difficulty scale through the application of Item Response Theory (IRT) and Rasch analysis in addition to expert judgement and CTS. It is anticipated that this enhanced LID scale will form the empirical basis for the alignment of all current and future assessment products to the same measurement scale that is itself aligned to the CEFR.

LTE tests are produced from the LTE item bank. At the time of analysis (early 2021), the bank contained a total of approximately 1,000 items from which four paper-based tests were produced to form the basis for the current study.

The report details how the test with the largest candidature (Test 3) was used as the starting point for the analysis required to establish a baseline measurement scale. Following this, the other three tests were calibrated in their own right in order to provide an initial perspective of the distribution of persons (candidates) and items. Linking items were then anchored against Test 3 logit values, after which the three tests were recalibrated.

Having calibrated the four tests onto a single scale using IRT, this scale was aligned to the original LID scale. Rescaling the calibrated scale from standard logit values to a mid-point of 80 with a spacing factor of 20 resulted in a scale which was comparable to the original LID/CEFR level scale.

The fact that the calibrated Rasch scale produced from the LTE paper-based tests has emerged as well aligned to the original LID scale provides support for further integration of LanguageCert products onto the common scale and validates the use of expert judgement and CTS in the original LID scale creation. Both the whole process and the successful outcome support the view that expert judgment, a time proven human element in assessment, and rigorous statistical modelling, can and should work hand in hand for the benefit of both.

Introduction

This report documents a study based on data gathered for the LanguageCert Test of English (LTE) in order to validate the LanguageCert Item Difficulty (LID) scale created in 2017. The *LanguageCert Test of English* (LTE) is an English 'for work' exam intended for people over the age of 18 in or about to enter the workplace, as well as those in higher or further education. The LTE has been accredited by the UK's Office of Qualifications and Examinations Regulation (Ofqual) and can, therefore be regarded as a high-stakes exam.

The report provides the basis for the development of the original LID scale, based on IRT, that will form the basis for the alignment of all current and future assessment products to the same scale that is itself aligned to the Common European Framework of Reference (CEFR).

Current Study: Purpose

The *LanguageCert Test of English* (LTE) comprises three products, as in Table 1 below.

Table 1: Three LanguageCert test products

Test product	CEFR levels aimed at
(1) a PB test measuring A1-B1	Test aimed at beginner to intermediate cohorts.
(2) a PB test measuring A1-C2	Test for candidates at all CEFR levels
(3) an adaptive test measuring CEFR A1-C2	Test for candidates at all CEFR levels

The purpose of the current study is to validate, link, and establish a common scale for paper-based variants (1) and (2). A follow up study will align this scale to the adaptive LTE test scale ensuring that candidates taking any variant (PB or adaptive) will be consistently placed at the same point on the LID scale. Given that the scores are interchangeable, consistency of measurement across modes of delivery and different versions of the same test is essential for the reliability and validity of LanguageCert tests.

Test Development and Test Administration

The LTE item bank contains a total of approximately 1,000 items, calibrated in line with the original *LanguageCert Item Difficulty* (LID) scale, as laid out in Table 2.

Table 2: LanguageCert Item Difficulty (LID) scale

CEFR Level	LID cut score
C2	160 +
C1	140- 159
B2	120-139
B1	100-119
A2	80-99
A1	60-79
Below A1	0-59

The LID scale was developed on the basis of the expert judgement of a group of assessment and item writing experts who are highly experienced in writing test materials and aligning them to the CEFR. The LID scale is aligned to the six CEFR levels measuring item difficulty in a 0-200 scale where 60 is the cut score level for A1, 80 for A2 and, moving up by 20 points per CEFR level, arriving at 160 at CEFR level C2. Items with a difficulty below 60 are included in the tests as these items measure at the Pre-A1 level. Items above 160 have a ceiling difficulty of 180.

Two PB tests measuring the range A1-B1, and two measuring A1-C2 were assembled as shown in Table 1 above. Table 3 below presents an overview of the four tests that were constructed and the number of candidates taking the tests in this study.

Table 3: Four paper-based tests

Study test name	Items	Number of candidates	Target CEFR levels
Test 1	72	721	A1-B1
Test 2	72	93	A1-B1
Test 3	110	1,161	A1-C2
Test 4	110	137	A1-C2
Total	364	2,112	

Tests 1 and 3 have considerably larger sample sizes than Tests 2 and 4, thus making the analysis for these tests more generalisable.

Common items were included across the four tests and it is on this basis that scale development and calibration were conducted. Table 4 shows the location of common items across the four tests.

Table 4: Common items across tests

	Test 1	Test 2	Test 3	Test 4	Total
Test 1			19	21	40
Test 2			22	20	42
Test 3	19	22			40
Test 4	21	20			42

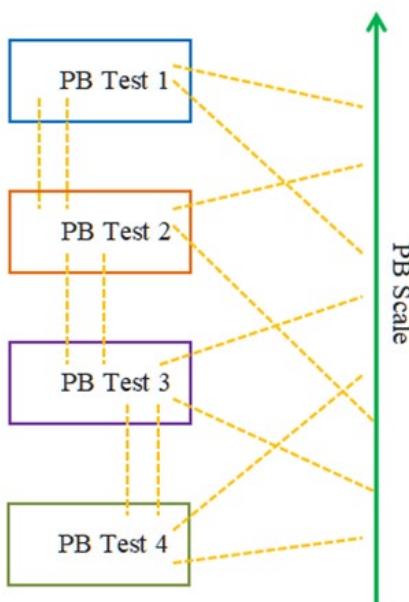
While there was a total of 364 items across the four tests, 82 of these were common items. This meant that there were 282 discrete items in the four-test database.

Calibrating the LTE tests

Given that the main statistical procedures used in this study involve Rasch, the reader is referred to the outline of the Rasch measurement model and the concept of “frame of reference” (FOR) provided in the Glossary of statistical terms.

This section describes the calibration of the four paper-based (PB) tests. The calibration exercise consisted of two key stages. The first involved establishing the internal consistency of the items in all four PB tests and linking these to a unified metric with a view to establishing internal consistency reliability. The second involved pulling all items together to form an initial PB test measurement scale.

In Figure 1 below, the four PB tests analysed in the current study are linked to the common PB scale by equating via common items. In the first instance, while they retain distinct FORs and item locations, and are hence legitimately placed on the PB scale, they have to be interpreted within their own respective FOR.

Figure 1: AT Frame of Reference

Scale Construction Strategy

The principal statistical tool used for calibrating the Rasch scale is the unidimensional Rasch measurement model, using Winsteps (Linacre, 2020). The common or linking items across the PB tests provide anchoring points where all items from the four PB tests form elements of the scale. In addition, the existing linkage to the CEFR levels in the original LID scale underlying the four PB tests, and whether and to what extent the scale is aligned to CEFR levels is investigated.

The scale construction included the steps laid out below.

Step 1

Given that Test 3 had the largest candidature ($N=1,161$) and number of items ($N=110$), Test 3 was taken as the starting point for the analysis to establish a baseline measurement scale. The larger sample size in Test 3 enables a higher degree of precision and stability for this baseline than is the case with smaller sample tests. (Following this, initial results were investigated to establish that the goodness-of-fit for Test 3 was adequate to provide the baseline and starting point of the scale construction.)

Step 2

Test 1 was first calibrated on its own in order to provide an initial view of the distribution of persons (candidates) and items. Linking items were then anchored at Test 3 logits, after which Test 1 was recalibrated. The results of the two calibrations were then compared for any significant distortions that may have emerged in the anchored results. Large discrepancies between the two would indicate 'disturbance' – that is, anchored item values being either under- or over-estimated in the recalibration of either items and/or persons.

Step 3

The same method as in the Step 2 process was followed with Test 2.

Step 4

The same calibration approach was then used for Test 4. Taking anchor items from both Tests 1 and 2 enabled Test 4 to be linked to Test 3 despite the lack of any direct links between the two tests. In a similar fashion, Tests 1 and 2 were linked via linking items obtained from Test 3.

Background to Analysis

The key Rasch analysis elements that form the basis for the analysis and discussion in this report are:

Overall Calibration Tables

Here reference is made to Infit Mean Squares (*IMNSQ*) and Outfit Mean Squares (*OMNSQ*).

Variable (Item / Person) Maps

In the Figures below, item/person maps are laid out such that the person spread (in logits) appears to the left-hand side of the ruler, while the item spread (in logits) appears to the right-hand side of the ruler. Higher level persons (candidates) appear towards the upper left side of the map, while lower level persons appear towards the lower left side of the map. Similarly, more difficult items appear towards the upper right side of the map, while easier items appear towards the lower right side of the map.

In standard Rasch output, logits are presented so that zero is the mid-point with an SD, or spacing factor, of 1 between logits. Under such an output, above zero (a positive value) means a person of higher level and a more demanding item; below zero (a negative value) means a person of lower level and an easier, less demanding item. To make the interpretation of logit values more user-friendly, logits may be rescaled – often with the intention of all values being positive. In the analyses of the initial calibrations of the four tests presented below, 100 was set as the initial mid-point of the scale (zero logits), with one SD rescaled as 20. The red lines in the Figures below indicate these calibration mid-points.

It should be noted that, of the four tests, Tests 1 and 2 aimed at A1-B1 candidates. Candidates sitting these tests have thus only been able to be graded up to B1.

Data Analysis and Interpretation

This section presents the calibration analyses for each test. Test 3 is calibrated first to produce anchor items against which the other three tests may be subsequently linked. The other three tests (Tests 1, 2 and 4) are then calibrated twice: firstly, in their own right and secondly, against Test 3.

Test 3: Initial Calibration

The overall calibration results are presented in Table 5.

Table 5: Test 3 – overall calibration results

PERSON	1161 INPUT		1161 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	58.6	108.9	115.87	4.47	1.00	.0	.99	.0
S.D.	14.0	4.1	13.08	.36	.11	1.3	.23	1.1
REAL RMSE	4.48	TRUE SD	12.29	SEPARATION	2.74	PERSON RELIABILITY	.88	

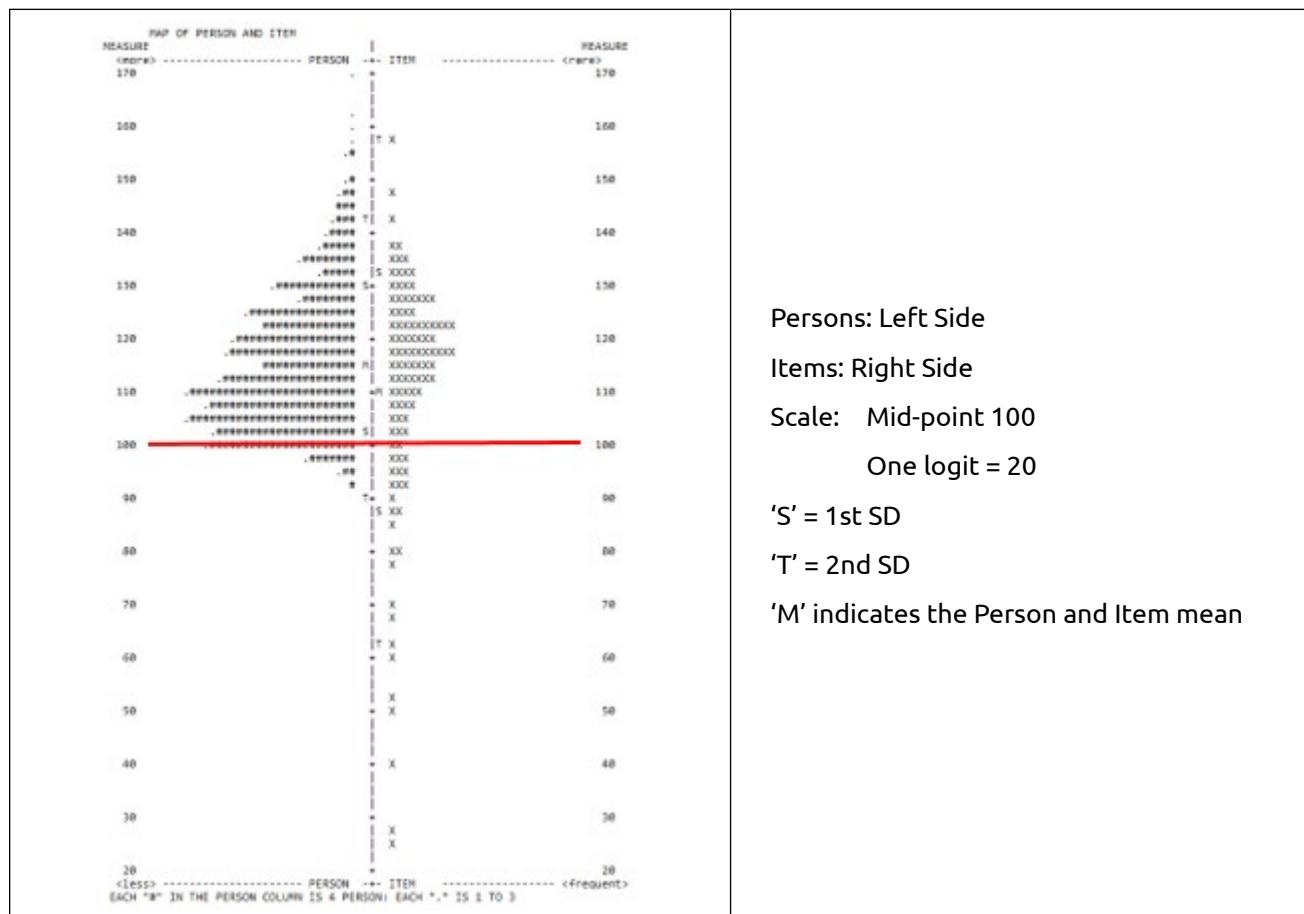
ITEM	110 INPUT		110 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	618.9	1149.2	110.00	1.48	1.00	-.1	.99	-.2
S.D.	221.1	23.8	23.23	.61	.10	4.7	.15	4.7
REAL RMSE	1.61	TRUE SD	23.17	SEPARATION	14.43	ITEM RELIABILITY	1.00	

The key indices to be noted in Table 5 are:

- Reliability (overall), which, for the test items, at 1.0 is very good.
- The Separation index of 14.43 indicates that the True SD (the amount of variance among items) is more than 15 times the error, indicating that there is a large separation and a small standard error in the item calibration.
- The Item Outfit Mean Square (OMNSQ) is the measure (in standard errors [SE]) of how items are grouped around the calibrated measure. In Table 6, item outfit at 0.99 is less than one SE, indicating there are no clear outliers among the items. This confirms that the items form a relatively coherent assessment.
- Item Infit Mean Square (IMNSQ) measures the SEs within an item. Table 6, below, shows item infit to be 1.0 SE, indicating good information (neither too wide nor too narrow) from the options in the items. This suggests that the items have been well constructed.

Figure 1 below lays out the Person/Item calibration map for Test 3. In the analysis of Test 3, logits have been rescaled to a mean of 100 and an SD of 20. The red line indicates the position of the mid-point of calibration.

Figure 1: Person / Item calibration map for Test 3



- From Figure 1, we see that both Person and Item distributions are quite wide and comparatively even in spread. Persons (on the left-hand side) extend from 90 to 150 (3 logits) while Items (on the right-hand side) extend from 90 to 140 (2.5 logits).
- Candidates are generally well matched with items, except for the most able candidates (to the top left of the figure) where there are very few items which match the person abilities.
- Items below 1st SD (85) are too easy as all or nearly all candidates in this sample were able to answer these items correctly.

Calibration of Tests 1, 2 and 4

Following the initial calibration of Test 3, the three remaining tests were analysed. In this procedure, each test was first calibrated in its own right to examine its initial goodness of fit. Tests 1, 2 and 4 were then recalibrated with items anchored to Test 3 through linked items. A number of items which appear in either Test 1 or Test 2 also appear in Test 3. When recalibration of Test 1 and Test 2 was carried out, the items common with Test 3 were anchored at the values set in Test 3. Similarly, re-calibration of Test 4 was anchored at the values of common items between Test 1 or Test 2 with Test 4. Pre- and post-anchoring results were then compared to explore whether any noticeable mismatch occurred between the two calibration exercises. Finally, anchored results were aligned to produce a common scale.

For the sake of efficiency, only the procedure adopted for Test 1 is discussed below.

Test 1: Analysis and Calibration

Test 1 when calibrated in its own right

The overall calibration results for Test 1 are presented in Table 6.

Table 6: Test 1 – overall calibration results.

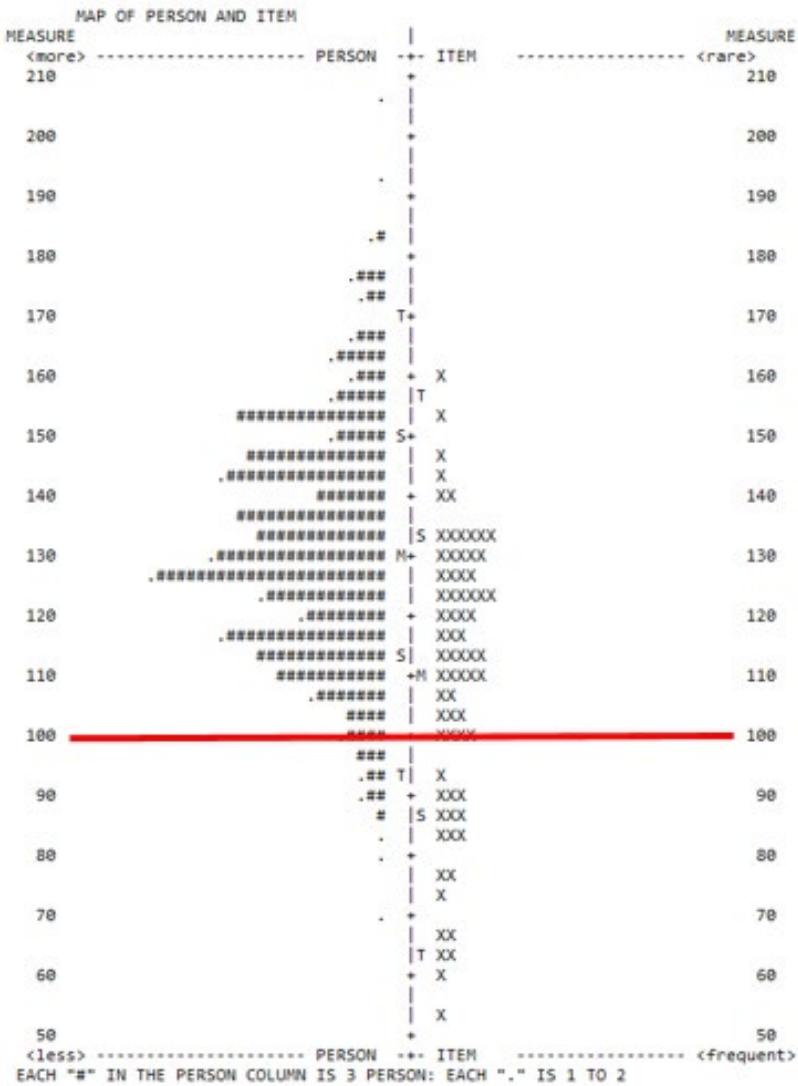
PERSON	721 INPUT		718 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	48.3	71.2	131.61	6.37	1.00	.1	.95	.0
S.D.	11.6	4.2	19.91	1.33	.18	1.3	.37	1.2
REAL RMSE	6.50	TRUE SD	18.82	SEPARATION	2.89	PERSON RELIABILITY	.89	

ITEM	72 INPUT		72 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	481.5	789.5	110.00	2.08	.99	.1	.94	-.2
S.D.	128.3	7.3	28.59	.59	.14	3.5	.30	3.7
REAL RMSE	2.17	TRUE SD	23.49	SEPARATION	10.84	ITEM RELIABILITY	.99	

- Overall item reliability of 0.99 is very high as is item separation at 10.84.
- Item Outfit Mean Square (OMNSQ) is 0.94 which is less than one SE and indicates there are no clear outliers among the items. Item Infit Mean Square (IMNSQ) is 0.99, indicating that good information was provided from the options in the items. This confirms that the items have been constructed well and the items form a coherent assessment.

Figure 2 presents the Person/Item calibration map for Test 1. Logits have been rescaled to a mean of 100 and an SD of 20.

Figure 2: Person / Item calibration map for Test 1



The Rasch analysis suggests the following:

- Both Person and Item distributions are quite wide and even in spread. Persons extend from 85 to 180 (4 logits) while Items extend from 60 to 150 – 4.5 logits.
 - Candidates are located higher on the scale than items, indicating that the test is relatively easy for this group of candidates. Items below the 1st SD (85) and especially the 2nd SD (65) are too easy as all or nearly all candidates were able to answer them correctly.

As mentioned above, Test 1 (and Test 2) was intended only for A1-B1 candidates. Candidates scoring above B1 are graded as B1 by default. This in part may help to account for the discrepancies in the two sets of analyses presented.

Some items may actually be at B2 level (120 in Figure 2 above), given that a few items appeared very difficult for the cohort.

Test 1 calibrated with Test 3 item anchors

The analysis presented below is a reanalysis of Test 1, anchoring it to Test 3 using the 16 common items in the two tests. Table 7 presents these results.

Table 7: Test 1 Calibration with Test 3 item anchors

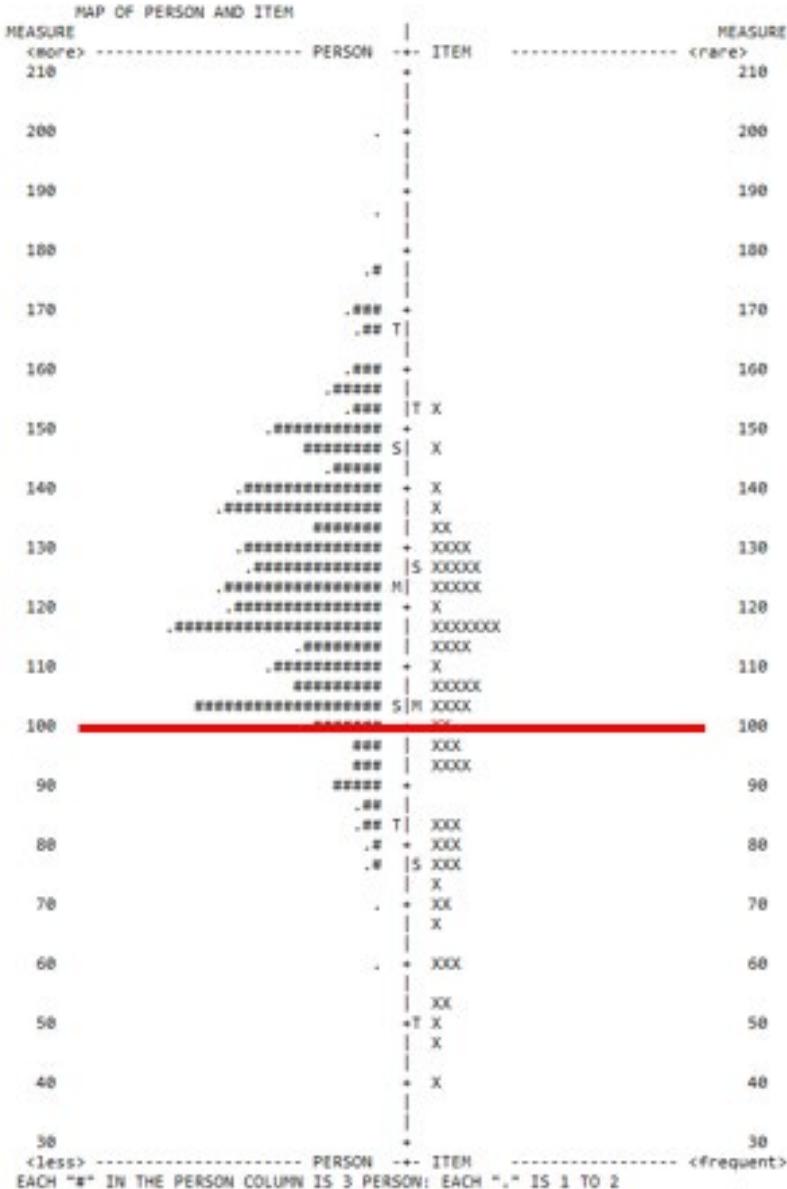
PERSON	721 INPUT		718 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	48.3	71.2	124.49	6.56	1.07	.6	1.16	.6
S.D.	11.6	4.2	20.55	1.29	.19	1.3	.46	1.2
REAL RMSE	6.68	TRUE SD	19.43	SEPARATION	2.91	PERSON RELIABILITY	.89	

ITEM	72 INPUT		72 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	481.5	709.5	102.08	2.31	1.13	1.1	1.16	.7
S.D.	128.3	7.3	26.05	1.17	.55	4.0	.98	4.0
REAL RMSE	2.59	TRUE SD	25.92	SEPARATION	10.01	ITEM RELIABILITY	.99	

- Item overall reliability at 0.99 and separation at 10.01 are both high.
- Item Outfit Mean Square (OMNSQ) is 1.16 – less than one SE; Item Infit Mean Square (IMNSQ) is 1.13, indicating good information being provided from the options in the items.

Figure 3 presents the Person/Item calibration map for Test 1 after anchoring. The logits have been rescaled to a mean of 100 and an SD of 20. The red line indicates the mid-point.

Figure 3: Person / Item calibration map for Test 1, after anchoring



This reveals that:

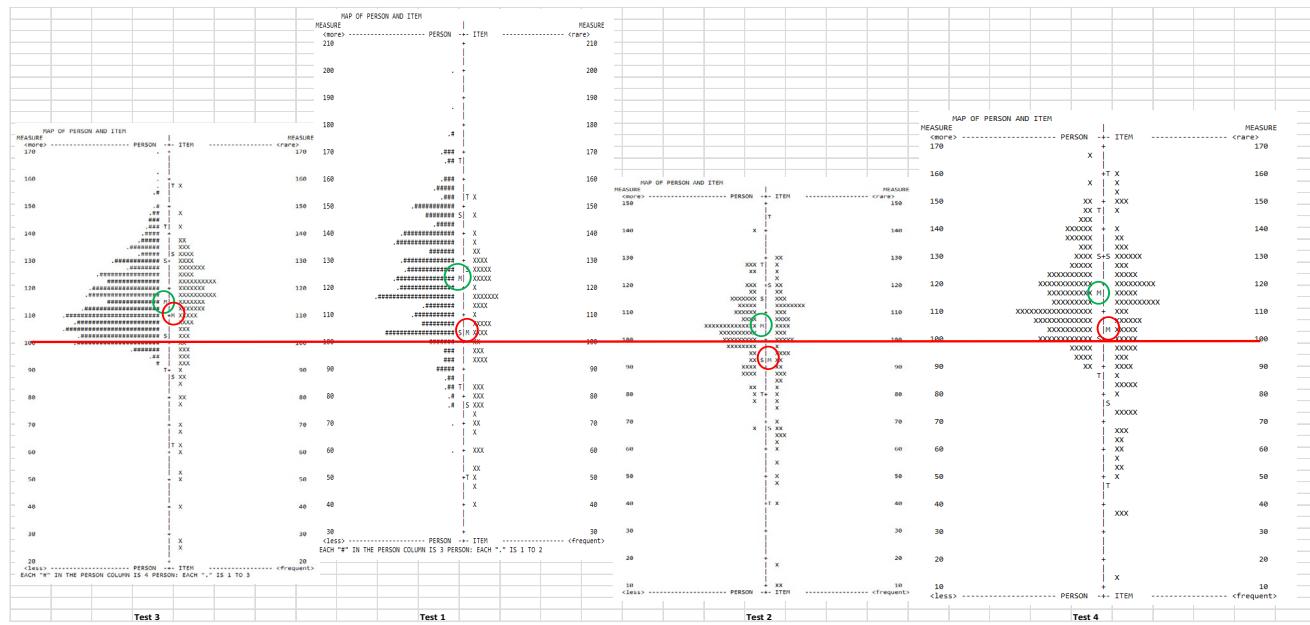
- After anchoring, the Person distribution remains relatively unchanged if slightly lower, ranging from 80 to 170 (4.5 logits).
- The item distribution also remains almost unchanged after anchoring, perhaps shifting slightly lower, with a range of 50 to 150 (5 logits). Quite a number of items to the bottom right of the scale are below the level of the candidates although this is to be expected with a slightly truncated sample.

As mentioned above, the procedure conducted with Test 1 regarding the initial calibration, and then recalibration with Test 3 items anchors was also conducted with Tests 2 and 4.

Recalibrating

Figure 4 now presents a composite picture of the Person/Item maps of the four anchored test calibrations. The mid-points for both Persons and Items are indicated by the circled ‘M’ – green for Persons and red for Items. The red horizontal line indicates the mid-point, the origin of the Rasch scale of 100, or zero logits. The order of presentation of the tests in Figure 4 follows the order in which the tests were calibrated; namely Test 3 first, followed by Test 1, Test 2 and Test 4.

Figure 4: Candidate and Item distributions across the four tests after anchoring

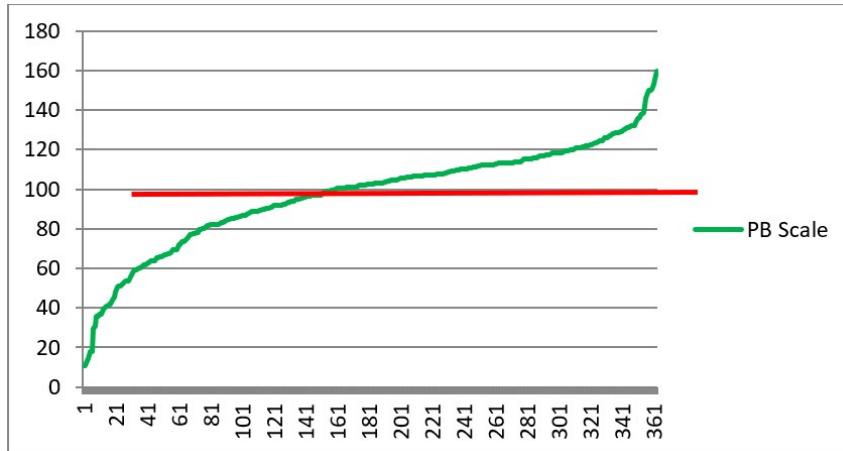


Legend: Red circled ‘M’ = Item mean; green circled ‘M’s = Person mean

Figure 4 illustrates, from the positioning of the mid-point of 100, that Persons and Items in the four tests were generally above the scale mid-point. With the exception of items in Test 2, candidates were generally slightly more able, and items slightly more demanding.

Figure 5 presents the relative difficulty of all 364 calibrated items in the four tests after anchoring.

Figure 5: TCC (Test Characteristic Curve) of 364 calibrated items



The vertical axis of the TCC in Figure 5 represents item difficulty levels and the horizontal axis represent the number of items. We can see that there is a steep progression of difficulty up to 60, indicating that there are relatively few items (about 40) between 10 to 60. Difficulty progression then moves upward steadily, until it reaches 130, when the slope becomes steeper with about 15 items in this section. There are, therefore, about 55 items at the two ends of the item difficulty spectrum, or about 15% of the total and covering the range A1-B2+. The majority of items (about 85%) fall between 60 and 130, across the mid-range of the scale.

Recalibrating the Scale

Having calibrated Tests 1-4 onto a single scale – taking Test 3 as the baseline – the next step involved examining the alignment of the newly-calibrated scale with the original LID scale.

To establish a baseline for Test 3, logit values had initially been rescaled to a mid-point of 100 with a spacing factor of 20. An advantage of Rasch is that, as long as the recalibration with the new mid-point does not alter the original calibration results, different mid-points may be used to suit specific calibration exercises (see <https://www.winsteps.com/winman/rescaling.htm> for an elaboration). Such a procedure may be viewed as being similar to changing individual tests' mid-points via anchoring.

Rescaling was subsequently conducted following discussion with the test development team such that a new mid-point of 80 was applied to match the initial LID scale, with the 20-point spacing factor maintained. Following this realignment, the whole test calibration process, with anchoring, was performed again with Test 3 as the starting point and the other three tests calibrated to Test 3 values. It must be pointed out here that, like all statistical procedures, Rasch calibration is content free. The interpretation of calibration results is guided by considerations beyond the statistical procedure as long as the principles underlying the statistical procedures are not violated. The anchored calibration of the paper-based tests based on Test 3, is the best amongst equals. The initial anchored calibration will gradually be refined as the item bank develops.

The final mapping of the four PB tests onto a single scale with the mid-point of 80 is shown in Table 8.

Table 8: Test 3 – Candidate distributions via LID and Rasch-calibrated (mid-point 80) scales

CEFR level	LID level cut scores	Candidates achieving grade via LID scale	Candidates achieving grade with Rasch (80) scale
C2	160	0%	0%
C1	140	2%	1%
B2	120	10%	10%
B1	100	21%	35%
A2	80	36%	48%
A1	60	28%	5%
pre-A1	40	2%	1%

With the mid-point of 80, the two scales are more closely aligned. Apart from some misalignment at the A level, the two scales are quite comparable.

With the mid-point of 80, Figure 6 now presents candidate distributions across all four tests after anchoring. The order of presentation follows the order of calibration, with Test 3 first.

Figure 6: Candidate and Item distributions across all four tests after anchoring

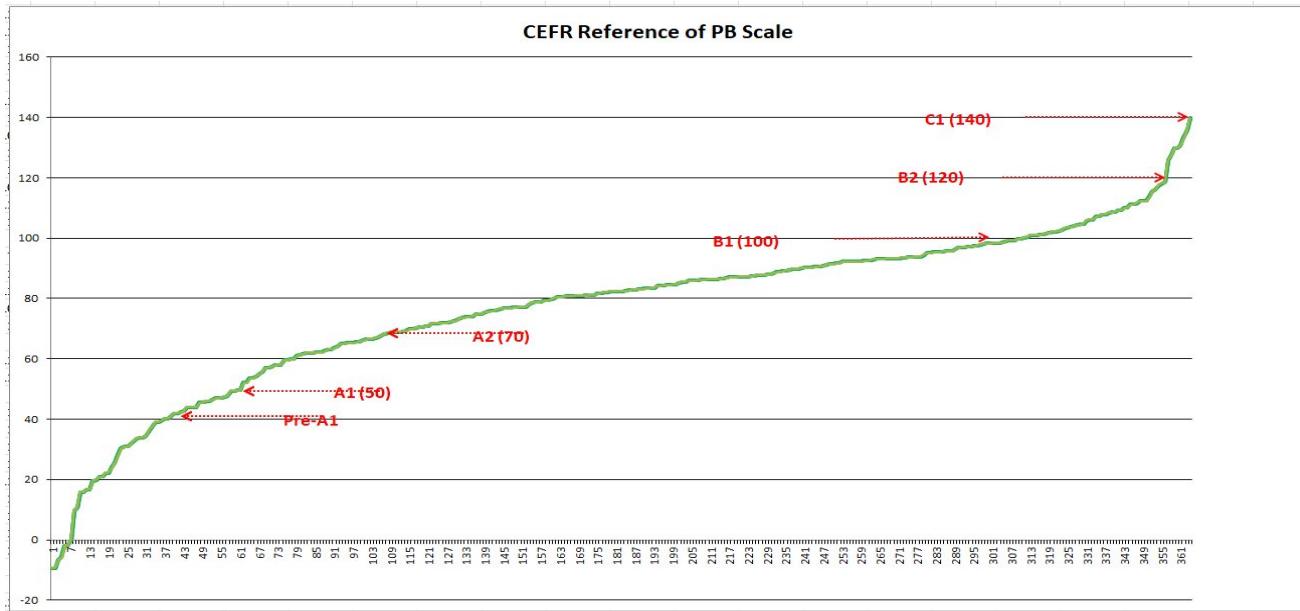
Figure 6 shows that the comparative standing of the four tests vis-a-vis one another has not changed with the use of the new scale mid-point of 80.

Table 9 and Figure 7 below present the finalised Rasch-calibrated PB scale, illustrating how the recalibrated scale matches CEFR levels and the original LID scale.

Table 9: LID and Rasch-calibrated PB scale cut score match

CEFR level	LID cut scores	Recalibrated Rasch scale cut scores
C2	160	
C1	140	140
B2	120	120
B1	100	100
A2	80	70
A1	60	50
Below A1		

In Figure 7 below, the red arrows and values represent the cut score points for the different levels.

Figure 7: Finalised Rasch-calibrated PB scale

The LID scale was initially developed as a linear scale, with the cut scores for each CEFR level as in Table 9 above. As can be seen, there is a very close fit between the original LID scale and the Rasch-calibrated scale. C1, B2 and B1 match exactly, with 20 points (or one logit) between each level. Between B1 and A2, the Rasch analysis suggests a slightly wider gap – 1.5 rather than one logit. Between A2 and A1 there is again a 20-point difference. Between A1 and pre-A1, the Rasch analysis suggests the gap should be only half a logit or 10 points.

In sum then, the Rasch-calibrated scale from pre-A1 up to C1 extends 100 points, or five logits, with the Rasch rescaling corresponding very closely – with the exception of A1 and A2 – to the original LID scale. The weaker alignment here needs to be investigated further.

Conclusions

The study reported above had two major objectives. The first was to calibrate, using Rasch measurement, the existing paper-based version of LanguageCert Test of English (LTE) onto a common scale; the second was to examine the subsequent alignment of the common scale produced, with the existing LanguageCert Item Difficulty (LID) scale developed on the basis of Classical Test Statistics (CTS) and expert judgement in order to lay the foundations for a single unified measurement scale, aligned to the CEFR, that would underlie all LanguageCert assessment products.

The report details how Test 3 with the largest candidature ($N=1,161$) and number of items ($N=110$) was taken as the starting point in terms of establishing a baseline measurement scale. The other three tests, after having been first calibrated in their own right, were then anchored to Test 3 via linking items drawn from Test 3, after which the three tests were then recalibrated. This process resulted in all four tests eventually being calibrated onto a single scale.

With all four tests on a single scale, the calibrated scale was rescaled to a mid-point of 80 with a spacing factor of 20 in order to align the calibrated Rasch scale and the original LID scale. The rescaling of the Rasch scale in this manner produced a comparable alignment between the two scales although there some differences detected at the A1, A2 and B1 levels require further exploration.

The next step is to calibrate the LTE adaptive test also generated from the LTE item bank, to the common Rasch scale produced in the current study. This will also entail a revisiting of the Frame of Reference concept. This is the subject of Chapter 9 that follows this chapter.

References

- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J.: Erlbaum.
- Fisher Jr, W. P. (1992). Reliability statistics. *Rasch Measurement Transactions*. Chicago, USA: MESA Press, 6, 238.
- Humphry, S. (2006). The impact of differential discrimination on vertical equating. ARC report.
- Linacre, J. M. (2006). A user's guide to WINSTEPS: Rasch-model computer programs. Chicago, IL: Winsteps.com.
- Linacre, J. M. (2006) Winsteps. Rasch measurement computer program. Chicago: MESA Press.
- McNamara, T. (1996). *Measuring second language performance*. New York: Longman.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing* 15 (2), 263-287.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndsills, UK: Palgrave Macmillan.
- Wright, B. D. (1992). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions*, 6, 196-200.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, Illinois: MESA Press.

Chapter 8: Validating the LanguageCert Test of English Scale: The Adaptive Test

David Coniam, Tony Lee, Michael Milanovic and Nigel Pike

Abstract

This paper reports on the calibration to a common scale of the LanguageCert Test of English (LTE) in its computer adaptive mode which in turn builds on the calibration of LTE paper-based tests reported in Coniam et al. (2021). The work described here now ensures that the LanguageCert Item Difficulty (LID) scale can be used for all LTE modes of delivery and that item difficulties in LTE align with all LanguageCert tests that make reference to the LID scale. The calibrated LTE item bank is therefore a robust source of materials for both the paper-based and computer-based adaptive tests.

Introduction

The LTE, which is accredited by the UK's Office of Qualifications and Examinations Regulation (Ofqual), is an English 'for work' exam intended for people over the age of 18 in or about to enter the workplace, as well as those in higher or further education.

The current study builds on Coniam et al. (2021), which documented the first phase of measurement scale development for the LanguageCert Test of English (LTE). That study described the validation of the LID scale via the LTE paper-based tests. The LID scale was created between 2017-2019 on the basis of classical test statistics and expert judgement. The LID scale is the empirical basis for the alignment of current and future LanguageCert assessment products to the same measurement scale that is itself aligned to the CEFR.

The Coniam et al. (2021) study focused on the LTE paper-based tests, which, after being calibrated, were placed on a common scale. The current study extends the LTE calibration process by demonstrating how the LTE adaptive test is calibrated to the same common LID scale as the paper-based tests. It demonstrates how candidates taking either a paper-based or an adaptive LTE test will be placed at more or less the same point on the LID scale regardless of which form of the test they take.

Current Study: Background

The LanguageCert Test of English (LTE) comprises three products, as in Table 1 below.

Table 1: Three LanguageCert test products

Test product	CEFR levels aimed at
(1) a paper-based test measuring A1-B1	Test aimed at beginner to intermediate cohorts.
(2) a paper-based test measuring A1-C2	Test for candidates at all CEFR levels
(3) an adaptive test measuring CEFR A1-C2	Test for candidates at all CEFR levels

The Coniam et al. (2021) study reported on the validation, linking, and establishing of a common scale for paper-based variants (1) and (2). The current study reports on the alignment of variant (3), the adaptive test, to the LID scale. The purpose of the current study, as mentioned, is to ensure that candidates taking any variant (paper-based or adaptive) will be consistently placed at the same point on the LID scale. Given that the scores are interchangeable, consistency of measurement across modes of delivery and different versions of the same test is essential.

Initial development and calibration of the LID scale had its origins in a compilation of the LTE paper-based tests (Coniam et al., 2021), with the latter study showing the four paper-based tests to be robust and the calibrated scale which emerged to be consistent with the data. The initial scale provided an acceptable basis for the development of the full LanguageCert scale on the basis of the adaptive test data – the focus of the study reported in the current paper.

While the current study reports on the LTE adaptive test, it must be restated that it is the LID scale, not the adaptive test, that is the focus of this study. For detail on adaptive testing and an overview of the LTE adaptive test, its algorithm and operation, the reader is referred to Pike & Coniam (2021).

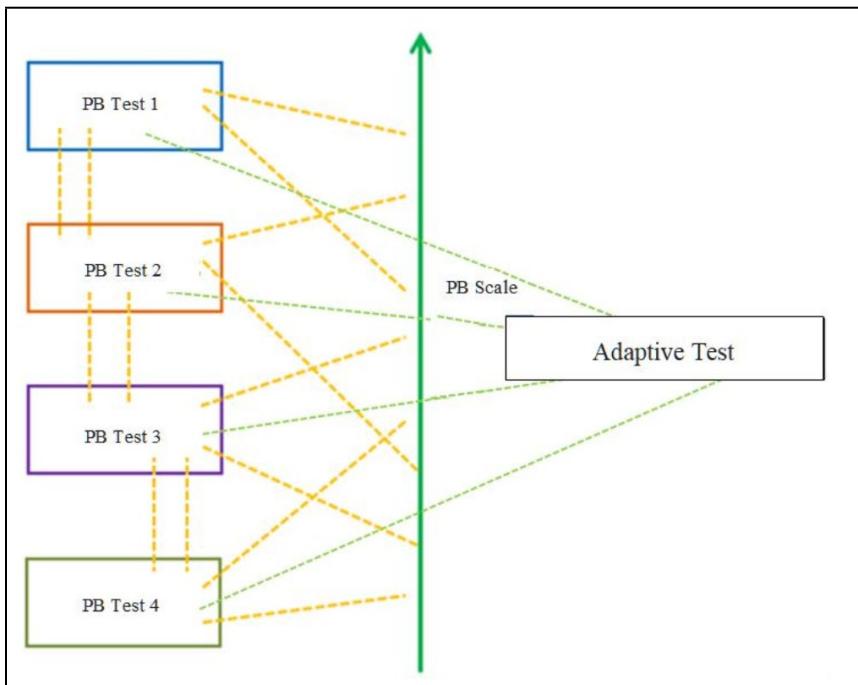
Since the calibration for both the paper-based tests and the adaptive test have made use of Rasch measurement, the reader is referred to the outline of the Rasch measurement model and the concept of the frame of reference (FOR) provided in the Glossary of statistical terms at the end of the volume.

Figure 1 illustrates the frame of reference in the context of the calibration of the paper-based LanguageCert Test of English (LTE) (Coniam et al., 2021). In that validation study, a common scale was constructed for the LTE via four paper-based (PB) –referred to as Tests 1-4 in Figure 1 below. The green arrow separating the two sets of tests is the calibrated LID scale.

In operational terms, two yardsticks indicate whether an item may be accepted within the FoR of two tests:

1. That item difficulty in both tests is comparable: there is less than 0.5 of a logit between item measures.
2. That item values occur in roughly similar positions in both tests; i.e., both items are, say, within the top 25th percentile.

Figure 1: LanguageCert adaptive test Frame of Reference



Key: PB = paper-based

The LTE adaptive test includes items which may also appear or have appeared in the paper-based tests. As a combined data matrix, however, the adaptive test constitutes a distinct and separate FOR from the paper-based tests. This is somewhat different from an adaptive test in the usual sense of the term. There, candidates are presented with items one item at a time from the interim paper-based scale and thus remain within the same FOR of the paper-based scale. Anchoring the items in the LTE adaptive test with values from previously-calibrated paper-based tests may not necessarily fit the new FOR: each test is an individual entity, and as such, values cannot simply be transferred from one to another. The two conditions laid out above first need to be satisfied.

In terms of analysis, the corollary is therefore that an FOR should be established for the paper-based tests, through linking via anchor items. A similar procedure is then conducted for the adaptive test – internal linking via anchor items. Once robust scales have been determined for both FORs, a merging of the two scales – of the two FORs, that is – may then be attempted. How this is achieved practically in the context of the LTE adaptive test in relation to the already existing calibration of the paper-based tests – with both being eventually merged onto a common scale – is discussed below.

The LanguageCert CAT

The LTE adaptive test assesses listening and reading from CEFR levels A1 to C2. Development began in 2019, with an initial item bank of approximately 400 items consisting of a range of listening and reading items and testlets (mini tasks of 2-5 connected items) which assessed different listening and reading constructs. The item bank furnishes test materials for both the paper-based and the computer adaptive tests. The adaptive

test was trialled in late 2019 and went live in April 2020. The trial adaptive bank had approximately 900 items and the first live adaptive bank approximately 800 items. Items are continually being added to the core LTE item bank, and by early 2021, the bank comprised over 1,500 items. As is necessary with all item banks, the item bank will continue to be refreshed and grow in the future.

Analysis

The analysis of the adaptive test was conducted in early 2021, at which point the dataset consisted of 827 items and 5,870 candidates. The analysis of this dataset via the Rasch analysis software Winsteps (Linacre, 2006) is described below. Table 2 below details the summary statistics for the calibration.

Table 2: Adaptive Test Calibration Details

PERSON	5870 INPUT		5870 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	38.1	56.9	134.92	6.66	1.00	.0	1.00	.0
P.SD	5.3	2.4	32.07	1.00	.12	.9	.38	.9
REAL RMSE	6.74	TRUE SD	31.35	SEPARATION	4.65	PERSON RELIABILITY	.96	

ITEM	827 INPUT		827 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	270.3	403.8	99.25	3.95	.99	-.2	1.00	-.2
P.SD	195.6	266.6	39.93	4.97	.11	2.3	.27	2.3
REAL RMSE	6.35	TRUE SD	39.42	SEPARATION	6.21	ITEM RELIABILITY	.97	

A total of 5,870 candidates and 827 items were included in the calibration. Candidates took on average 57 items, from which a mean raw score of 38.1 emerged. Item reliability is high at 0.97, as is person reliability at 0.96, the latter being the equivalent of classical test theory reliability (Anselmi et al., 2019). Person infit mean-square (1.00) and outfit mean-square (1.00) fit statistics are both within the acceptable range of 0.5 to 1.5, suggesting that the calibration of persons may be taken as acceptable. By the same token, item infit mean-square (0.99) and item outfit mean-square (1.00) fit statistics are also acceptable. The overall summary calibration statistics point, therefore, to a test that may be viewed as sound.

The overarching LanguageCert Item Difficulty (LID) scale lays out item difficulty levels generally adopted in LanguageCert assessments (Coniam et al., 2021). These are presented in Table 3.

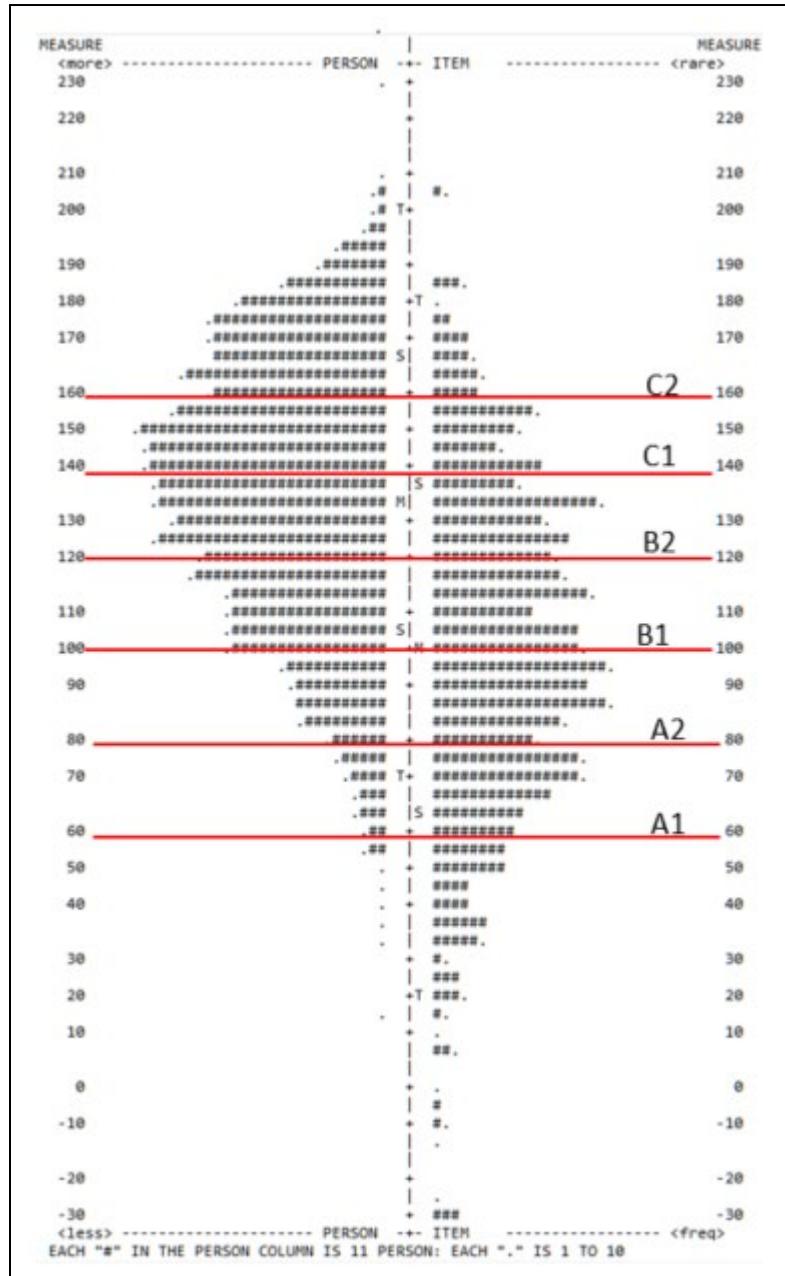
Table 3: LID scale

CEFR level	LID scale range	Mid point
C2	151-170	160
C1	131-150	140
B2	111-130	120
B1	91-110	100
A2	71-90	80
A1	51-70	60

To give a visual overview of the measurement, the vertical ruler (the ‘facet map’) produced in the Winsteps output is presented below in Figure 2. This is a visual representation of where facets (items and candidates)

are located on the scale. In Figure 2 below item/person maps are laid out such that the person spread (in logits) appears to the left-hand side of the ruler while the item spread (in logits) appears to the right-hand side of the ruler. Higher level persons (candidates) appear towards the upper left side of the map while lower level persons appear towards the lower left side of the map. Similarly, more difficult items appear towards the upper right side of the map while easier items appear towards the lower right side of the map.

Figure 2: Person-Item facet map



As Figure 2 illustrates, person and item distributions are quite wide and comparatively even in spread. Both extend approximately 120 points, or six logits – the rule-of-thumb operational range (Bond et al., 2020). Persons (on the left-hand side) extend from 60 to 200 while Items (on the right-hand side) extend from 30 to 170.

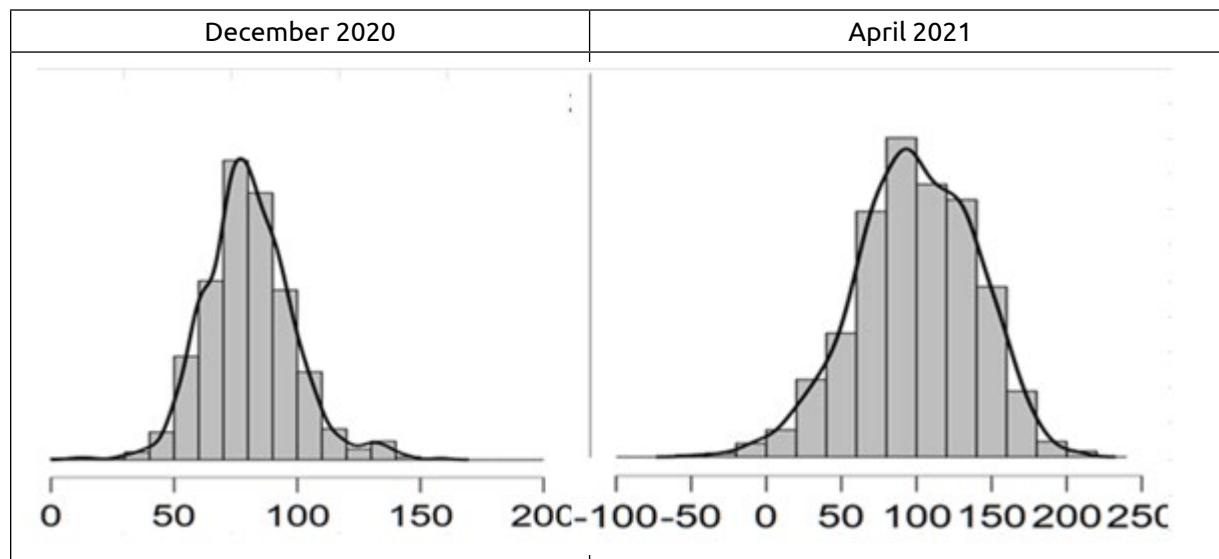
Candidates generally match with items. The midpoint of the item curve may be seen to be around B2; with persons, the midpoint of the curve may also be seen to be around B2. The Person distribution is, however, dependent upon the nature of the test population in this sample. It is known, for example, that there were a considerable number of high ability candidates in the sample.

Brief Comparison with Earlier Calibration

The analysis described in this paper was conducted in early 2021 when the adaptive test consisted of 827 items and 5,870 candidates. A previous, exploratory analysis had been conducted in late 2020, at which point the dataset consisted of 820 items and 1,575 candidates. The section below presents some comparative analyses of the two datasets, in order to give a sense of how, with the increase in size, dataset robustness has, unsurprisingly, improved.

Figure 3 presents a virtual anchoring of items in each dataset, with the curve indicating the peak, the mid-point of each dataset.

Figure 3: Item spread in datasets



While the Pearson correlation between the two sets of data was 0.96, the mid-point has shifted slightly upwards – from items centring around 80 (B1) to around 100 (B2)

Table 4 presents an elaboration of the April 2021 dataset, with the percentiles indicating CEFR levels, and LID scale values.

Table 4: Percentiles indicating CEFR LID scale position (as of April 2021)

		Level	LID scale value
No. of items	816		
Mean	100.76		
Std. Deviation	37.98		
Maximum	204.47		
		C2	150
		C1	130
75th percentile	129.42		
		B2	110
50th percentile	99.92		
		B1	90
25th percentile	73.64		
		A2	70
		A1	50
Minimum	6.95		

The expanded calibration of the adaptive test, in terms of both item and candidate numbers, has shown improvement in the rigour of the LID scale from two key aspects:

1. The scale mid-point (the 50th percentile) is now 100 (99.92), which closely matches the item distribution mean (100.76).
2. Levels A1 and A2 now occur in the bottom 25th percentile, levels B1 and B2 in the central 50th percentile, and C1 and C2 in the top 25th percentile. Such a distribution might possibly be expected of any large candidate sample size. Everything else being equal, the mid-range ability group would be expected to occupy the major central region of the distribution while the higher and lower ability groups would be expected to occupy the upper and lower narrower range of ability.

Conclusion

As outlined in Coniam et al. (2021), the LanguageCert LID scale for all LanguageCert tests, was developed and calibrated initially against a set of paper-based tests. The initial calibrated scale that emerged provided a validation of the paper-based tests, showing them to be robust and consistent with the data. The initial scale therefore provided an acceptable basis for the further development of the LanguageCert Item Difficulty scale and the integration of LTE on to the overarching LID scale on the basis of the adaptive test data.

The focus of the current study has been to calibrate the expanded set of items in the item bank against the cohort of candidates who have thus far taken adaptive tests from the LTE item bank.

With the extension and expansion of the scale and the item bank, measurement statistic configurations necessary to achieve the goal of a robust calibrated scale have had to be taken account of. Specifically, the concept of the frame of reference for measurement has been instructive in setting parameters for co-configuring the paper-based tests as one entity, and subsequently incorporating the expanded item bank dataset and adaptive test into a single frame. It is now possible to see, post hoc, after anchoring, that the different tests match up. All the tests may now be viewed – and may operate – within the same frame of reference.

The calibrated LanguageCert Item Difficulty (LID) scale may now be considered to be a comprehensive scale, linked to an item bank which provides both anchoring from individual tests with different FORs and individual item-based adaptive tests.

With a coherent LID LTE scale having been developed, two further projects are now being undertaken. The first of these involves a comparative study of both versions of the LTE. This involves administering – to a representative sample of candidates – versions of both the LTE paper-based test and the adaptive test. The second project, which is ongoing, involves an expansion in the size of the item bank, with concomitant confirmatory re-analysis. Currently, as reported in this paper, the item bank comprises 827 items. Once the candidate cohort reaches 10,000, further analysis will be conducted and the robustness of the calibration revisited.

References

- Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, 10, 2714.
- Bond, T., Yan, Z., & Heene, M. (2020). Applying the Rasch model: Fundamental measurement in the human sciences. Routledge.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). Validating the LanguageCert Test of English scale: The paper-based tests. London, UK: LanguageCert.
- Goodman, L. (1990). Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. *Sociological Methodology*, 20, 249-294.
- Humphry, S. (2006). The impact of differential discrimination on vertical equating. *ARC report*.
- Humphry, S. M., & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement*, 9(3), 249-264.
- Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum.
- Pike N., & Coniam, D. (2021 in press). Adaptive testing and the LanguageCert Test of English adaptive test. ELTNEWS.

Chapter 9: Externally-Referenced Anchoring: Equating Expert Judgement and Rasch Measurement Values in LanguageCert IESOL English Language Tests

Tony Lee, Michael Milanovic, David Coniam and Nigel Pike

Abstract

This paper reports on the use of *externally-referenced anchoring* by LanguageCert as a methodology for calibrating language test materials and aligning test forms. The datasets used in this paper are taken from tests at each of the six levels of LanguageCert IESOL suite, all of which have been aligned to the CEFR through expert judgement. We illustrate in this paper the extent to which externally-referenced anchoring, using Item Response Theory (IRT) but based on expert judgement, can be used as an effective, reliable and valid methodology. The approach is based on the premise that successful anchoring may be achieved by reference to well-targeted, expertly-written test forms aligned to the underlying traits of a particular CEFR level by expert judgement and verified through the use of IRT [Note 1].

This study focuses on the analysis of 18 LanguageCert test forms, three at each CEFR level. The LanguageCert Item Difficulty (LID) scale, which underlies all LanguageCert test materials, is linked empirically to the CEFR and each test was placed on the LID scale based at the midpoint of its distribution. This was then set as the externally-referenced anchor for a given CEFR level.

The findings of this study indicate that, while the match between the distribution of items in the selected LanguageCert IESOL tests and the LID scale was not perfect, in general, a relatively close match between the

items in the tests and the LID scale was found and, as a consequence, their link to the corresponding CEFR level. For each test, most of the items fell between the 25th and 75th percentile of any given level: this range representing the lower and upper bounds of LID scale values for each CEFR level. These results demonstrate that LanguageCert IESOL test items are well set and appropriately positioned at respective CEFR levels on the basis of expert judgement. The study illustrates that *externally-referenced anchoring* based on expert judgement may be used as a methodology for aligning test forms to an external frame of reference, in this case the CEFR.

Expert Judgement and Test Setting

'Expert judgement' is a key factor in language assessment test development both in the area of item writing and test setting as well as in the estimation of item difficulty, which in turn impacts level setting and cut scores. In the case of test setting, the use of experts is a critical requirement. Rodriguez (1997) refers to item writing as an art, while Bristol (2015) describes the creation of examination questions as both an art and a science. Haladyna & Downing (1989) provide a set of seven ground rules originally selected for good item setting, some of which are echoed in Alderson et al. (1995) where the qualities of an expert item writer are cogently discussed. What is clear however, is that training and experience are necessary characteristics of successful item writing. Coniam (1997) suggests that well-trained and competent item writers may be expected to achieve a 'quality setting' rate of around 70% and above; that is, 70% of the items such writers produce make their way into a live test or examination. In a follow up article, Coniam (2009) observes that barely-trained item writers are unlikely to achieve a quality setting rate of more than 20%. This leads us to the conclusion that good tests – with good items and an accurate reflection of a given proficiency level – can be produced efficiently by well-trained and experienced writers.

There has been considerable discussion of the use of expert judgement in standard setting, with difference of opinion in some quarters – Alderson & Kremmel (2013), for example. Generally, however, the use of expert judgement has been widely employed in the field of language assessment for test validation and standard setting – see Lumley, 1993; Bachman et al, 1995; Gable & Wolf, 1993. Recent validation studies involving expert judgement include VanderVeen et al. (2007), Song (2008), Gao and Rogers (2011), and van Steensel et al. (2013), studies in which judges were reported to have reached high levels of agreement.

LanguageCert and the CEFR

There are six examinations in the *LanguageCert* International ESOL suite, all aligned to the six CEFR levels: Preliminary (A1), Access (A2), Achiever (B1), Communicator (B2), Expert (C1) and Mastery (C2). The *LanguageCert* examination specifications reflect the requirements of the CEFR; test materials writers represent the highest international standards and have extensive expertise in, and knowledge and understanding of, the CEFR.

All *LanguageCert* test setters meet minimum requirements in terms of professional qualifications and experience in order to be eligible for consideration as an item writer. There is an extensive item writing manual which lays out in detail how to write items and how to achieve appropriate quality standards.

Each IESOL test has a designated CEFR level, with, as mentioned, all test forms carefully set using expert judgement and reviewed by other expert staff.

The LanguageCert Item Difficulty (LID) scale is the metric against which items are linked to the CEFR on the basis of item difficulty. The LID scale was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement by a group of assessment and item writing experts who are highly experienced in writing test materials and aligning them to the CEFR. The LID scale is presented in Table 1.

Table 1. LID scale

CEFR level	LID scale range
C2	170-150
C1	150-130
B2	130-110
B1	110-90
A2	90-70
A1	70-50

Studies by Coniam et al. (2021) have validated the LID scale and extended it beyond its original CTS origins to a Rasch-based calibration where all levels are statistically validated and linked.

Given that the main statistical procedures used in this study involve Rasch, the reader is referred to the outline of the Rasch measurement model and the concept of “frame of reference” (FOR) provided in Glossary of statistical terms.

To achieve meaningful test anchoring, it is important to consider a fundamental tenet: that the starting point of a Rasch calibration is the mid-point of the calibration. This is the estimation of the point in a test at which a candidate has a 50/50 chance of answering the item/s correctly. A test, if specified to measure at a particular level of ability, should have the mid-point of the item distribution of the test in question anchored at a position in a scale representing that level of ability.

LanguageCert, the CEFR and Externally-referenced Anchoring

Coniam & Lampropoulou (2020), in their analysis of 62 LanguageCert IESOL Listening and Reading tests using classical test statistics, showed that LanguageCert IESOL tests are well constructed and robust. However, despite being robust and comparable, these IESOL tests had not been calibrated using IRT and anchor items to a single scale.

The most frequent manner of calibrating tests onto a single scale generally involves using common items between the different tests and cross-calibrating them via the Rasch scale, or via persons found in both tests and the Rasch scale. At times, however, the construction of the tests is such that there are no common elements – test items, person, or even examiners, through which linking via Rasch scale locations may be

established. An alternative approach, which is investigated in this study may be referred to as ‘virtual’, or ‘externally-referenced’ linking. Linacre (2018) outlines situations where no common (or identical) items exist although items do exist that might be defined as measuring the same trait.

Boone & Staver (2020) exemplify the concept of virtual linking – or ‘virtual anchoring’ – in the context of mathematics where two simple additional items are presented as being construed to share the same underlying trait. While there has been some research reported on the use of virtual anchoring, this has only been in the context of test equating (Longford ,2015; Boone & Staver ,2020; Lupescu, 1996, 2005. Further, in the latter two studies, the focus has been on the tracking of persons, with the methodology essentially being that of regression onto a latent variable from raw scores. In contrast, the current research presents externally-referenced anchoring in the context of test items. Following the use of fit statistics to first explore the robustness of the measurement, the focus of the study is on revealing the latent trait.

A similar use of externally-referenced anchoring to that used in the current study was conducted by Humphry et al. (2014). In the context of standard setting, and the use of a modified Angoff approach, Humphry et al. (2014) used a form of externally-referenced anchoring to explore how, via use of Rasch measurement, the expert rater scale might be aligned with the test taker scale.

In the current context, externally-referenced anchoring may, therefore, be seen through the lens of expert setters. Test forms have no common items but comprise items which have been set at predefined and well-accepted CEFR levels. The six well-understood CEFR levels may therefore be construed as being six ‘items’, with each item sharing the same underlying trait. This can be seen to be akin to how content-related items (as with the two mathematics addition items referred to above) share the same underlying trait.

As mentioned above, in line with Rasch principles, a test should ideally be anchored at the mid-point of the item distribution of a given scale. The mid-points of the LID scale for the six CEFR levels are presented in Table 2.

Table 2. LID scale

CEFR level	LID scale range	Mid point
C2	170-150	160
C1	150-130	140
B2	130-110	120
B1	110-90	100
A2	90-70	80
A1	70-50	60

While there are many IESOL test forms at each CEFR level, typically there are no linking items or candidates by which cross-calibrating may be conducted. Externally-referenced anchoring using the calibrated mid-point of a given CEFR scale is therefore the method used in the current study in order to anchor the different IESOL tests onto the LID scale. The frame of reference in this case does not constitute the items but rather the CEFR scale locations calibrated through the items involved. The critical anchoring parameters in this instance are therefore the expert-rated CEFR levels of the items in a given test and the calibrated CEFR locations on the LID scale.

In order to investigate the extent to which such expertly-written yet uncalibrated test forms were indeed equivalent in terms of difficulty and level, the externally-referenced anchoring approach was applied whereby each test's midpoint was taken as an accurate representation of the level in question. The midpoint of each test in this context would then:

1. enable an effective calibration of the items in each of the IESOL tests given that no other restrictions are imposed on the items.
2. reveal the goodness of fit between the calibrated item distributions and the expertly assigned CEFR levels. The fit is determined by whether a broadly bell-shaped distribution of item measures emerges where the majority of item measures are clustered around the mean and fall between the 25th to 75th percentile and so largely within a given level.

When test development takes place, the mid-point of an individual test is intended by the test developers to represent a given CEFR ability level. It was decided to anchor the tests to the LID scale level via the mid-point for each test, which, it is argued, in turn anchors each test to the CEFR. The goodness of match of the anchoring is evaluated by the extent to which the mid-range of the items in the tests coincides with the CEFR levels on the LID scale and the extent to which the mid-range of the test item distribution includes most of the items in each test.

In this study, three IESOL tests randomly selected for each CEFR level – 18 test forms in total – are anchored by external referencing following the procedure described.

Analysis

There are a number of key analytics usually conducted when doing Rasch measurement – and which have been reported on in previous LanguageCert studies (see e.g., Coniam et al., 2021). The first of these involves the 'fit' of the data to the Rasch model, referring, in essence, to how well obtained values match expected values. 'Fit' itself is divisible into a number of related, if slightly different, categories. A perfect fit of 1.0 indicates that obtained values match expected values 100%. Acceptable ranges of tolerance for fit range from 0.7 to 1.3 (Bond et al., 2020). Key statistics usually reported on are item outfit mean squares, item infit mean squares, and Reliability.

A summary of the analysis of the 18 tests – three at each CEFR level, with each test comprising approximately 50 items – is presented below.

Item Infit and Outfit

The majority of the items in all tests fell within the acceptable fit range of 0.7-1.3, indicating good fit to the Rasch model. Misfit, where it occurred, was only in a small percentage of items, and not more than 5% (2-3 out of 50) items on any one test. Appendix 1 presents fuller details.

At A2, B1, C1 and C2 levels, all test item infit and outfit mean-square values were within the 0.7 and 1.3 range, indicating that the items performed well.

With A1, all infit and outfit mean-square values were within the 0.7 and 1.3 range, except for a marginally higher outfit figure on Test A1-T1, indicating a slight outlier effect.

With B2, all infit and outfit mean-square values were within the 0.7 and 1.3 range except for an outfit of 2.26 on Test B2-T2, and 2.01 on Test B2-T3 – although these relatively high outfits only occurred at the 90th percentile.

Reliability

Test reliability, for a 50-item test, is proposed to be 0.7 or above (Ebel, 1965). For an 80-item test, 0.8 or better is the projected figure, and it is this which is taken as the baseline in the current study. For the 62 tests reported on in the Coniam & Lampropoulou (2020) study, almost all test reliabilities – via the KR20 statistic – were above 0.8.

The equivalent of classical test measures of reliability in Rasch is person reliability (Anselmi et al., 2019); this benchmark statistic is currently reported for all 18 tests in the current study. As Appendix 1 illustrates, the target of 0.8 or better was achieved by externally-referencing all tests for all levels apart from one A1 test with a reliability of 0.75, and one A2 test with a reliability of 0.77.

Together, these sets of background statistics are illustrative of a set of robust, well-constructed tests. The picture of test robustness confirms that the externally-referenced anchoring is being conducted against a backdrop of reliable tests.

A fuller picture of the data is available in Appendices 1 and 2. Appendix 2a illustrates test C1 T1, for which the midpoint is 140 on the LID scale. As can be seen the item distribution is quite regular and bell-shaped. Appendix 2b, which illustrates test A1 T1 for which the midpoint is 60, is not quite as regular, being somewhat bimodal with a set of more demanding items towards the upper end of the scale. In general, however, as discussed below, the results reflect more the picture presented by the C1 than the A1 test.

Externally-referenced Anchoring Results

An analysis of the 18 tests from two perspectives is presented below. First, tables are presented with test means and measures that emerged after externally-referenced anchoring, in particular at the means recorded at the 25th, 50th and 75th percentiles. Second, graphs are presented which provide a more visual representation of the outcome of the externally-referenced anchoring (ERA).

Table 3a: Item distributions in A1 externally-referenced anchored IESOL tests

A1 ERA midpoint = 60	A1 T1	A1 T2	A1 T3
No. of items	52	52	50
Mean	60	60	61.3
Std. Deviation	37.68	24.48	24.45
Minimum	3.35	4.37	9.8
Maximum	145.22	108.68	119.41
25th percentile	32.1	38.83	50.43
50th percentile	53.05	63.93	62.35
75th percentile	71.51	76.21	75.4

The externally-referenced anchoring midpoint for A1 was 60. For Test A1 T1, the mean measure at the 50th percentile was 53.05, a third of a logit (i.e., 6 points) below the midpoint; for Test A1 T2 and T3, the mean measure at the 50th percentile was very close to the midpoint of 60. 70 is the top end of the A1 cut score; figures recorded at the 75th percentiles for all three tests were very close to this figure of 70. This confirms the fact that the majority of items at this level are at their appropriate level.

Table 3b: Item distributions in A2 externally-referenced anchored IESOL tests

A2 ERA midpoint = 80	A2 T1	A2 T2	A2 T3
No. of items	52	52	52
Mean	80	80	80
Std. Deviation	20.38	21.21	21.75
Minimum	36.12	26.35	37.6
Maximum	116.75	135.86	139.04
25th percentile	68.36	69.94	64.79
50th percentile	78.08	82.78	78.32
75th percentile	97.63	90.89	92.8

The externally-referenced anchoring midpoint for A2 was 80. At the 50th percentile, all three tests were very close to this figure with 90 as the top end of the A2 cut score; the 75th percentiles of A2 T2 and T3 had means very close to this figure; A2 T1 had some rather more demanding items, with a slightly higher mean measure of 97.63.

Table 3c: Item distributions in B1 externally-referenced anchored IESOL tests

B1 ERA midpoint = 100	B1 T1	B1 T2	B1 T3
No. of items	45	47	50
Mean	98.88	91.29	98.18
Std. Deviation	22.18	17.72	19.04
Minimum	49.88	44.45	61.75
Maximum	136.74	134.66	131.99
25th percentile	81.48	81.84	81.25
50th percentile	106.2	91.87	100.09
75th percentile	116.18	101.63	111.41

The externally-referenced anchoring midpoint for B1 was 100. Tests B1 T1 and T3 were very close to this figure at the 50th percentile; items in B1 T2 were slightly easier. With 110 as the top end of the B1 cut score, a similar picture emerged: B1 T1 and T3 were very close to the 75th percentile, while B1 T2 had items of slightly easier values at 101.63.

Table 3d: Item distributions in B2 externally-referenced anchored IESOL tests

B2 ERA midpoint = 120	B2 T1	B2 T2	B2 T3
No. of items	48	44	47
Mean	120	117.9	120
Std. Deviation	20.35	37.72	29.42
Minimum	66.43	61.34	71.77
Maximum	154.61	196.11	186.13
25th percentile	106.09	91.98	101.13
50th percentile	119.11	113.78	118.11
75th percentile	133.67	151.88	137.63

The externally-referenced anchoring midpoint for B2 was 120. Tests B2 T1 and T3 were very close to this figure at the 50th percentile; items in B2 T2 were slightly easier. With 130 as the top end of the B2 cut score, a similar picture emerged. At the 75th percentile, the B2 T1 and T3 mean measures were very close to this cut score, while B2 T2 had items which were rather more demanding 151.88.

Table 3e: Item distributions in C1 externally-referenced anchored IESOL tests

C1 ERA midpoint = 140	C1 T1	C1 T2	C1 T3
No. of items	51	52	52
Mean	139.06	140	140
Std. Deviation	16.2	17.41	19.09
Minimum	97	98.44	105.91
Maximum	170.8	188.08	194.25
25th percentile	128.7	128.24	122.89
50th percentile	141	141.59	140.78
75th percentile	149	149.55	149.44

The externally-referenced anchoring midpoint for C1 was 140. All three were almost exactly at this figure, showing an extremely close fit. Similar pictures were recorded at the 25th and 75th percentiles. With 150 being the top end of the C1 cut score, a very similar picture emerged, with all three tests having mean measures almost exactly at this figure.

Table 3f: Item distributions in C2 externally-referenced anchored IESOL tests

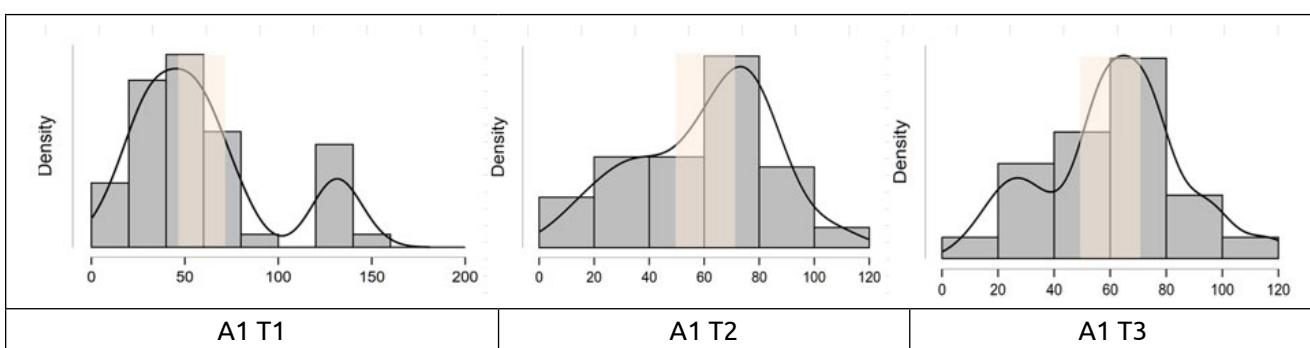
C2 ERA midpoint = 160	C2 T1	C2 T2	C2 T3
No. of items	50	50	50
Mean	158.22	158.59	158.43
Std. Deviation	18.71	14.68	16.04
Minimum	106.4	121.6	129.86
Maximum	197.56	186.2	196.44
25th percentile	143.51	147.5	146.75
50th percentile	160.73	157.6	158.64
75th percentile	172	171.2	169.13

Figures recoded for C2 were very similar to those returned for C1. With the externally-referenced anchoring midpoint for C2 being 160, all three C2 were almost exactly at this figure. Similar pictures were recorded at the 75th percentiles. With 170 the top end of the C2 cut score, a similar picture to C1 again emerged, with all three tests having mean measures almost exactly at the 170-point C2 top end cut score figure.

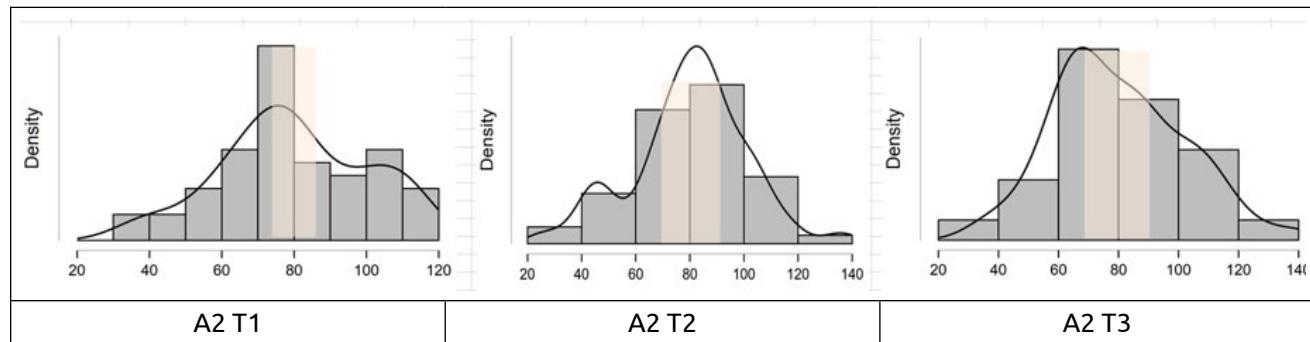
As a parallel view, and a reframing of the data presented in the tables above, the charts in Figure 1 below contain the results of anchoring as matched visually against the LID CEFR levels.

The grey bars and the trend graphs represent the IESOL item distributions; the shaded areas are the LID CEFR ranges. The density represents the frequency of items at a given LID scale range.

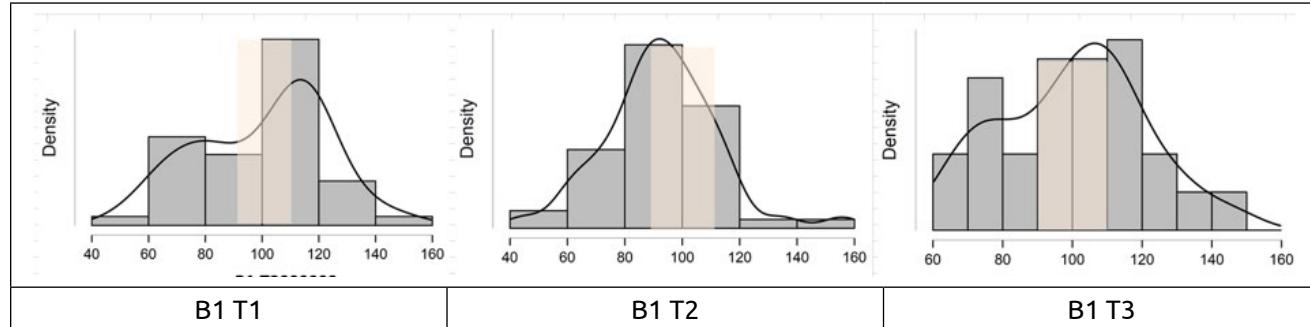
Figure 1a: Externally-referenced anchoring of A1 level tests



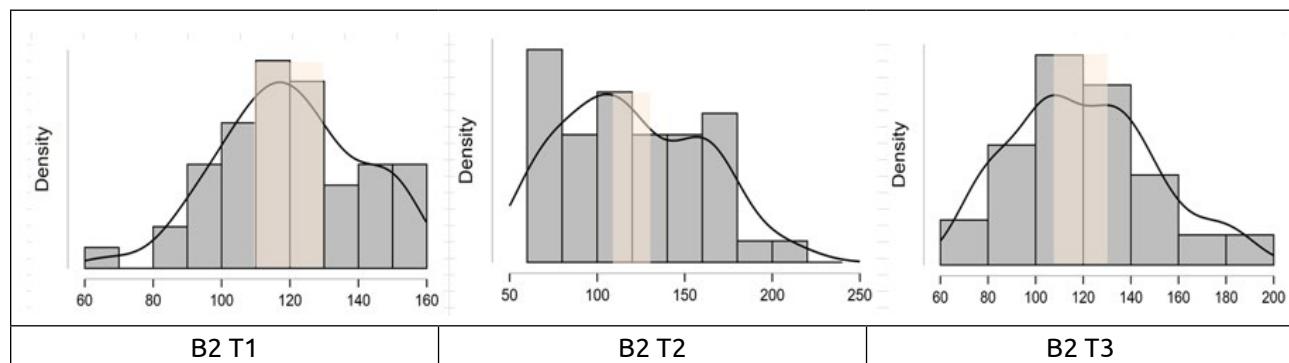
60 is the externally-referenced anchoring midpoint for A1, repeated by the orange shading. Tests A1 T2 and T3 show quite a normal distribution, in particular with test A1 T3. Test A1 T1 is less regular – being somewhat bimodal with a number of items which are more demanding than might be expected at A1 level.

Figure 1b: Externally-referenced anchoring of A2 level tests

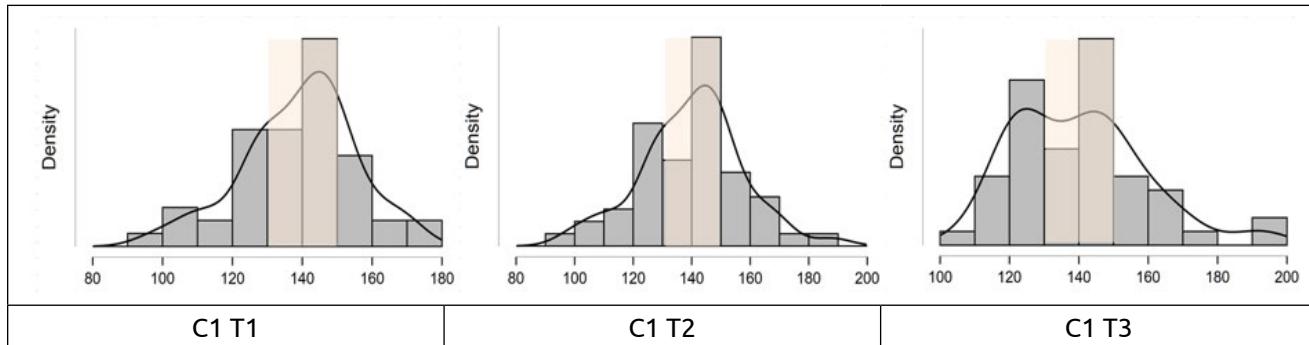
80 is the externally-referenced anchoring midpoint for A2. A2 T2 fits a normal distribution well, as does A2 T1 although this test has quite a large number of items exactly around the midpoint of the A2 scale.

Figure 1c: Externally-referenced anchoring of B1 level tests

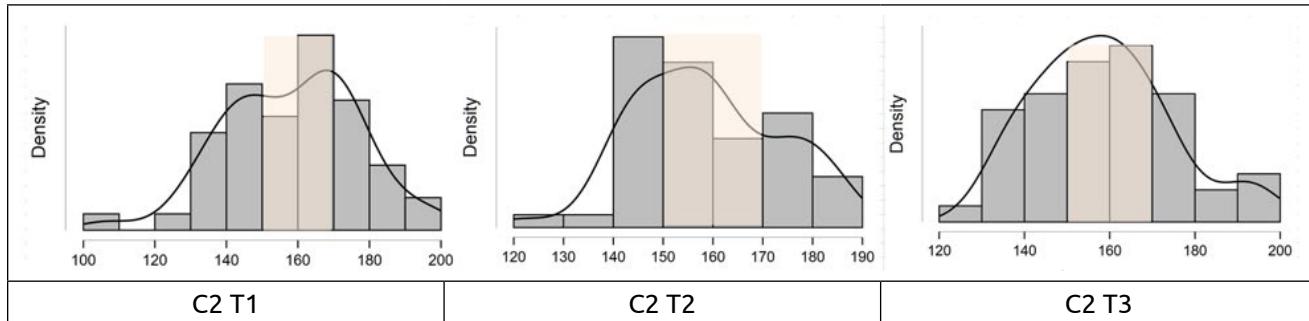
100 is the externally-referenced anchoring midpoint for B1, repeated by the orange shading. B1 T2 shows quite a normal distribution. The B1 T1 and T3 tests are slightly negatively skewed towards more demanding items.

Figure 1d: Externally-referenced anchoring of B2 level tests

120 is the externally-referenced anchoring midpoint for B2. B2 T1 and T3 show quite normal distributions; B2 T2 items are distributed in a slightly narrower range.

Figure 1e: Externally-referenced anchoring of C1 level tests

140 is the externally-referenced anchored midpoint for C1. All three tests show generally normal distributions.

Figure 1f: Externally-referenced anchoring of C2 level tests

160 is the externally-referenced anchoring midpoint for C2. All three tests again show generally normal distributions.

It can be seen that the LID CEFR zones in general occupy the centre of IESOL item distribution, with this distribution including a substantial number of the items in a given test. The expert-rated CEFR levels for the IESOL tests match well with the calibrated LID scale CEFR levels. The IESOL tests may therefore be considered to be acceptably well anchored onto the LID scale.

Conclusion

This paper has reported on the externally-referenced anchoring of LanguageCert IESOL tests against the LanguageCert LID scale CEFR levels. Calibrating tests onto a single scale generally involves using common items between different tests and cross-calibrating them using Rasch measurement. When there are no linking items available, other methods, however, need to be used. One of these, proposed by Linacre (2018), involves the use of items that measure the same trait, i.e., externally-referenced anchoring. In the current context, externally-referenced anchoring is illustrated through the lens of expert setters who have been producing quality items (see Coniam & Lampropoulou, 2020) at predefined and well-understood CEFR levels for many years.

Two related hypotheses regarding the validity of externally-referenced anchoring have been investigated.

The first hypothesis is that good Rasch infit and outfit statistics from the externally-referenced anchoring process are achieved. At each of the six CEFR levels, three different test forms were selected at random for analysis and good Rasch infit and outfit statistics are indeed found for each test. The first hypothesis is therefore confirmed.

The second hypothesis is that broadly bell-shaped item measure distributions would emerge from the analysis. All analyses generally recorded a good match between IESOL-assigned CEFR levels and the LID scale CEFR levels, with sets of items, for the most part, showing generally balanced distributions. The majority of items in almost all tests fell within the 25th to 75th percentiles: the points at which these percentiles broadly match the upper and lower end of the cut scores determined for a given CEFR level. Hypothesis two is also confirmed.

As indicated – in particular by Appendix 1b – not all matches between the items distributions and the LID scale are perfect; in general, however, a close match is reported, with the majority of items falling between the 25th and 75th percentiles – the lower and upper bounds of LID scale values for a given CEFR level. Consequently, while the results indicate that LanguageCert IESOL test items are generally appropriate for the respective CEFR level, the concept of externally-referenced anchoring as a methodology is also validated.

The match in the current study between externally-referenced anchored levels and LID scale CEFR levels reinforces the argument that LanguageCert IESOL tests have been well set, and statistically verifies the experts' judgements. The fact that the majority of the items fall within the 25th to 75th percentiles confirms that the items in the IESOL tests are well-targeted at the appropriate CEFR level by expert setters. The present study lends further support to the use of expert ratings in assessment.

While the externally-referenced anchoring outcomes obtained from the current study confirm the robustness of LanguageCert tests reported elsewhere (Coniam & Lampropoulou, 2020), only three tests were analysed at each CEFR level. Using externally-referenced anchoring principles, a study is currently underway to analyse a single dataset containing 15 tests at any given CEFR level. Results from this study will supplement the picture of the current study and will be reported on in due course.

Notes

1. The term "test" is used as generic form for an individual test or subtest. The term "test form" is used to indicate parallel, or multiple, versions of the same test.

References

- Alderson, C. J., Alderson, J. C., Clapham, C., Wall, D., & Swan, M. (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, 535–556.
- Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, 10, 2714.

- Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). An investigation into the comparability of two tests of English as a foreign language. Cambridge: Cambridge University Press.
- Boone, W. J., & Staver, J. R. (2020). Externally-referenced equating of test forms. In advances in Rasch analyses in the human sciences. Springer, Cham.
- Bristol, T. (2015). Test item writing: 3Cs for successful tests. *Teaching and Learning in Nursing*, 10(2), 100-103.
- Coniam, D., & Lampropoulou, L. (2020). A review of LanguageCert IESOL listening and reading test reliabilities 2018-2020. London: LanguageCert.
- Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System*, 37(2), 226-242.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). Validating the LanguageCert Test of English scale: The paper-based tests. London: LanguageCert.
- Coniam, David. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, 14(2), 5-22.
- Gable, R. K., & Wolf, M. B. (1993). Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings (2nd ed.). New York, NY: Springer Science & Business Media.
- Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing*, 28(1), 77–104.
- Goodman, L. (1990). Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. *Sociological Methodology*, 20, 249-294.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2(1), 37-50.
- Humphry, S. (2006). The impact of differential discrimination on vertical equating. *ARC report*.
- Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education*, 27(1), 1-18.
- Linacre, J. M. (2018). Winsteps Rasch measurement computer program user's guide. Beaverton, OR.
- Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, 10(3), 211–234.
- Luppescu, Stuart (1996). Externally-referenced equating: An approach to reading test equating by concept matching of items. Doctoral dissertation, University of Chicago.
- Luppescu, S. (2005). Externally-referenced equating. *Rasch Measurement Transactions*, 19(3), 1025.
- Rodriguez, M. C. (1997). The art & science of item writing: A meta-analysis of multiple-choice item format effects. In annual meeting of the American educational research association, Chicago, IL.
- Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, 25(4), 435–464.
- 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.
- VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. C. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT reasoning test. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies*. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix 1: Fit Statistics and Person Reliabilities

A1	A1-T1 Items	A1-T1 Infit	A1-T1 Outfit	A1-T2 Items	A1-T2 Infit	A1-T2 Outfit	A1-T3 Items	A1-T3 Infit	A1-T3 Outfit
Valid	52	52	52	52	52	52	52	52	52
Missing	0	0	0	0	0	0	0	0	0
Mean	60	0.96	1.25	60	1	0.88	57.85	0.99	0.96
Std. Deviation	37.68	0.11	0.7	24.48	0.15	0.39	29.97	0.13	0.29
Minimum	3.35	0.77	0.29	4.37	0.79	0.18	-51.68	0.75	0.36
Maximum	145.22	1.22	3.69	108.68	1.44	2.09	119.41	1.4	1.85
25th percentile	32.1	0.89	0.84	38.83	0.89	0.67	44.85	0.91	0.74
50th percentile	53.05	0.96	1.05	63.93	0.97	0.81	60.43	0.97	0.95
75th percentile	71.51	1.04	1.43	76.21	1.05	1.03	73.83	1.05	1.08
90th percentile	130.12	1.09	2.48	86.04	1.19	1.41	91.44	1.17	1.32
Reliability	0.84			0.75			0.82		

A2	A2-T1 items	A2-T1 Infit	A2-T1 Outfit	A2-T2 items	A2-T2 Infit	A2-T2 Outfit	A2-T3 items	A2-T3 Infit	A2-T3 Outfit
Valid	52	52	52	52	52	52	52	52	52
Missing	0	0	0	0	0	0	0	0	0
Mean	80	0.99	1.04	80	0.99	0.96	80	1	0.98
Std. Deviation	20.35	0.17	0.42	21.24	0.17	0.41	21.76	0.13	0.33
Minimum	36.17	0.72	0.37	26.38	0.72	0.41	37.59	0.72	0.4
Maximum	116.7	1.34	2.1	135.93	1.5	2.58	139.06	1.25	1.79
25th percentile	68.37	0.87	0.74	69.98	0.89	0.72	64.79	0.93	0.75
50th percentile	78.08	0.97	0.93	82.81	0.97	0.86	78.31	1.02	0.96
75th percentile	97.6	1.1	1.32	90.93	1.04	1.09	92.8	1.09	1.17
90th percentile	109.16	1.24	1.72	103.88	1.12	1.5	109.11	1.16	1.4
Reliability	0.88			0.88			0.77		

B1	B1-T1 items	B1-T1 Infit	B1-T1 Outfit	B1-T2 items	B1-T2 Infit	B1-T2 Outfit	B1-T3 items	B1-T3 Infit	B1-T3 Outfit
Valid	46	46	46	52	52	52	52	52	52
Missing	6	6	6	0	0	0	0	0	0
Mean	100	1	1.04	100	1	1.28	100	1	1.01
Std. Deviation	23.21	0.21	0.39	32.1	0.14	1.23	20.81	0.14	0.43
Minimum	49.93	0.67	0.57	44.45	0.72	0.5	61.73	0.74	0.51
Maximum	150.45	1.48	2.08	196.14	1.29	7.97	146.06	1.39	2.55
25th percentile	81.55	0.8	0.7	83.33	0.9	0.8	83.27	0.92	0.76
50th percentile	106.72	0.99	0.93	94.31	1.02	0.97	101.38	0.99	0.92
75th percentile	116.78	1.12	1.27	108.11	1.1	1.23	112.67	1.08	1.07
90th percentile	123.42	1.28	1.52	133.01	1.17	1.57	128.1	1.15	1.47
Reliability	0.88			0.85			0.84		

B2	B2-T1 items	B2-T1 Infit	B2-T1 Outfit	B2-T2 items	B2-T2 Infit	B2-T2 Outfit	B2-T3 items	B2-T3 Infit	B2-T3 Outfit
Valid	48	48	48	48	48	48	47	47	47
Missing	0	0	0	0	0	0	1	1	1
Mean	120	0.99	1	737.44	0.97	1.22	120	0.99	1.13
Std. Deviation	20.35	0.18	0.27	2416.94	0.24	1.63	29.42	0.17	0.7
Minimum	66.43	0.72	0.5	61.34	0.41	0.21	71.77	0.73	0.61
Maximum	154.61	1.5	1.67	9999	1.6	9.9	186.13	1.46	3.97
25th percentile	106.09	0.84	0.76	93.48	0.85	0.53	101.13	0.86	0.75
50th percentile	119.11	0.98	1	117.83	1	0.71	118.11	0.97	0.88
75th percentile	133.67	1.12	1.19	161.12	1.1	1.11	137.63	1.1	1.19
90th percentile	148.53	1.22	1.35	183.83	1.28	2.26	155.81	1.19	2.01
Reliability	0.86			0.88			0.87		

C1	C1-T1 items	C1-T1 Infit	C1-T1 Outfit	C1-T2 items	C1-T2 Infit	C1-T2 Outfit	C1-T3 items	C1-T3 Infit	C1-T3 Outfit
Valid	52	52	52	52	52	52	52	52	52
Missing	0	0	0	0	0	0	0	0	0
Mean	140	1	0.98	140	0.99	0.93	140	0.99	0.98
Std. Deviation	19.99	0.1	0.3	17.41	0.1	0.25	19.09	0.14	0.3
Minimum	80.93	0.82	0.26	98.44	0.83	0.43	105.91	0.78	0.64
Maximum	191.16	1.32	2.04	188.08	1.21	1.7	194.25	1.5	2.16
25th percentile	131.01	0.93	0.79	128.24	0.92	0.77	122.89	0.91	0.76
50th percentile	137.29	1	0.93	141.59	0.97	0.91	140.78	0.97	0.91
75th percentile	153.25	1.05	1.06	149.55	1.05	1.06	149.44	1.02	1.06
90th percentile	160.67	1.13	1.34	160.17	1.13	1.24	164.47	1.17	1.31
Reliability	0.85			0.81			0.88		

C2	C2-T1 items	C2-T1 Infit	C2-T1 Outfit	C2-T2 items	C2-T2 Infit	C2-T2 Outfit	C2-T3 items	C2-T3 Infit	C2-T3 Outfit
Valid	52	52	52	52	52	52	52	52	52
Missing	0	0	0	0	0	0	0	0	0
Mean	160	1	0.92	160	0.99	0.93	160	0.99	0.95
Std. Deviation	20.46	0.13	0.27	16.06	0.12	0.24	17.6	0.14	0.27
Minimum	106.4	0.79	0.47	121.56	0.8	0.32	129.86	0.76	0.51
Maximum	211.29	1.42	1.91	195.82	1.36	1.52	199.4	1.4	1.71
25th percentile	144.8	0.91	0.75	148.12	0.92	0.82	147.18	0.88	0.76
50th percentile	161.54	0.97	0.92	157.96	0.96	0.92	159.58	0.99	0.92
75th percentile	172.22	1.08	1.05	172.2	1.05	1.05	170.69	1.05	1.06
90th percentile	182.76	1.13	1.2	181.82	1.14	1.21	188.96	1.19	1.32
Reliability	0.83			0.81			0.84		

Appendix 2: Sample Outputs from IESOL Calibrations

Appendix 2a: C1 T1 (midpoint = 140)

PERSON	670 INPUT		670 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	39.4	53.7	172.55	8.83	1.00	.1	.98	.1
P.SD	8.8	2.4	24.41	3.58	.13	.7	.35	.7
REAL RMSE	9.53	TRUE SD	22.48	SEPARATION	2.36	PERSON RELIABILITY	.85	
ITEM	54 INPUT		52 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	507.3	666.6	140.00	2.30	1.00	.0	.98	-.1
P.SD	90.4	2.9	19.80	.65	.10	2.0	.30	2.3
REAL RMSE	2.39	TRUE SD	19.65	SEPARATION	8.24	ITEM RELIABILITY	.99	

MEASURE	PERSON	MAP	ITEM
<more> <are>			
230	.	+	
220		T+	
210		+	
200		+	S
190		C2.2.5	
		+ C1.4.7	
180		+T	
170		M	
		+ C2.2.4	
		C1.4.2	
		C2.3.5	
160		+S C1.3.5 C1.4.4 C2.4.2 C2.4.8	
		C2.4.7	
		C1.2.2 C1.3.6	
		C1.1.5 C2.3.4	
150		+ C1.2.6 C1.4.6	
		S C2.3.3	
		C1.1.4 C1.3.2 C2.3.6 C2.4.1	
		C1.3.4	
140			
		C1.1.1 C1.3.7 C2.2.1 C2.2.2	
		C1.1.6 C1.3.3 C2.3.7 C2.4.4	
130		C1.1.2 C1.1.3 C1.2.5 C1.4.5 C2.1.4 C2.1.5 C2.2.6	
		+ C1.2.3 C1.4.1 C2.1.3 C2.4.3	
		C2.1.1	
		C1.2.1	
120		T C2.2.3 C2.4.5	
		+S	
		C1.3.1	
		C1.2.4 C2.4.6	
		C1.4.3 C2.3.2	
110			
		+	
100		+T	
		C2.1.2	
90			
		+	
80		+ C2.3.1	
		<class> <freq>	

Appendix 2b: A1 T1 (midpoint = 60)

PERSON	S18 INPUT		S18 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	INHSQ	ZSTD	OHNSQ	ZSTD
MEAN	35.3	51.9	82.98	9.14	.98	-.2	1.24	-.1
P.SD	7.2	.5	23.48	2.63	.56	1.7	1.46	1.5
REAL RMSE	9.51	TRUE SD	21.46	SEPARATION	2.26	PERSON RELIABILITY	.84	

ITEM	S4 INPUT		S2 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	INHSQ	ZSTD	OHNSQ	ZSTD
MEAN	351.7	516.7	60.00	2.80	.96	-.2	1.25	1.2
P.SD	143.7	1.2	37.32	.65	.18	1.5	.69	3.1
REAL RMSE	2.88	TRUE SD	37.21	SEPARATION	12.93	ITEM RELIABILITY	.99	

```

MEASURE    PERSON - MAP - ITEM
          <more>|<rare>
159        .   +
               |
               #####
149        +   1.2.1
               #
139        +   1.1.4
               .   T 2.1.3
               .   | 1.1.5 1.1.6
129        T+  1.1.1 1.2.2
               .
               .   1.1.2
               |
               ##### 2.1.5
119        +   .
               .##|
               |
109        #####|+
               S|
               .#####
               .#####
99         #   +
               .#####|S
               .
               #####
89         #####|+
               #####|M| 1.3.3
79         .#   +
               #####| 1.2.4 1.3.1 2.3.4
               .#####
               #####| 2.2.2 2.3.2
69         .#####|+
               #####| 1.1.3
               #####| 1.2.3 1.2.5
59         ####|S+M 1.3.2 2.2.3
               .###| 1.2.6
               .##| 2.1.2 2.2.4 2.2.5 2.3.7 2.4.7
               ##| 2.1.1
49         #####|+ 1.3.5 2.4.8
               .##| 1.4.1
               .#| 1.3.6 1.4.4
               .| 2.3.6 2.4.6
39         .|+ 1.2.7 2.4.3
               .###|T| 2.3.1
               #| 1.4.3 2.3.5 2.4.5
29         .|+ 1.4.5 2.1.6
               .| 1.4.2 2.3.3 2.4.4
               .| 1.3.4
               .|S 2.1.4
19         +| 2.4.1 2.4.2
               | 2.2.1
               |
               +| 1.1.7
               |
               +| 1.4.6
0          +| 1.4.6
               |
               <less>|<freq>
EACH "#" IS 3: EACH "." IS 1 TO 2

```



SECTION 4: ORIGINAL RESEARCH



Chapter 10: Identifying Guessing in English Language Tests via Rasch Fit Statistics: An Exploratory Study

David Coniam, Tony Lee and Leda Lampropoulou

[Coniam, D., Lee, T., & Lampropoulou, L. (2021). *Identifying guessing in English language tests via Rasch fit statistics: An exploratory study*. *English Language Teaching*, 14(4).]

Abstract

This article explores the issue of identifying guessers – with a specific focus on multiple-choice tests. Guessing has long been considered a problem due to the fact that it compromises validity. A test taker scoring higher than they should through guessing does not provide a picture of their actual ability. After an initial description of issues associated with guessing, the article then outlines approaches which have been taken to either discourage test takers from guessing or which attempt, statistically, to handle the problem. From this, the article moves to a novel way of identifying potential guessers: from the post hoc use of Rasch fit statistics. Two datasets, each consisting of approximately 200 beginner level English language test takers were split into two. In each dataset, half the test takers' answers were randomised – to approximate guessing. Results obtained via a Rasch analysis of the data was then passed to an analyst who used the Rasch fit statistics to identify possible guessers. On each dataset, 80% of guessers were identified.

Keywords: guessing, Rasch, fit statistics, English language

Introduction

The key concept in assessment is generally considered to be *validity*. Validity (see, e.g., Messick, 1989; Bachman & Palmer, 2010) may be framed as constituting the extent to which a given test score can be interpreted as an indicator of the key abilities or constructs being measured. There are a number of issues which may be

construed as “construct irrelevant variance” (Downing, 2002): cheating, testwiseness, teaching to the test, flawed test design, to name but a few. One construct irrelevant variance (CIV) that many have long grappled with, especially in multiple-choice (MC) tests is that of guessing. Guessing essentially increases measurement error in that it raises the possibility of correct test taker responses, and hence compromises validity.

Two issues that are interconnected relate to what can be done in terms of: one, identifying guessing; two, dealing with guessing.

While the focus of the current paper is on the first issue, it will be necessary to put both issues into perspective to give a sense of direction to the current paper. In a seminal paper on the effect of random guessing on test validity, Lord, some 50 years ago commented on the reality that while test takers may be advised not to guess, guessing on the part of test takers cannot really be militated against.

Identifying Guessing

While a considerable amount of effort has been expended into dealing with possible guessing, the literature on identifying guessing is surprisingly thin. Lord and Novick (1968) described an item's discriminating power as its effectiveness in discriminating among higher and lower achievers, and stated that the correlation between an item score and the overall test score (i.e., an item-total correlation) provides “a rough index of item discriminating power” (p. 331).

In a somewhat similar vein, Downing (2002) examined the issue of construct-irrelevant variance from the perspective of what he termed “flawed test questions”. In his study, in which experts identified flawed items in a science test, the discrimination indices of the items and the overall KR20 reliability figure and the passing scores were lower for the flawed items than for well-constructed items. While Downing concluded with the observation that the use of flawed test questions may well have a negative impact on student performance, the study is a useful indicator of how statistics may have a role to play in identifying CIV issues, with guessing being one.

Boone & Staver, J.R. (2020) illustrate how the point measure correlation – a statistic provided by many Rasch analysis programs – may be used to identify possible errors in answer keys. While this may not be guessing per se, the construct-irrelevant variance that a flawed key brings to an analysis, trends in a similar direction.

The work of Attali and Bar-Hillel (2006) focuses on the fact that with many MC tests, the correct answers will tend to be in the “middle positions”, with test takers who guess then tending to select the middle options as their guesses.

A recent approach to identifying guessing – in the context of computer-based tests – has been that of response time – i.e., the time taken to record a response. The work of Wise (e.g., 2017, 2019) has illustrated how rapid-guessing (in computer-based tests) may be identified, and in part therefore dealt with. Certain studies have, however, found that the correctness of rapid guesses exhibits little relationship to overall test performance (Goldhammer et al., 2016).

Handling Guessing

Since guessing has long been an issue which has been seen to threaten test validity, methods of dealing with, or discouraging, guessing have been approached from a number of perspectives. One major approach which has morphed through various approaches – with its supporters and opponents – has been that of “formula scoring”. Formula scoring involves penalising test takers for incorrect answers, deducting marks, thus attempting to get test takers to only attempt questions they feel sure of answering correctly. However, certain researchers (for example, Lord, 1964) have recommended that all test takers answer all questions – in effect forcing an equal willingness to guess on all test takers.

MC tests have generally been scored using a conventional number-correct scoring method (Kurz, 1999; Bereby-Meyer et al., 2002) where a test taker gets the score for the number of questions they have answered correctly, irrespective of whether they have guessed.

Kurz (1999) presents an overview of some of the different scoring formulas used to correct for guessing. In this method of negative marking, test takers are penalised for incorrect responses.

Another method proposed by Traub et al. (1969) rewards a student for not guessing – by awarding points for omitted items rather than penalising for incorrect responses.

Since it might be expected that in a four-option MC item, test takers will get 25% correct and where; in a 3-option MC item, test takers will get 33% correct, one proposal has been that the passing score should be raised to reflect this effect (see e.g., Lesage et al., 2013).

In summary, however, as Lesage et al. (2013) point out, in the 50 years that guessing on MC tests has been researched, evidence as to which method is the most robust for dealing with guessing is still lacking.

Guessing has also long been a concern in English language tests, having been explored and discussed in a wide range of assessment situations and contexts – from vocabulary to reading; see e.g., Haynes, 1984; Huibregtse et al., 2002; Vanhove & Berthele, 2015; Gyllstad et al., 2015).

One statistical model for dealing with guessing is the three-parameter IRT model (see e.g., Birnbaum, 1968; Waller, 1989). Using this model, it is possible to make corrections on the basis of correct answers. As LeBeau & McVay (2017) point out, however, large sample sizes at times inhibit the use of the model: Hulin et al. (1982) suggest a minimum of 1,000 subjects for accurate measurement in the three-parameter logistic (3PL) model. While such samples may be accessible in large-scale tests, they will not be available to teachers in schools who may well be looking at class or, at most, school-size groups, and will be limited to a sample of a couple hundred students at most.

Background and Methods

Given that reducing guessing increases test validity (see Kurz, 1999), the question is therefore: To what extent can guessing, and in particular wild guessing, be identified?

The current study build on the use of statistics, with the working hypothesis centred around the concept of model fit. Given that the main statistical procedures used in this study involve Rasch, the reader is referred to the outline of the Rasch measurement model provided in the Glossary of statistical terms at the end of the volume.

Expected Values

The central concept in Rasch is that of the 'fit' of the data to the Rasch model; i.e., the extent to which obtained values match expected values. A 'perfect' fit of the data to the model may be interpreted from three perspectives.

- good point measure correlations
- outfit and infit mean squares of 1.0
- standardised Z-scores of 0.0

Such figures would indicate that obtained values exactly match expected values.

Mean square values less than 1.0 generally indicate that observations are too predictable. In contrast, mean squares values above 1.0 indicate over-dispersion, the possibility of pure guessing; over-fit indicates near uniformity in response, the possibility of giving invariant answers – which can be construed as another way of guessing. While acceptable ranges of tolerance for fit vary, acceptable ranges are generally taken as from 0.7 (30% below expectations) to 1.3 (30% above expectations) (see Linacre, n.d. (a)).

Regarding standardised Z-scores, figures above 2.0 indicate considerable distortion or degradation in the measurement system (Linacre, n.d. (b)).

Data and Analysis

The data in the current study was drawn from tests produced by the international language assessment organisation *LanguageCert*. *LanguageCert* produce and administer a suite of tests – the International ESOL suite – which are aligned to the six CEFR levels: Preliminary (A1), Access (A2), Achiever (B1), Communicator (B2), Expert (C1) and Mastery (C2). The examination specifications reflect the requirements of the CEFR; test materials writers employ the highest international standards and have extensive expertise in, and knowledge and understanding of, the CEFR [Note 1].

Two datasets were constructed, both with beginner-level test takers. This level of ability was selected because test takers at this level are beginners, have a more restricted grasp of English and therefore are possibly more

prone to guess. It should be noted that in the tests they were administered, there is no negative marking formula: no marks are deducted for incorrect answers.

The first dataset comprised a sample of 203 test takers who had been graded at A1 level in terms of their English language standard under the CEFR. The second comprised a sample of 287 test takers who been graded at A2 level.

The datasets were first split into two, with the two groups in each dataset identified by their infit and outfit mean squares values.

Approximately half of the test takers' scores were left untouched on the basis of good fit to the Rasch model of the '1' threshold: all test takers' infit and outfit scores were between 0.95 – 1.05; that is, that the mismatch between expected-to-observed scores was only minimally (5% either way) under- or over-estimated. This group was labelled '1' – Unchanged.

The responses for the other half of the test takers were then randomised – to simulate random guessing. This group was labelled '2' – Randomised.

On t-tests run with each set of tests, the two groups were reported to be of equal ability, with no significance reported in group mean scores.

The data was then run through the software program *Winsteps* (Linacre, 2020), following which the results were passed to an analyst who attempted to identify the guessers, using the infit and outfit mean square output.

Hypothesis

The hypothesis in the current study is that – on the basis of high outfit or infit, or high standardised Z-score statistics – 80% of test takers may be identified as guessers.

Results

An analyst was then given test takers' mean square outputs; they were not given access to test takers' scores – which would have given the analyst hints and hence invalidated the exercise.

The following guidelines were given to the analyst in terms of coming to a decision about guessing:

1. Investigate negative person point measure correlations.
2. Investigate outfit, then infit.
3. Investigate mean squares, then standardised Z-scores.
4. Investigate high values, then low or negative values.

The analyst, in their examination of the mean squares, was instructed to assign one of three labels to each test taker. These, as mentioned, were: '1' – Not a Guess, '2' – Definite Guess, '3' – Unsure.

Cohen's kappa was used to report on coder agreement. According to McHugh (2012), a level of 0.6 for kappa indicates 'moderate' and a level of 0.8 or better 'strong' agreement.

Descriptive analyses are presented below, separately for each test. These are then followed by the crosstabulation picture, along with the figure for kappa. Tables 1 and 2 first present the results for Test A1.

Test A1

Table 1 presents the labels assigned in the original Test 1A dataset, with Columns 2 and 5 showing the number of cases in the different categories.

Table 1: Category descriptives – Test A1 (N=203)

Original	Cases	Analyst	Cases
Unchanged	107 (52.7%)	Not a Guess	97 (47.8%)
Randomised	96 (47.3%)	Definite Guess	87 (42.9%)
		Unsure	19 (9.3%)

As can be seen, the analyst suggested that 97 (47.8%) of the dataset were not guessers and 87 (42.9%) definite guessers. They were unsure about 19 test takers.

Table 2 presents a crosstabulation of the two sets of data, with Kappa calculated for inter-category agreement. The key cell showing agreement are in bold font.

Table 2: Category crosstabs – Test A1

		ORIGINAL		Totals
		Unchanged	Randomised	
ANALYST	Not a Guess	97 (90.1%)	0 (0.0%)	97
	Definite Guess	3 (2.8%)	84 (87.5%)	87
	Unsure	7 (7.2%)	12 (12.5%)	19
Totals		107	96	203

In the dataset of 107 original unchanged answers, the analyst labelled 97/107 (90.1%) as 'non-guessers'. Of the 96 randomised answers, they labelled 84/96 (86.6%) as 'guessers'.

Kappa for inter-category agreement – that is, between the original categories and the analyst's verdicts – was 0.81 ($p<.000$) – 'strong' agreement in Hughes' (2012) terms.

Tables 3 and 4 present the results for Test A2.

The hypothesis in this case was, therefore, proven.

Test A2

Table 3: Category descriptives – Test A2 (N=287)

Original	Counts	Analyst	Counts
Unchanged	135 (47.0%)	Not a Guessers	147 (51.2%)
Randomised	152 (53.0%)	Definite Guessers	116 (40.4%)
		Unsure	24 (8.4%)

The analyst labelled 116 of the 287 subjects (40.4%) as definite guessers, and 147 (51.2%) as not guessers. They were unsure about 24 test takers.

Table 4 presents a crosstabulation of the two sets of data, with Kappa calculated for inter-category agreement.

Table 4: Category crosstabs – Test A2

		ORIGINAL		Totals
		Unchanged	Randomised	Total
ANALYST	Not a Guessers	115 (85.2%)	32 (25.6%)	147
	Definite Guessers	7 (5.2%)	109 (71.2%)	124
	Unsure	13 (9.6%)	11 (7.2%)	24
		Totals	135	152
				287

In the dataset of 135 original unchanged answers, the analyst labelled 115/135 (85.2%) as 'non-guessers'. Of the 152 randomised answers, they labelled 109/152 (71.2%) as 'guessers'.

Kappa for inter-category agreement – that is, between the original categories and the analyst's verdicts – was, for this test, lower – 0.59 ($p < .000$), i.e., 'moderate' agreement in Hughes' (2012) terms.

The hypothesis in this case was, therefore, not proven.

Discussion

This study was exploring how guessers might be identified via the use of fit statistics produced via one-parameter Rasch analysis. The hypothesis was that 80% of test takers who exhibited high infit and outfit statistics might be identified as guessers. This hypothesis was accepted for Test A1, where 87.5% of random guessers were identified. This hypothesis was not accepted for Test A2, however, where only 71.2% of random guessers were identified – and where the 80% target threshold was not achieved.

While the study is limited in its scope in that it only involved two groups of beginner-level English language test takers, the methodology does illustrate the potential for identifying guessers. The one-parameter Rasch model does not require a very large sample, unlike its three-parameter cousin. The methodology in the current study has limited itself to low-ability test takers who might be more inclined to guess than more able ones; this is an issue that will need to be explored in further studies with other higher ability groups. It may be the

case that the hypothesis was accepted for the absolute beginner test sample A1, because absolute beginners, having very little language to start with may be forced to guess more than learners with a higher proficiency of language, who may be making more ‘judicious’ attempts at items.

The value of the current study is in its relevance to small scale studies, where, for example, a teacher needs to identify guessers in a school exam to consider remedial work for students who may be experiencing difficulties. (The work of Ho et al. (2012) in using Rasch measurement with Hong Kong teachers is instructive here.) Another area of possible use is where teachers are pretesting exam material, and only want to include answers by ‘bona fide’ respondents. Being able to eliminate certain guessers would enable pretest results to be seen to have greater validity. This is the key contribution of the current study: that reducing the amount of guessing in a test makes for better validity in terms of how the scores may be interpreted.

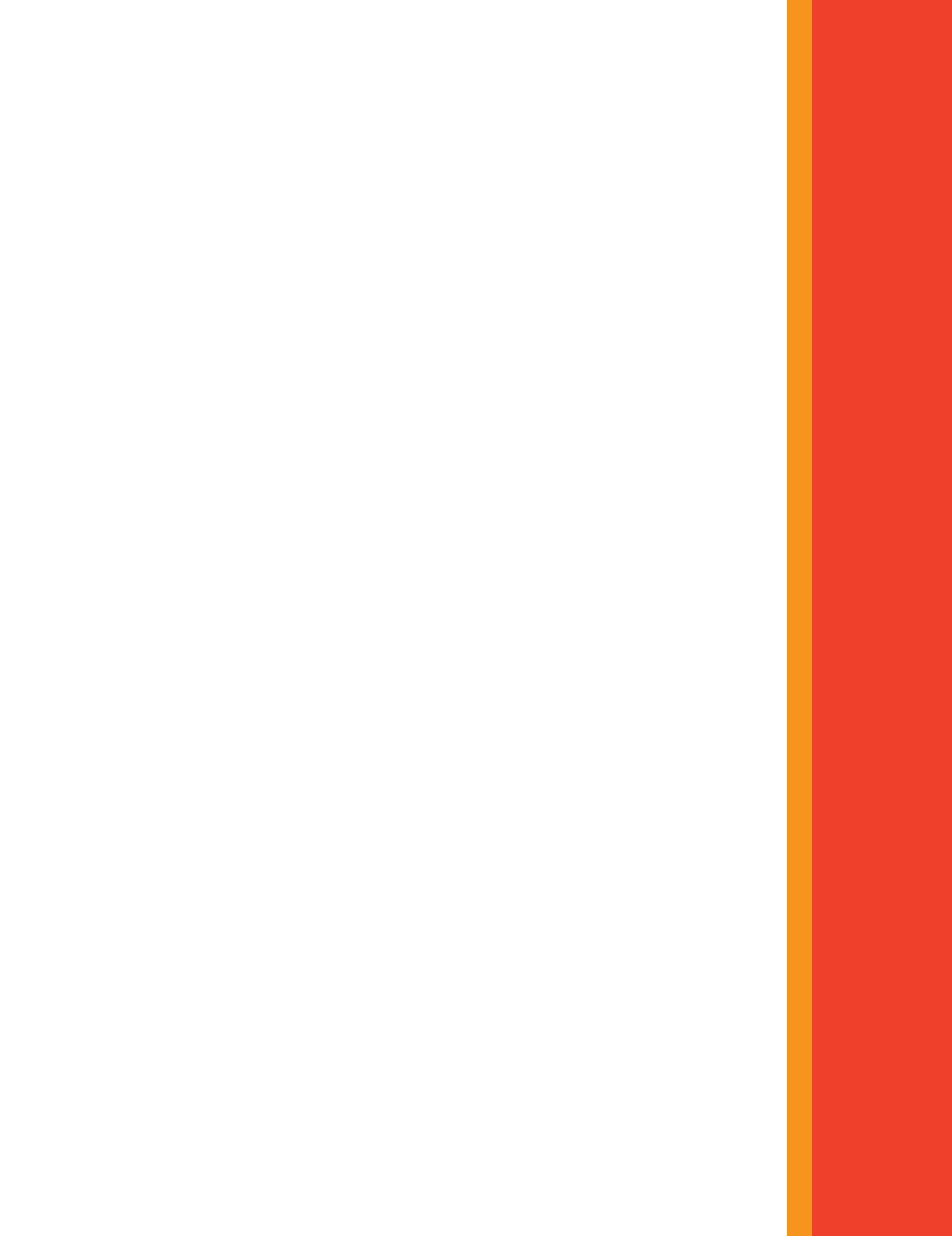
Notes

1. The CEFR has, over the past two decades, come to be accepted across Europe (and indeed beyond, with many countries linking their language curricula, syllabuses and examinations to the CEFR) as a specification of common standards across many different European languages. The CEFR lays out a set of common standards which permit employers and educational institutions to evaluate the language qualifications of test takers applying for employment or admission to education.

References

- Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, 40(2), 109-128.
- Bachman, L. & Palmer, A. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.
- Bereby-Meyer, Y., Meyer, Y. & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, 15: 313–327.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J.: Erlbaum.
- Boone, W.J., & Staver, J.R. (2020). *Advances in Rasch Analyses in the Human Sciences*. Springer: Cham, Switzerland.
- Downing, S. M. (2002). Construct-irrelevant Variance and Flawed Test Questions, *Academic Medicine*, 77(10), 103-104.
- Fleiss, J., (1981). *Statistical Methods for Rates and Proportions*. Wiley VCH, New York.
- Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics*, 166(2), 278-306.
- Haynes, M. (1984). Patterns and perils of guessing in second language reading. *on TESOL*, 83, 163-176.
- Ho, C.M., Leung, A.W.C., Mok, M.M.C., & Cheung P.T.M. (2012). Informing Learning and Teaching Using Feedback from Assessment Data: Hong Kong Teachers’ Attitudes Towards Rasch Measurement. In: Mok M. (eds) *Self-directed Learning Oriented Assessments in the Asia-Pacific*. Education in the Asia-Pacific Region: Issues, Concerns and Prospects, vol 18. Springer, Dordrecht. http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-94-007-4507-0_17

- Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language testing*, 19(3), 227-245.
- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: a monte carlo study. *Applied Psychological Measurement*, 6 (3), 249–260.
- Jiang, S., Wang, C., & Weiss, D.J. (2016). Sample Size Requirements for Estimation of Item Parameters in the Multidimensional Graded Response Model. *Frontiers in Psychology* 7.
- Kurz, T. B. (1999). A review of scoring algorithms for multiple-choice tests. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.
- LeBeau, B., & McVay, A. (2017). Validity of the three parameter logistic item response theory model for field test data. ITP Research Series: University of Iowa. <https://itp.education.uiowa.edu/ia/documents/Validity-of-the-Three-Parameter-Item-Response-Theory-Model-for-Field-Test.pdf>
- Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking?. *Studies in Educational Evaluation*, 39(3), 188-193.
- Linacre, J. M. (2020). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com
- Linacre, J. M. (n.d. (a)). Rasch power analysis: Size vs. significance: Infit and Outfit mean-square and standardized Chi-Square fit statistics. <https://www.rasch.org/rmt/rmt171n.htm>.
- Linacre, J. M. (n.d. (b)). Reasonable mean-square fit values. <https://www.rasch.org/rmt/rmt83b.htm>.
- Lord, F. M. (1964). The effect of random guessing on test validity. *Educational and psychological measurement*, XXIV(4), 745-747.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276-282.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Rogers, H. J. (1999). Guessing in multiple choice tests. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 235-243). Amsterdam: Pergamon.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of response to test items. *Applied Psychological Measurement*, 27(3), 159-203.
- Vanhouve, J., & Berthele, R. (2015). The lifespan development of cognate guessing skills in an unknown related language. *International Review of Applied Linguistics in Language Teaching*, 53(1), 1-38.
- Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement*, 13(3), 233–243. h
- Wise, S.L. (2019) An Information-Based Approach to Identifying Rapid-Guessing Thresholds, *Applied Measurement in Education*, 32(4), 325-336.



Chapter 11: The Development and Delivery of Online-Proctored Speaking Exams: The Case of LanguageCert International ESOL

Leda Lampropoulou and Yiannis Papargyris

[*Papargyris, Y. and Lampropoulou, L. (2021). The development and delivery of online-proctored speaking exams: The case of LanguageCert International ESOL, in ALTE (2021). Safeguarding the future of multilingual assessment in the post-covid world – Proceedings of the ALTE 7th International Conference, 22-24 April 2020, Madrid.*]]

Abstract

Responding to commercial requests for the online delivery of its exams, LanguageCert embarked on an attempt to develop an online-proctored (OLP) equivalent to its established International English for Speakers of Other Languages (IESOL) Speaking exam suite. To this extent, we considered the practical aspects of replicating the exam in an online environment, the potential need to adjust the content and format of the test, the applicable variants of the exam registration and exam administration processes, security and test integrity issues, as well as any potential impact on the assessment methodology the exam employs. Most importantly, we needed – and still do – to ensure that the accuracy of assessment outcomes is not compromised by the mode of delivery in any way.

Background

In mid-2018, LanguageCert began developing an online equivalent of the face-to-face International English for Speakers of Other Languages (IESOL) Speaking exam. Several factors had led to this decision, the implementation of which eventually coincided with the severe impact the COVID-19 pandemic had on paper-based exam delivery globally; among others (Marshall, Shannon, & Love, 2020), the growing dominance of online

communication and the popularity of online learning (Lim & Wang, 2016), both of which increasingly dictate the need for online assessment solutions. Furthermore, LanguageCert, a member of the PeopleCert Group of Companies, already had access to an electronic delivery system, through which PeopleCert has successfully and securely delivered over 390,000 exams in an online-proctored environment, mainly in the fields of Business and IT. The online-proctoring platform, offered by PeopleCert, is manageable in terms of technical competence and equipment required, and familiar and straightforward in terms of navigation and candidate expectations. It is worth noting that LanguageCert had also been delivering its IESOL (Listening, Reading, Writing) exam suite through PeopleCert's proprietary online-proctoring system since the beginning of 2017. On that occasion, it was only a system check and a test integrity check that were required for the effective implementation of the computer-based exam delivery, using the same, already tested and successfully used, online-proctored environment. The IESOL (Speaking) exam suite, though, was a different case, as it both relied on and effectively assessed live human interaction.

Objective

The main objective of this paper is to outline our rationale for the development of the online version of the IESOL Speaking exam, to identify the potential implications for the comparability between the different modes of delivery, and to explain how these were addressed in the context of LanguageCert's test development processes.

The LanguageCert IESOL Speaking Exam

The LanguageCert IESOL Speaking exam suite has been available since 2015 in a paper-based format. Each test consists of a short, spoken interview between one candidate and an interlocutor, who manages the interaction and is responsible for recording the session, but does not assess the candidate. Test format is consistent across all CEFR levels, comprising four parts: personal information, situational roleplay, interactive task, and long turn [Note 1]. Prescribed exam duration varies depending on the level of the exam from six (A1) to 17 (C2) minutes.

The Transition from Face-to-face to Online

The majority of studies conducted during the 1980s and 1990s, when computer-based testing was gradually introduced, supported the view that the computer-based testing process was comparable to paper-administered exams in terms of candidate experience, construct validity, scoring, etc. (Burke & Normand, 1987; Vincino & Moreno, 1988; Levin & Gordon, 1989; Taylor, Kirsch, Jamieson, & Eignor, 1999; Wise, Boettcher Barnes, Harvey, & Plake, 1989). LanguageCert's experience of the delivery of the online-proctored IESOL (Listening, Reading, Writing) exam echoes findings reported in research such as the above.

The study reported in this paper extends research into online-proctored delivery of tests in that the focus is on the skill of Speaking and on the implications which online communication may bring to the test.

It is important to point out that all of LanguageCert's claims to comparability presented below are based on the decision that both exam versions – face to face and online – are delivered by a human interlocutor; the

automated delivery of the spoken exam would pose an undisputed limitation to the capturing of any measurable extent of interaction (Galaczi, 2010).

In preparation for the transition from a face-to-face to an online environment, we needed to ensure that the mode of delivery did not exert an impact on the existing assessment arrangements. To this extent, a team of experts reviewed all assessment descriptors and criteria for references to paralinguistic elements and/or other features or interactional competences which might presuppose or rely upon face-to-face communication (e.g., eye contact, body language, gesture, etc.). We eventually concluded that the criteria used in the LanguageCert scales (i.e., Task Fulfilment and Coherence, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, Pronunciation, Intonation and Fluency) could be applied equally effectively – regardless of the mode of delivery.

In terms of exam content, a review of LanguageCert's Item Bank was conducted, to identify items which might be adversely affected by the mode of delivery. With that in mind, a panel of experts reviewed references to places (e.g., *We are sitting in a restaurant and you want to order a drink.*) and references to actions which presuppose proximity (e.g., *Can you pass me this magazine, please?*). In the same sense, we decided to amend certain instances of the latter, especially certain actions reflecting the practicality of the exam – i.e., instructions to the interlocutor – which may include passing on paper and pencil for notetaking during certain parts of the test. In terms of exam content, very few and minor amendments were undertaken specifically for the online delivery of certain items which denoted proximity between interlocutor and candidate (e.g., *How did you get here?* became *How did you get there?*).

Following the initial stages of the qualification review with the objective of administering it online, we proceeded with a pretesting phase, which would give us the opportunity to investigate a series of issues pertaining to the comparability of the two delivery streams. More specifically, we needed to ensure the comparability of assessment outcomes and to collect feedback from candidates, exam personnel and stakeholders as to the suitability of the proposed solution. The pilot began in mid-March 2019 and lasted two months. Approximately 180 spoken interviews were conducted at CEFR Levels B1–C2. We decided to exclude lower levels (A1, A2) from online delivery, as we felt that the English-medium onboarding process (a series of brief exchanges between the interlocutor and the prospective candidate to ensure that test takers have set up their systems appropriately and are ready to start the exam, e.g. 'Please show me your desk with your camera') would possibly disadvantage candidates at those levels. In brief, the pilot objectives were the following: i) to ensure error-free test administration (systems check); ii) to identify any potential linguistic limitations during the onboarding process for candidates at B1; iii) to identify any irregular trends in terms of assessment or any indication of irreconcilable lack of comparability.

It was anticipated that online delivery would affect interlocutor conduct in one respect, namely the management of potential sound delays and/or minor connection issues during the interview. As opposed to cases where actual technical support is needed, we are referring here to very minor connection and/or bandwidth issues which do not necessarily interrupt the interview but may, nonetheless, cause a momentary disconnection or delay. On such occasions, interlocutors were advised to actively encourage candidates to ask for repetition, whereas accommodating short delays, pauses and break-ups became part of standard interlocutor training. For instance, to accommodate minor technical issues, interlocutors were advised to repeat a question as many times as requested if there was a break-up in the signal and the candidate requested repetition. In contrast,

for a similar request during a face-to-face interview, the interlocutor would be advised to change the question instead of repeating it a third time. Backchanneling is also a technique which interlocutors have been trained to perform differently. In Part Three of the test, for example, where a dialogue takes place between interlocutor and candidate, the former would be instructed to encourage the candidate with backchanneling sounds, such us ‘um-hum’. It was noticed that such interjections might, however, reach the candidate a bit later than intended, or cause them to think that the interlocutor might want to start speaking, and interlocutors are no longer encouraged to interact in such a way, but to prioritise clearer turn-taking instead.

Observations and Next Steps

Upon completion of the pretest analysis, we found convincing evidence that scores were comparable between the face-to-face and the online delivery of the exam. This came as the result of analysing the average scores from candidate performances on the online pretest and comparing them to the averages of candidates participating in the paper-based speaking exams at the respective level. As a next step, the comparability study aims to investigate the effect of the mode of delivery on score accuracy and on interrater reliability.

In parallel to the – ongoing – quantitative analysis of candidate performance, we also conducted a qualitative study in order to identify candidate reactions to the online test-taking experience. This consisted of a candidate survey and one focus group consisting of assessment experts. Feedback was overwhelmingly positive both from candidates and stakeholders, who highlighted aspects of practicality and accessibility. A comment which stood out and seemed to be widely shared among survey participants is that the online interface bestows a ‘more democratic’ power relationship between the candidate and the interlocutor. Another described the online exchange between candidate and interlocutor as ‘less intimidating than actually visiting a test centre’.

Candidates – especially those at B1 level – were able to follow the onboarding instructions without any noticeable difficulty. 88% of candidates described the experience as good (51.28%) or excellent (36.90%).

No higher levels of malpractice or cases of compromised exam delivery have been detected since the online delivery of Speaking exams commenced, either. Unsurprisingly, though, as online testing remains something of a predominantly solitary activity for now, malpractice activities seem to revolve around individual test takers rather than centres, making the occurrence of malpractice qualitatively different to centre-based malpractice that affects several test-takers during a single administration (Draaijer, Jefferies, & Somers, 2017).

Overall, the digitisation and online delivery of existing assessments contribute to the modernisation of the assessment landscape and to its being in step with the developments in communication, technology and learning. It is imperative, nonetheless, that assessment development and certification principles be adhered to, and that the electronic delivery of traditional assessments is comparable and consistent with standard delivery methodologies.

Notes

1. In terms of turn-taking, a 'short turn' may be seen to be a single short answer. A 'long turn' is when a candidate provides a more extensive answer to a prompt (see Kasper & Youn, 2018).

References

- Burke, M. & Normand, J., (1987). Computerized Psychological Testing: Overview and Critique. *Professional Psychology: Research and Practice*. 18. 42-51. 10.1037/0735-7028.18.1.42.
- Draaijer, S., Jefferies, A., & Somers, G. (2017). Online proctoring for remote examination: a state of play in higher education in the EU. In International Conference on Technology Enhanced Assessment (pp. 96-108). Springer, Cham.
- Galaczi, E. D., (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities, in Araújo, L (Ed.) *Proceedings of the Computer-based Assessment (CBA) of Foreign Language Speaking Skills*, 29-51.
- Levin, T. & Gordon, C., (1989). Effect of Gender and Computer Experience on Attitudes toward Computers. *Journal of Educational Computing Research*, 5(1), 69-88. Retrieved January 19, 2021 from <https://www.learntechlib.org/p/141027/>.
- Kasper, G., & Youn, S. J. (2018). Transforming instruction to activity: Roleplay in language assessment. *Applied Linguistics Review*, 9(4), 589-616.
- Lim, C. P., & Wang, L. (Eds.). (2016). Blended learning for quality higher education: Selected case studies on implementation from Asia-Pacific. Bangkok: UNESCO Bangkok Office.
- Marshall, D. T., Shannon, D. M., & Love, S. M. (2020). How teachers experienced the COVID-19 transition to remote instruction. *Phi Delta Kappan*, 102(3), 46-50.
- Taylor, C., Kirsch, I., Jamieson, J. & Eignor, D., (1999). Examining the Relationship Between Computer Familiarity and Performance on Computer-Based Language Tasks. *Language Learning*.
- Vincino, F. & Moreno, K., (1988). Test-taker's attitudes toward and acceptance of a computerised adaptive test. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, USA.
- Wise, S. L., Boettcher Barnes, L., Harvey A. L. & Plake, B.S., (1989). Effects of Computer Anxiety and Computer Experience on the Computer-Based Achievement Test Performance of College Students, *Applied Measurement in Education*, 2:3, 235-241, DOI: 10.1207/s15324818ame0203_4



Chapter 12: Online Proctoring of High-stakes Examinations: A Survey of Past Candidates' Attitudes and Perceptions

David Coniam, Leda Lampropoulou and Angeliki Cheilaris

Abstract

This paper reports reactions by candidates to the use of online proctoring (OLP), 'invigilation', in the delivery of high-stakes English language examinations.

The paper first sets the scene in terms of the move from face to face to online modes of delivery. It explores the challenges and benefits that both modes offer, in terms of accessibility, fairness, security and cheating.

Evidence is then presented from a survey exploring the reactions to and perceptions of OLP by candidates who had taken an English language examination via OLP. A strong endorsement of OLP was generally recorded. Feedback revealed that respondents perceived OLP to be a more personal as well as a more efficient way of taking a test. Some pertinent negative comments from a smaller number of respondents could be construed as constructive and are also discussed. The results are indicative of a broad acceptance of OLP, pointing to strong future uptake of the OLP mode of test delivery.

Background

Online Delivery of Learning and Teaching

The mode for 'delivery' of both teaching and assessment has long been accepted as the teacher standing at the front of a class of students, to whom she provides input / facilitates (see Wiesenbergs & Stacey, 2008). In line with developments in technology and its uptake and acceptance across all facets of society (Lim & Wang, 2016) views of how education may be delivered are changing, spurred on considerably by the 2020 covid-19

pandemic (TPD@Scale for the Global South, 2020). The traditional mode of delivery is consequently reducing as the use of more innovative and interactional methods grows.

The acceptance of online learning and teaching is possibly clearest in the case of the development of MOOCs (Massive Online Open Courses) over the past decade – see e.g., Bonk et al. (2015).

Todd (2020) presents a cogent discussion of how covid-19 has forced teachers to consider – and to immediately handle – online teaching.

Over the past decade in particular, developments in technology have spurred the move to a ‘blended’ mode in terms of learning and teaching (see Lim & Wang, 2016). Many countries have moved to embrace, or are at least experimenting with, different forms of synchronous and asynchronous modes of teaching (Lim & Graham, 2021).

While the world is perhaps undergoing a degree of what Williamson (2015) terms ‘datafication’, the “objective quantification of all kinds of human behaviour and sociality”, this new paradigm in education is opening doors to many different sectors of society, as the work of Lim & Graham (in press) illustrates.

Nonetheless, while greater use of technology, and blended learning in particular, has brought about a change in mindset in terms of the delivery of teaching content using online instruction and facilitation, the delivery of examinations is still broadly expected to occur in a face-to-face situation (see Ahlawat et al., 2014).

Assessment – and in particular high-stakes assessment – has long been accepted as being conducted in a pen-and-paper medium, in front of an examiner in a centre such as a school or university hall. Following completion of the test, the test paper is then submitted to the marker(s), and the result emerges often after a considerable period of time. There has been some take-up of technology in the area of assessment, although less than with teaching. A brief examination of other test administration permutations now afforded by technology will now be presented. Figure 1 elaborates different potential test administration permutations.

Figure 1: Test administration permutations

Test taker location	Invigilator location	Invigilation conducted
at a centre	in person at a centre	by examiner
at home	remotely (via video link)	by invigilator
at home	–	by self

Alternative modes of test administration – in addition to the traditional – involve a test taker taking a pen-and-paper (or a computer-based) test at their own home under the invigilation of an examiner in another location (via video link). Another permutation involves a test taker being administered an oral test remotely by an examiner.

A final permutation is that of a test taker taking an exam under their own invigilation – as in the form of take-home exams (Bengtsson, 2019).

Hussein (2020) comments that while typical learning and teaching may be conducted quite competently by current online learning technologies, the conducting of assessment is fraught with challenges and problems. One major issue centres around the ability of teachers – and institutions – to ensure academic integrity when examinations are taken remotely, and from a private location such as the test taker's home.

Online delivery of assessment. There are a number of issues to be examined with regard to the online delivery of assessment; some of which may be seen as positive while others are viewed as negative.

The covid-19 pandemic has forced a major rethink of how education is delivered, with many educational institutions moving extremely quickly to partial if not total online delivery of classes – see Gardner (2020). Khan and Jawaid (2020) discuss the issue of technology enhanced assessment (TEA) during the covid-19 pandemic in Pakistan. They conclude with the observation that it is not now possible to "shy away from online teaching, learning and assessment" – the key comment here being assessment: assessment needs a deep rethink, but it has faced much less of a sea change than has learning and teaching. A major part of this issue involves strong concerns about security.

Clark et al. (2020) comment on the 'fit' of instructional practices within the course in terms of online vs face to face teaching and assessment. They comment on the issue of 'continuity' stating that where a course is intended at the outset to be a distance learning one, then online assessment should naturally fall into place. During covid, while teachers managed to adapt to an extent to online teaching, assessment was still a bolt-on. The mismatch between intended course outcomes and assessment conducted online tended to be greater than paper-based assessment which was more aligned with intended course outcomes. Clark et al. (2020) suggest that if the intended course outcome-assessment gap is to be narrowed, classes need begin with a distance learning format. In this way, students will become acclimatised to such an environment and be prepared for taking online exams, and will more clearly see the fit between online assessment and course content.

In Ardid et al.'s (2014) study of groups of students given online non-proctored exams, the participants scored higher than those who were given online-proctored exams. In the context of non-proctored assessments, two key issues then come to the fore (which will be discussed below) –security and honesty.

Alexander, et al. (2001) found no significant difference in student performance on proctored online exams and proctored paper and pencil exams in a computer technology course.

Computer adaptive testing has to an extent begin to change the face of assessment – in part because of the adaptive nature of the assessment which purports to offer test takers test material at their appropriate level rather than have test takers work though a linear test, where a considerable amount of the material will, quite possibly, not be at their level. (Thompson, 2017)

Advantages

On the positive side, test takers may take a test in the comfort (and safety) of their own home – an important factor in times of a pandemic where movement are restricted or, particularly, for a person who is disabled in some manner. Convenience and speed are two other factors to be considered; a test may be delivered via computer, and results may therefore be obtained in a more speedy manner.

Finally, as stated earlier, a majority of assessments –high-stakes school and university tests for example – involve test takers sitting in a hall and writing by hand for two hours. Given that the majority of assignments which test takers will have written over the course of an academic year will have involved multiple drafts via a computer word processor, it well may be argued that the traditional mode of administering exams actually compromises validity because the realities of traditional examination conditions do not reflect real life (and are thus, in a sense, invalid (Mogey et al. 2012). The possibility of completing a test via the word processor on a locked-down computer actually offers a test taker a more valid mode in which to complete an examination.

Downsides

One possible downside concerns expectations of teaching outcomes vs. expectations of assessment outcomes. Online teaching strongly stresses collaborative principles, such as discussion, peer support, learning tailored to individuals, self-regulated learning, getting students to set their own goals, plan, monitor and control their cognition (Boekaerts & Corno, 2005). In contrast, expectations of assessment (and in particular high-stakes assessment) are that the result will be generated by one test taker, working on their own, with no recourse to any form of external support. For comparability's sake, this generally means that the same test needs to be delivered to the same group of individuals at the same time. This requirement therefore involves security, honesty, fairness and reliability issues, which leads to considerations of test takers gaining an 'advantage', i.e., cheating.

Security

There has been considerable discussion in the literature about levels of security for different types of online examinations. Foster (2013) presents a cogent overview of security in online proctoring, which is useful as a lens through which high-stakes assessment can be viewed. Foster (*ibid*) defines online proctoring as emphasising "the critical use of the Internet and automated processes to produce a secure solution in monitoring test takers." Remote proctoring, he generally defines as the human-invigilation of examination, often in lower-stakes situations.

Foster presents a comprehensive list of 18 key security features, ranging from the management and training of the proctor, to interaction with the test taker, to the stability of the internet, to data transfer encryption. The list is laid out in Figure 2.

Figure 2. Key Security Features in OLP examinations (after Foster, 2013)

Key Feature	Subsumed under each feature
1. Online proctor during exam	
2. Continuous internet	
3. Encryption for data transfer	
4. Schedule availability	
5. Proctor management	Supervised / training / career path /certification
6. Interaction with test taker	Live chat / canned messages / live instruction to examinee / proctor views examinee screen / proctor as collusion threat
7. Prevent proctor view of screen	
8. Later video review proctoring	
9. Later video review capable	
10. Control during test session	Test launch / pause test / suspend test / cancel test
11. Automated proctoring	Inappropriate keystroke / audio levels / real-time data forensics
12. Lockdown	
13. Authentication	
14. Webcam	
15. Logs/records	Video storage / session review / time-stamped incident / incident logs
16. Program customization	Levels of security decisions / allowed/specify aids
17. Effectiveness research	

Foster describes how systems can provide levels of security which make online proctoring of examinations viable. He comments on the disadvantages that may be associated with traditional proctoring, where the proctors may be corrupt or may want to influence candidates scores in some way. Indeed, a number of studies report how exam security may be stronger as a result of the technologies associated with monitoring of online examination than in traditional face-to-face settings (Rose, 2009; Watson & Sottile, 2010).

LanguageCert has a rigorous set of regulatory principles for online-delivered and online proctored examinations and assessment applicable to test security, format and personnel involved. These adhere closely to the guidelines and recommendation laid down by the UK's Qualifications and Curriculum Authority (2007), and predate Foster's (2013) set of security features provided in Figure 1 above. To exemplify, upon first logging on, candidates need to follow a thorough 'onboarding' process; this includes an ID check, locking down their computer, checking there are no second monitors, and a room check through their webcam to show that the room is secure and that no other person or aids are present. The behaviour of candidates during the examination is then monitored in a number of ways: via qualified, specially trained proctors, and the use of video and audio recording, as well as advanced video and audio analytics and surveillance software.

Cheating and Academic Dishonesty

Cheating in exams is not a new phenomenon. Before the advent of the digital age and much easier access to the internet and plagiarism, comments about candidates cheating in examinations were not new – see e.g., Wright & Kelly, 1974; Bushway & Nash, 1977; Sierles et al., 1980. The Carnegie Council Report (1979) some forty years ago made reference to a growing “ethical deterioration” in academic life in terms of the number of college students cheating to get their desired grades.

The internet, access to digital documents and networks of people willing to facilitate paid cheating has certainly brought issues of cheating more to the fore over the past decade (Harper et al., 2020). Cheating in online examinations is becoming more salient, and has been explored in numerous studies, (Harmon & Lambrianos, 2008; Grijalva et al., 2006; Watson & Sottile, 2010).

Corrigan-Gibbs et al., (2015), provide an extensive debate on the extent of cheating and academic dishonesty. They discuss the vulnerability of online tests as being a salient concern. Much of the research in this area focuses on ways to make an online exam secure and discourage or prevent cheating, and how some test-takers attempt to undermine these efforts.

Reference was made above to the work of Rose, 2009 and Watson & Sottile, 2010. These researchers suggest that compared with traditional face-to-face settings, if adequate protocols are in place, online tests may be as, if not more, secure than face to face tests.

Data

The data in the current study involves a survey administered to past candidates of LanguageCert’s International ESOL suite of English language tests. There are six tests in the IESOL suite, all of which are aligned to one of the six CEFR levels, A1 – C2. Due to language constraints, examinations offered in OLP mode are only available for candidates at B1 level and above.

The Survey

This section reports the data collection procedure, with the survey administered via the Internet (and the surveymonkey online facility) from February 2021 onwards.

The research team met in late 2020 and early 2021 to discuss issues related to the survey’s questionnaire design, with the questionnaire worked on and revised during and after the meetings via email. The questionnaire was then trialled on members of staff who had taken OLP examinations themselves. After moderating the questionnaire and making modifications, the survey instrument was finalised.

Items were posed on a 6-point Likert scale, with ‘1’ indicating a negative response or agreement, and ‘6’ a positive response or agreement. A 6-point scale was deliberately chosen to prevent respondents sitting on the fence, thus avoiding committing themselves to an opinion. The questionnaire (see Appendix 1) consisted of 22 items in two sections. Section 1 (items 1-10) comprised respondents’ personal details; Section 2 (items

12-21) comprised 10 items, probing respondents' views of their experiences, reflections on the OLP process, and their preference for taking tests by traditional means or via OLP. A final question solicited respondents' assistance in following up the questionnaire with a structured interview, followed by an open question asking if they had any additional comments that they wished to make on any aspects of the OLP process.

Ethical Considerations

The intention was to send an email link to the survey to all past LanguageCert candidates who had taken LanguageCert examinations via OLP, and who had agreed that they would be prepared to receive communications from LanguageCert.

In line with European Union General Data Protection Regulations, LanguageCert candidates state whether they are prepared to be contacted, or to receive any form of communication from LanguageCert, subsequent to having taken an examination. This consideration was therefore borne in mind when past candidates were approached – i.e., were sent the link – to participate in the survey.

Data

The number of candidates that have taken IESOL examinations from March 2019 to the running of survey in March 2021 is presented in Table 1. As mentioned, OLP is only offered to candidates at levels B1 and above.

Table 1: Survey – Past candidates contacted

	Number
LanguageCert IESOL candidates	19,923
LanguageCert IESOL B1 – C2	15,806
Candidates having taken tests via OLP	8,898
Past candidates sent email link to survey	7,170
Emails opened	2,917
Responses	920
Response rate	31.5%
Agreed to be interviewed	306

As Table 1 illustrates, almost 20,000 candidates had taken IESOL examinations in the two-year period up to March 2021. Excluding A1 and A2 candidates leaves a figure of 15,806 candidates, of whom 8,898 (56.3%) had taken IESOL examinations via OLP since the commencement of OLP exams in March 2019.

Nulty (2008), in a summary of studies of both online and paper surveys, reports that online surveys in general achieve a rather lower response rate than paper-based surveys. He cites a figure of, on average, 33% for online, as against 56% for paper surveys. The current response rate of 31.5% comes very close to Nulty's figure.

Data Analysis

An analysis of the survey data will now be presented.

First, the robustness of the questionnaire is gauged through reliability analysis and factor analysis. A presentation of key descriptives is then made – followed by an exploration of the inferential data.

Reliability

The first step in assessing the reliability of a questionnaire involves using the Cronbach alpha statistic. The analysis of the ten attitudinal items on the questionnaire via Cronbach's alpha returned a figure of 0.88. Given that a level of 0.8 is generally recommended as desirable in a questionnaire (e.g., Trobia, 2011), this suggests that the construction of the questionnaire – given that there were only ten items – was acceptable.

Factor Analysis

An exploratory factor analysis using Principal Component Analysis (PCA) with varimax rotation was conducted (working on the assumption that the underlying factors in the survey are related) to explore how the major constructs of the questionnaire were patterned, and whether these fitted the attitudinal questions that comprised the questionnaire.

In line with Kaiser's (1974) recommendations regarding Sampling Adequacy Measures – the KMO (Kaiser-Meyer-Olkin) statistic – the figure of 0.91 indicated that the sample size was clearly adequate for factor analysis. Results from Bartlett's test of sphericity – $\chi^2 (55) = 4,529.03$, $p < .000$ – indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Only two components had eigenvalues over Kaiser's criterion of 1, which in combination explained 64.2% of the variance.

Taking loadings above 0.4 as being indicative of a cut-off point appropriate for interpretative purposes (Stevens, 2002), two possible factors emerge in the Component Matrix. Table 2 elaborates, with the items grouped together by factor.

Table 2: Component Matrix

Components	1	2
##13. How clear were OLP setup instructions?	.839	
##14. How straightforward was the OLP setup process?	.791	
##15. How was the online connection with the interlocutor?	.722	
##16. How clear was the interaction with the interlocutor?	.814	
##17. How was the overall OLP experience?	.775	
##18. Preference for tests by traditional means or by OLP?		.813
##19. "Taking tests by OLP is a more personal experience"		.785
##20. "Taking tests by OLP is more efficient"		.836
##21. Your score: better on traditional or OLP tests?		.715
##12. How anxious were you before the OLP test?		

As can be seen from Table 2, two clear factors emerge, with each factor, or latent trait, having at least four indicators (i.e., items), the minimum cutoff recommended (see e.g., Yan, 2020). The factors that emerge may be defined as:

- (1) 'institutional' – comprising items 13, 14, 15, 16, 17, and centring around delivery of the test by OLP means.
- (2) 'personal' – comprising items 18, 19, 20, 21, and centring around personal reactions to taking a test by via OLP.

Item 12 – "anxiety" – appears to be in a category of its own, not being included in either of the two factors identified. Such 'isolation' is mirrored in the responses below, as will be seen. The factor analysis would therefore appear to bear out the validity of the questionnaire.

Descriptive Statistics

In the analysis below, responses are presented by the 920 respondents in the sample. Where possible, figures are matched against the general demographic trends of LanguageCert IESOL tests.

Demographics

This section presents a comparative picture of survey respondents versus the bigger picture of the entire cohort of LanguageCert candidates of IESOL tests. The IESOL test registration form asks candidates for detail on gender, age, and mother tongue. Since not all supply their details, there is consequently a degree of missing data in the IESOL whole test figures. The survey, however, requested that respondents provide this demographic data, and all complied. Table 3 presents a comparative picture of respondents to the survey and candidate demographics from the 15,806 candidates who have thus far taken the B1 – C2 tests.

Table 3: Comparative picture of demographics in whole test and survey cohorts

	IESOL whole test cohort (%)	Survey cohort (%)
Test level		
B1	20.7	18.1
B2	38.3	29.8
C1	26.6	25.9
C2	14.3	26.3
Gender		
Female	53.5	63.9
Male	38.5	36.1
Age		
<21	35.1	10.1
21-30	36.0	27.0
31-40	16.6	28.4
41-50	7.9	24.3
>50	4.4	10.1
Mother tongue		
Greek	19.7	29.9
Italian	24.3	23.1
Chinese	13.9	7.0
Spanish	12.9	10.3
Polish		7.6
Other languages	29.2	22.1

As can be seen from Table 3, the distribution of candidates across tests is broadly comparable, if slightly less at B1 – possibly not surprising since the medium of engagement with the online proctors for all tests is English.

Comparatively more females have taken IESOL tests than males, a proportion which is echoed in responses to the survey. Concerning age, the whole test cohort shows a skew towards the younger age where 70% are under 21 and between 20-30 years or age. This skew is not reflected in the response to the survey, where there is a more even response pattern, with 10% of responses by the youngest and oldest, and a response rate in the twenty percents by the 20-50-year olds.

The mother tongues of approximately 70% of LanguageCert IESOL candidates are Italian, Greek, Chinese and Spanish. This pattern is broadly mirrored in the response rate to the survey, although higher response rates were recorded by speakers of Greek and Chinese. The response rate of the former is perhaps understandable, given that the survey was sent out from LanguageCert's Athens centre.

Attitudinal Items

This section is in two parts. First, items which diverge greatly from the mid-point of 3.5 are discussed. This relates to the issue of 'consumer validity' (Coniam, 2013), whereby a mean score considerably above (or below) the mean indicates strong acceptance (or rejection) of the proposition. In the current study, a '6' indicated a positive and '1' a negative response; in the current dataset (see Table 4 below), strong positive responses are defined as those above '4.5' while strong negative responses would be below '2.5' (although there are none of the latter in the dataset). Following the factor analysis, items are grouped into the factors identified.

Table 4: Survey item means

Survey item (N=835)	Mean	SD
##12. How anxious were you before the OLP test?	3.3	1.7
##06. Assessment of personal computer literacy	4.9	1.1
'institutional' items		
##13. How clear were OLP setup instructions?	5.1	1.3
##14. How straightforward was the OLP setup process?	4.8	1.4
##15. How was the online connection with the interlocutor?	4.7	1.5
##16. How was the interaction with the interlocutor?	5.3	1.2
'personal' items		
##17. How was the overall OLP experience?	5.1	1.2
##18. Preference for tests by traditional means (1) or tests by OLP (6)?	4.8	1.6
##19. "Taking tests by OLP is a more personal experience"	4.5	1.5
##20. "Taking tests by OLP is more efficient"	4.7	1.4
##21. Your score: better on traditional (1) or OLP (6) tests?	4.6	1.6

Item 12 had the lowest mean score, just below the central mean of 3.5. This is perhaps understandable given that, for many candidates, this was the first time they had ever taken an examination via OLP. Item 6 – a demographic question asking for an assessment of personal computer literacy – shows that candidates felt that they did not have problems working with computers or in interacting online. This suggests that the anxiety they felt may be attributed more to the looming examination than to how to respond via a computer. Nonetheless, the wide SD in item 6 (anxiety) is illustrative of the fact that, despite being computer literate, many are still concerned as they begin to take the examination.

Despite the anxiety many candidates clearly feel, the responses to the attitudinal items are all very positive. 4.5 has been proposed as a benchmark for endorsement of a proposition (Coniam, 2013). The positive nature of the responses may be seen by virtue of all the 'institutional' items having means in the high 4's or above 5. For the majority of respondents, the setup process was felt to be unproblematic; online connections were good; OLP setup instructions were clear; and interaction with the interlocutor was rated very highly indeed.

Responses to the 'personal' items were also, on the whole, very positive, with the overall OLP experience in particular rated above 5. Respondents showed a clear preference for taking tests by OLP as opposed to by traditional means. The extent to the preference for OLP has been spurred on by Covid, or is a sign of the times, will remain to be seen. Nonetheless, the fact that taking tests by OLP was seen to be a more personal experience may give an indication as to the greater update and longer term acceptance of OLP.

Looking towards the future, on the issue of preference for tests by traditional means (1/6) or via OLP (6/6), a mean of 4.8/6 was recorded, indicative of very positive acceptance of OLP and strong future uptake of this means of test delivery. Respondents also felt that OLP was more personal and a more efficient way of taking a test.

Inferential Analysis

A chi square analysis of items where significant differences emerged will now be presented.

55 chi square analyses were conducted – the 11 attitudinal items against five background demographic variables. In general, little significance emerged on the majority of the analyses, indicating that respondents were in agreement with items irrespective of backgrounds such as gender, age, grade obtained, the level of test sat, or what test type – Speaking or LRW test – had been taken.

There were only six instances of statistical significance, where 5% is taken as the level of significance. Table 5 elaborates, with the attitudinal item and variable affected identified.

Table 5: Attitudinal items – significant differences (sorted by item)

Attitudinal items	Variables	Significance
##06. personal computer literacy	Gender (females)	$\chi^2(5) = 15.94, p = .007$
##06. personal computer literacy	Age (>50)	$\chi^2(20) = 42.56, p = .002$
##06. personal computer literacy	Level (B1)	$\chi^2(15) = 37.48, p = .001$
##06. personal computer literacy	Grade (Fail)	$\chi^2(15) = 55.41, p < .001$
##12. anxious before the OLP test	Grade (High pass)	$\chi^2(15) = 42.10, p < .001$
##17. Evaluation of the OLP experience	Age (> 50)	$\chi^2(20) = 31.63, p = .047$

Of the six instances of statistical significance, four instances were related to perceived personal computer literacy – despite this item receiving an overall comparatively high mean of 4.9/6. Females felt they were less technologically able than men, as did older respondents. Candidates who took a B1 test were in general less positive. As mentioned, the OLP process is conducted through the medium of English. This indicates that for lower-level candidates – B1 in particular – the OLP experience via English puts extra demands on candidates with a lower ability in English. The fact that candidates who failed were more negative in their perceptions is perhaps understandable. It is possibly surprising that significance only emerged for fail grade candidates on one variable.

Significance was only reported against grade received, with, interestingly enough, high pass candidates responding the most negatively. This is suggestive of examination apprehension

Regarding reactions to the overall OLP experience, on this variable, significance was recorded by the oldest age group (above 50 years) being more negative than other age groups. This response echoes this age group's perception that they are also less computer literate than younger age groups.

Comments

In response to the final item requesting any comments respondents had, a total of 716 comments were received. After undergoing a thematic analysis, the comments have been categorised, and totals tabulated. Table 6 presents a picture of the results.

Table 6: Written comments provided (N=716)

Total positive: 234		Total negative: 68	
admin	9	admin	7
computer issues	6	computer issues	5
connection issues	6	connection issues	21
interlocutors	32	interlocutors	3
test materials	2	test materials	2
convenient	17	covid issues	3
efficient	7	test delivery	20
'minimalist'	157	other	7
total	234	total	68
"No comment": 394		Off topic: 20	

Of the 716 total written comments received, 394 were framed simply as "no comment", "nothing to add" etc. 20 comments were classified as "topic", e.g., "A more accepted institute will be great. like Canada immigration." The latter two categories have been disregarded in the following discussion. 157 'minimalist' positive comments such as "very good", "thanks" were received; these have also been disregarded.

While the total number of positive comments outweighed the total number of negative comments by a considerable margin, in terms of *substantive* comments, the balance was about equal. In this context, O'Cathain & Thomas (2004) discuss how open questions at the end of a survey help to "redress the power balance between researchers and participants". This is where respondents raise issues not covered by the closed questions – their "safety net" as Biemer et al. (2011) describe the open questions. In this light, the considerable number of negative comments – as compared to the general tenor of positive responses to the closed questions – is understandable. This is the forum for negatively-oriented respondents to air any specific grievances they may have.

Figure 3 present a sample of comments on some of the key themes. Some themes received comments from both sides of the spectrum; some themes – connection issues, test delivery – were themes for negative comments.

Figure 3: Major positive and negative comments themes: Examples

Administration of exams	
<i>Positive tenor</i>	<i>Negative tenor</i>
Just very happy with the service, they kept me informed about any changes, the people at the test centre very friendly and they followed the covid strict standards.	I didn't get my results in 3 business days as it is written in the home page of the website so I couldn't use this certificate for an exam and so I couldn't get a better mark in this examination.
	The most challenging and conflict part of my test was being told in the morning that my test has been cancelled on the short notice but also rewriting subject that I already passed.

Computer issues	
<i>Positive tenor</i>	<i>Negative tenor</i>
My overall opinion on OLP tests is very positive, however I would like to report that my initial experience was a bit troublesome due to incompatibility of the program with Mac IOS. This should be made clear previously and maybe require the use of another system.	Earphones should not be necessary

Interlocutors	
<i>Positive tenor</i>	<i>Negative tenor</i>
The lady who examined me was very nice and helpful. The internet connection wasn't very good, but thanks to her everything was alright. Thank you!	I didn't expect proctors speaking with an inflection different from British or American one, so I struggled to understand her.
	I think you should solve the connection quality. The examiner was really nice and understood the situation but I felt really nervous cause there were a lot of breaks cause of the Internet. Anyway thanks a lot for this kind of certifications which allow us being prepared despite of COVID. Best regards.

Certain themes were skewed in terms of negative comments.

Connection issues	
<i>Positive tenor</i>	<i>Negative tenor</i>
	The internet connection in China mainland for online test is very poor. No matter I tried to use a VPN or not. And I have encountered several times that the application crashed. It would be appreciated if you could try to improve your application regarding the internet connection and crashing issues. The examiner is kind and professional.
	I hope the examiner can consider the network factors of both sides in the oral test. If the examinee's performance is not good due to the network reasons, should we give a chance to retake the test

Test delivery issues	
<i>Positive tenor</i>	<i>Negative tenor</i>
	At the start of my test, another candidate was quite close and I could hear a fair amount of background noise. This was not great.
	Listening part felt really loud, ExamShield wouldn't allow me to permanently lower the volume.
	Difficulties in the test:–the fact that we cannot print the test so that we can read it. It is not allowed to use a draft notebook thanks

Convenience	
<i>Positive tenor</i>	<i>Negative tenor</i>
For me it was a great occasion to take this certification during a pandemic.	
I consider OLP exams a very efficient way to take exams, especially in this period of pandemic	

Efficiency	
<i>Positive tenor</i>	<i>Negative tenor</i>
I would like to congratulate Language cert for the innovation way to examine students' English proficiency during covid-19 lockdown. I also need to say a huge thank you to the personnel for the kindness ,the patience and the excellent service.	

Certain of the issues raised above – such as scrolling, adjusting volume, seeing the amount left to be covered is outlined in sample material and practice sessions on the LanguageCert website. If test takers have issues, comments, complaints following a test they have taken, there is a *Special Considerations* section which handles and investigates such problems. Following a review of the OLP video, for example, scores may be adjusted if a test taker is considered to have been disadvantaged.

As mentioned above, it is in this open comments section that the disaffected may make their voice heard. This has been the case with the survey administered in this study. The negative comments do nonetheless provide for thought, and follow-up action. One issue, which is under constant review, is that of internet connections. These have recently been upgraded; and it is hoped that some of the issues raised by respondents here will have been addressed.

Conclusion

This paper has explored the reactions to and perceptions of OLP by candidates who had taken an examination via online proctoring, specifically in the context of English language examinations delivered by LanguageCert. A survey was sent out to all past candidates of LanguageCert IESOL examinations who had agreed to be contacted. Of those who saw the request, a 31.5% response rate was achieved, in line with what might be expected for online surveys.

Demographically, responses were broadly comparable with the cohorts who have taken LanguageCert examinations over the two-year period in which the examination body has been operating.

Responses to all attitudinal questions returned high positive means, endorsing all aspects of the OLP process. Instructions, the setup process, interactions with the interlocutor – all received very positive responses. A comparatively high number of negative written comments were received compared to the general positive tenor to the responses to the closed questions. These have been passed to the relevant body, with some issues, such as internet connections, constantly receiving attention from the systems section of the company.

Inferential analysis revealed computer literacy to be significant against certain demographic factors. Females and older respondents felt they were less able technologically – findings reported in previous studies (see e.g., Yau & Cheng, 2012). B1 level candidates were less positive: the fact that the OLP process is all conducted, and explained, through the medium of English is an issue that needs to be considered. LanguageCert is currently looking at providing detail to candidates in major languages other than English; this issue possibly also needs to be given consideration regarding the OLP logon, onboarding and setup process.

Regarding preference for tests by traditional means or via OLP, a strong endorsement of OLP was recorded. Respondents felt that OLP was more personal and a more efficient way of taking a test – possible an effect of test delivery via OLP having continued throughout the covid pandemic. All these positive signals are clearly indicative of the broad acceptance of OLP, pointing to strong future uptake of the OLP mode of test delivery.

References

- Alexander, M. W., Bartlett, J. E., Truell, A. D., & Ouwenga, K. (2001). Testing in a computer technology course: An investigation of equivalency in performance between online and paper and pencil methods. *Journal of Career and Technical Education*, 18(1), 69-80.
- Ardid, M., Gómez-Tejedor, J. A., Meseguer-Dueñas, J. M., Riera, J., & Vidaurre, A. (2015). Online exams for blended assessment. Study of different application methodologies. *Computers & Education*, 81, 296-303.
- Berkey, D., & Halfond, J. (2015). Cheating, student authentication and proctoring in online programs. *New England Journal of Higher Education*.
- Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., & Sudman, S. (2011). Measurement errors in surveys, (Vol. 173). Hoboken.
- Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology*, 54(2), 199-231.
- Bonk, C. J., Lee, M. M., Reeves, T. C., & Reynolds, T. H. (Eds.). (2015). MOOCs and open education around the world. Routledge.
- Carnegie Council on Policy Studies in Higher Education, & Carnegie Commission on Higher Education. (1979). Fair practices in higher education: Rights and responsibilities of students and their colleges in a period of intensified competition for enrollments: a report of the Carnegie Council on Policy Studies in Higher Education. Jossey-Bass.
- Clark, T. M., Callam, C. S., Paul, N. M., Stoltzfus, M. W., & Turner, D. (2020). Testing in the time of COVID-19: A sudden transition to unproctored online exams. *Journal of chemical education*, 97(9), 3413-3417.
- Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E., & Thies, W. (2015). Deterring cheating in online environments. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(6), 1-23.
- Exam.net 2020, A robust, easy-to-use and secure exam platform, viewed June 10th 2020, <<https://exam.net/>>.
- Foster, D., & Layman, H. (2013). Online proctoring systems compared. Webinar. <http://www.slideshare.net/caveonweb/caveon-webinar-series-online-proctoring-best-practicesoct-2013-slideshare-final>.
- Gardner, L. (2020). Covid-19 has forced higher ed to pivot to online learning. Here are 7 takeaways so far. *The Chronicle of Higher Education*, 20.
- Hillier, M., & Fluck, A. (2013). Arguing again for e-exams in high stakes examinations. In ASCILITE-Australian Society for Computers in Learning in Tertiary Education Annual Conference (pp. 385-396). Australasian Society for Computers in Learning in Tertiary Education.
- Hollister, K. K., & Berenson, M. L. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision Sciences Journal of Innovative Education*, 7 (1), 271–294.
- Hussein, M. J., Yusuf, J., Deb, A. S., Fong, L., & Naidu, S. (2020). An evaluation of online proctoring tools. *Open Praxis*, 12(4), 509-525.
- Khan, R. A., & Jawaaid, M. (2020). Technology enhanced assessment (TEA) in COVID 19 pandemic. *Pakistan Journal of Medical Sciences*, 36(19), 108-110.
- King, C. G., Guyette Jr, R. W., & Piotrowski, C. (2009). Online exams and cheating: An empirical analysis of business students' views. *Journal of Educators Online*, 6(1), 1–11.
- Lim, C. P., & Wang, L. (Eds.). (2016). Blended learning for quality higher education: Selected case studies on implementation from Asia-Pacific. Bangkok: UNESCO Bangkok Office.

- Lim, C. P. & Graham, C.. (eds.) in press 2020. Blended Learning in Asia.
- Marshall, D. T., Shannon, D. M., & Love, S. M. (2020). How teachers experienced the COVID-19 transition to remote instruction. *Phi Delta Kappan*, 102(3), 46-50.
- Mogey N, Cowan J, Paterson J, Purcell M. 2012. Students' choices between typing and handwriting in examinations. *Active Learning in Higher Education*, 13(2):117-128. doi:10.1177/1469787412441297.
- Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & evaluation in higher education*, 33(3), 301-314.
- Oaks, H. R. (1975). Cheating attitudes and practices at two state colleges. *Improving College and University Teaching*, 23(4), 232-235.
- O'Cathain, A., & Thomas, K. J. (2004). "Any other comments?" Open questions on questionnaires—a bane or a bonus to research?. *BMC medical research methodology*, 4(1), 1-7.
- Proctorio 2020, A Comprehensive Learning Integrity Platform, viewed June 10th 2020, <<https://web.proctorio.com/>>.
- Qualifications and Curriculum Authority. (2007). Regulatory principles for e-assessment. <https://publications.parliament.uk/pa/cm200607/cmselect/cmeduski/memo/test&ass/ucm3102paper4.pdf>.
- Respondus 2020, Assessment tools for learning systems, viewed June 10th 2020, <<https://web.respondus.com/>>.
- Rose, C. (2009). Virtual proctoring in distance education: An open-source solution. *American Journal of Business Education (AJBE)*, 2(2), 81-88.
- SEB 2020, Safe Exam Browser, viewed June 10th 2020, <<https://safeexambrowser.org/>>.
- Sierles, F., & Hendrickx, I. (1980). Cheating in medical school. *Academic Medicine*, 55(2), 124-5.
- Tan-Choi, A., Tinio, V. L., Castillo-Canales, D., Lim, C. P., Modesto, J. G., & Pouzevara, S. R. (2020). Teacher's guide for remote learning during school closures and beyond. Quezon City, Philippines: Foundation for Information Technology Education and Development.
- TPD@Scale for the Global South, 2020. Teacher's guide for remote learning during school closures and beyond. Information Technology Education and Development, Inc. <https://tpdatscalecoalition.org>.
- Truszkowski, D. (2019). Proctored versus non-proctored testing: a study for online classes (Doctoral dissertation, the American College).
- Watson, G., & Sottile, J. (2008, March). Cheating in the Digital Age: Do students cheat more in on-line courses?. In Society for Information Technology & Teacher Education International Conference (pp. 798-803). Association for the Advancement of Computing in Education (AACE).
- Weiner, J. A., & Hurtz, G. M. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology*, 18(1), 13-20.
- Wibowo, S., Grandhi, S., Chugh, R., & Sawir, E. (2016). A pilot study of an electronic exam system at an Australian University. *Journal of Educational Technology Systems*, 45(1), 5-33.
- Wiesenber, F. P., & Stacey, E. (2008). Teaching philosophy: Moving from face-to-face to online classrooms. *Canadian journal of university continuing education*, 34(1).
- Wright, J. C., & Kelly, R. (1974). Cheating: Student/faculty views and responsibilities. *Improving College and University Teaching*, 22(1), 31-34.
- Yau, H. K., & Cheng, A. L. F. (2012). Gender difference of confidence in using technology for learning. *Journal of Technology Studies*, 38(2), 74-79.

Online-Proctored Tests: Experiences and Reflections

We would be very grateful if you could take a few minutes to reflect on the online-proctored English language test that you took with LanguageCert. Please click on the circle, or select the number of stars as appropriate. **You do not need to identify yourself. All information collected is for research purposes only, and will be kept in the strictest confidence.**

OLP = online proctored; A' Traditional Test ' = a test by pen and paper; in a regular school or Test Centre setting
 LRW = Listening, Reading and Writing

Section 1: Personal Details

##01. I am	<input type="checkbox"/> Male <input type="checkbox"/> Female
##02. I am years old	<21 21-30 31-40 41-50 >50
##03. I live in (country)	
##04. My mother tongue is ...	
##05. My education level is ...	<input type="checkbox"/> Primary <input type="checkbox"/> Secondary <input type="checkbox"/> Bachelor Degree <input type="checkbox"/> Higher Degree
##06. How computer literate do you consider yourself?	<input type="checkbox"/> not at all <input type="checkbox"/> very
##07. The last OLP test I took was at level ...	<input type="checkbox"/> B1 <input type="checkbox"/> B2 <input type="checkbox"/> C1 <input type="checkbox"/> C2
##08. How many LRW tests have you taken by OLP?	1 2 3 4 >4
##09. How many Speaking Tests have you taken by OLP?	1 2 3 4 >4
##10. What grade did you get on your last OLP test?	<input type="checkbox"/> Fail <input type="checkbox"/> Pass <input type="checkbox"/> High Pass <input type="checkbox"/> Prefer not to say

Section 2: Experiences and Reflections

##11. Respond to the questions below EITHER (1) about Speaking; OR (2) about Listening, Reading and Writing (LRW)	I am responding about <input type="checkbox"/> Speaking <input type="checkbox"/> LRW	
##12. How anxious did you feel before your OLP test?	very anxious	not anxious at all
##13. How clear were the OLP setup instructions for the test?	not clear at all	very clear
##14. How straightforward was the OLP setup process?	very troublesome	very straightforward
##15. How was the online connection between you and the interlocutor during the test?	very poor	very good
##16. How clear were the interlocutor's instructions and directions during the test?	not clear at all	very clear
##17. How was the overall OLP experience?	very poor	very good
##18. Do you prefer to take tests by traditional means or by OLP?	prefer traditional	prefer tests by OLP
##19. "It is a more personal experience to take tests by OLP than to take tests by traditional means"	strongly disagree	strongly agree
##20. "It is more efficient to take tests by OLP than to take tests by traditional means"	strongly agree	strongly disagree
##21. Do you think that you score better in tests taken by traditional means or in tests by OLP?	better on traditional better on OLP tests	
##22. Would you be available for a short follow-up (online) interview?	YES / NO. If yes, please leave your email or phone number.	
Do you have any comments that you would like to add?		

Chapter 13: Automated Writing Assessment (AWA) and the Carnegie Speech Writing Assessment System

David Coniam and Tony Lee

Abstract

This paper reports on a study to validate the use by LanguageCert of the Carnegie Speech (CS) Automated Writing Assessment (AWA) system. Following a brief introduction to the computer assessment of writing, the report details a study using a reference dataset from the Hong Kong Year 11 public exam (Coniam, 2009) comprising 300 scripts – largely at CEFR B2 level – where the marker statistics and candidates' overall subject score on the public examination were known background variables.

The study explores two key issues. The first is the reliability of the CS AWA system compared with that obtained in the previous study. Second, the study explores a methodology where the output of the CS AWA system (linked to the nine-point IELTS scale) can be equated with LanguageCert's Writing test scale where both scales are also aligned to the CEFR.

The analyses performed confirm that the scores produced by Hong Kong (human) markers correlate at a moderate-to-high level with those produced by the CS AWA system. Furthermore, the Rasch methodology used to equate the two scales (the six-point Hong Kong scale and the nine-point CS AWA / IELTS scale) illustrates that the scores produced on the two scales might be successfully aligned.

Introduction

The use of computers for assessment purposes has grown considerably since the 1980s, with much research and development in computer-based testing (Chapelle and Douglas, 2006). Studies demonstrate the advantages of computers in ease of assessment, marking etc. (Alderson, 2000), and in computer adaptive tests using

tailored tests (Chalhoub-Deville & Deville, 1999). Item types have nonetheless tended to focus on limited selection types such as multiple-choice or fill-in-the blanks (Alderson, 2000; Clapham, 2000).

Some 30 years ago, Warschauer & Healey predicted an *Intelligent CALL* phase where software should offer “easy interaction with the material to be learned, including meaningful feedback and guidance” as well as “comprehensible information in multiple media designed to fit the learning style of individual students” (1998). In the 2020s, Warschauer & Healey’s vision is coming much closer to reality with the use of natural language processing (NLP) in language assessment. While NLP represents a long-term aim in assessment, a considerable number of NLP elements are beginning to be incorporated into language assessment (see Lu, 2017; Yannakoudakis et al. 2018)

One particular area where the incorporation of the use of computers in language assessment has been generating a great deal of interest is in the automated rating of written essays. Automated Writing Assessment (AWA) has seen substantial development over the past two decades, with a considerable number of programs now available – see e.g., Shermis, 2014.

Two major reasons are advanced for using computers to score essays. The first concerns time and money where access to a reliable computer program may save raters, or assessment bodies, hours grading papers (Chapelle & Douglas, 2006). The second claim is that rating essays is a very subjective task – as Hughes (2003) noted some three decades ago. Raters may be using a marking scheme, or rubric, where they are not totally au fait with the scoring domains and levels. How raters arrive at a score, even when using a marking rubric, is still not totally clear (see e.g., Attali, 2013). Consequently, it is argued that the use of a computer rater – and ‘unbiased’ or objective software – avoids essays being graded or evaluated by human assessors who may be more prone to variability.

Shermis (2014) reports on the large-scale *Hewlett Foundation: Automated Essay Scoring Competition* in the USA in 2012. This competition examined the performance of some of the leading commercially available AWA systems, comparing the computer ratings with those of human raters. How such programs operate, their reliability etc, provide a useful backdrop for the current study, thus some background is provided to aid discussion. The key programs in the competition (discussed below) were:

Table 1: AWA systems in the Hewlett Foundation: Automated Essay Scoring Competition

AWA system	Developer
AutoScore	American Institutes for Research
Bookette	CTB McGraw-Hill
CRASE	Pacific Metrics
e-rater	Educational Testing Service
Intelligent Essay Assessor	Pearson Knowledge Technologies
IntelliMetric	Vantage Learning
Lexile Writing Analyzer	MetaMetrics
LightSIDE	Teledia Laboratory
Project Essay Grade	Measurement Inc.

Analytic Systems Used by AWA in Analysis

Earlier (i.e., prior to the 2010s), the majority of essay-grading software functioned by analysing sentences and paragraphs, looking for keywords as well as relationships between terms. The *Intelligent Essay Assessor* (IEA), for example, used latent semantic analysis (LSA) for most of its analysis (Foltz et al., 1998). Through training with LSA, after producing a matrix of words and documents, pieces of writing were then scored to see how well they matched the matrix.

The *Project Essay Grade* (PEG) system – to some extent the ‘grandparent’ of AWA which dates back to the 1960s – functioned by first identifying elements such as sentence length, number of paragraphs and elements of punctuation, and then used regression to determine how well the different variables correlated with the scores of human raters.

Educational Testing Service’s *E-rater* (Burstein, 2003) was possibly one of the first AWA systems to make extensive use of NLP to identify and examine specific linguistic categories such as grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage. It functioned on similar lines to PEG, using regression to predict performance. Chen et al. (2017)

The *Bayesian Essay Test Scoring sYstem* (BETSY), drew on a broad set of essay features for analysis. After a set of parameters had been specified, the system functioned by probabilistically assigning occurrences of content and stylistic features to a set of specified levels and determining the probabilistic chance of features in input essays falling into the levels stipulated (see Rudner & Liang, 2002).

Table 2 presents a snapshot of the analytic techniques and the main language focus of the key AWA programs which participated in the Hewlett Foundation (HP) 2012 Automated Essay Scoring Competition.

Table 2: Key AWA systems [adapted from Shermis (2014)]

AWA system	Analytic technique	Main focus	Essays for training
AutoScore	NLP	Style and content	n/a
Bookette	NLP	Style and content	250-500
CRASE	NLP	Style and content	100 per score point
e-rater	NLP	Style and content	100-1,000
Intelligent Essay Assessor	LSA	content	100-300
IntelliMetric	NLP	Style and content	300
Lexile Writing Analyzer	NLP	Style and content	None, uses fixed model
LightSIDE	Statistical	Style and content	300
Project Essay Grade	Statistical	Style	100-400

Issues of Reliability and Validity in the Use of AWA

In terms of reliability, many studies of AWA systems over the past two decades have reported favourably on different computerised essay-scoring programs, with many AWA studies reporting scores that correlate as highly with human raters as raters do with each other (Chung & O’Neil, 1997; Coniam, 2009; Shermis, 2014).

In the context of the HP Automated Essay Scoring Competition, referred to above, Shermis (2014) concludes “in a high-stakes testing environment machine-predicted scores came close to matching the distributional and agreement characteristics of scores assigned by human raters.” That is, all the systems may be seen as having high reliability.

While high reliability may be reported by AWA systems, the same cannot be said for validity.

One criticism of AWA systems is that the computer rating process is essentially a “black box” (see Weigle, 2002). Attali & Burstein (2006) stated:

Whereas human raters may directly evaluate various intrinsic variables of interest, such as diction, fluency, and grammar, in order to produce an essay score, AES systems use approximations or possible correlates of these intrinsic variables.

Two decades ago, Drechsel (1999), for example, argued that AWA systems did not read and understand essays like humans. While the extensive use of NLP features may, to an extent, mitigate Drechsel’s “black box” concerns, most AWA systems arrive at their analytic decisions after either an opaque set of analyses such as latent semantics (IEA), or slightly less controversially, via linguistic feature analysis such as that produced by e-rater – as Table 2 above illustrated.

Bennett and Zhang (2016) contest Shermis’s (2014) claim that machine and human scores had “virtually identical levels of accuracy”. They discuss how little was known about how the top performers in the HP competition optimised their scoring systems to the indicator by which they were judged and elaborate how some of the scoring systems have virtually nothing to do with candidates’ levels of language, and hence are very low in terms of validity.

A further issue – as reported in Table 2 above (see also Coniam, 2009) – is that AWA systems have traditionally required training with, in some cases, large number of scripts for each topic: e-rater, for example, required up to 1,000 training scripts. As computer systems advance, however, this issue is becoming less of a problem.

In summary, and putting concerns about validity aside for the moment, many studies have illustrated that the scoring capability of many AWA systems produce scores that correlate highly with those of human raters with high reliability and validity (Coniam, 2009; Lee, Gentile, & Kantor, 2010). LC is considering using an AWA system either as a second rater, with an already-established set of rating scales, or as a single rater in the context of medium stakes assessments such as LTE. The current study focuses on the reliability of the AWA system.

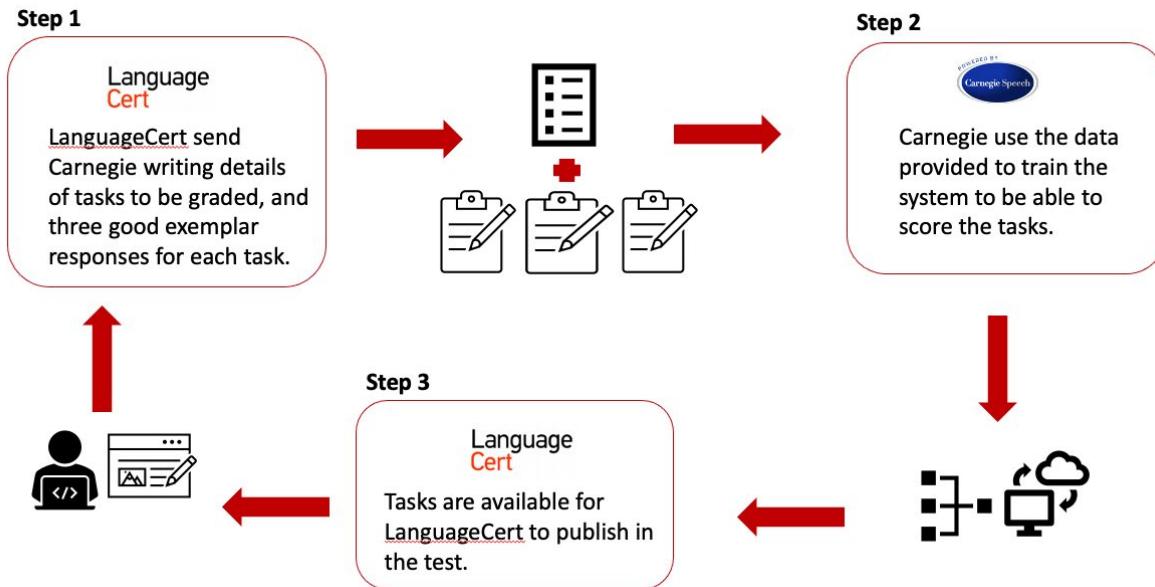
Background to the Current Study

LanguageCert is considering using the Carnegie Speech (CS) AWA engine. This section will first present a brief overview of the CS system, with the key focus being the reliability and robustness, of the CS AWA system. A comparative study will be described. This study involved scripts with known values and which had been scored in a previous study (Coniam, 2009) being rated by the CS system.

The engine will first be described, its analytic technique and output, after which a dataset which has been validated and reported on in another study (Coniam, 2009) will be presented. Finally, a methodology for recalibrating relating scores from the CS AWA system to the *LanguageCert* Writing scores will be presented.

The process for setting up test content in the CS AWA system is as follows in Figure 1.

Figure 1: Operation of the CS AWA system



The CS AWA system gets sent three 'good' responses for each topic to be rated – a considerable advance from the 100-1,000 training scripts needed previously by some systems. The CS system is then fed scripts, which it grades on a version of the IELTS 9-point marking scheme – i.e., four scoring domains, with each domain marked on nine-points or levels. The system returns a score per domain, as well as an overall average across the four domains.

The Previous Study

The reference dataset (reported in Coniam, 2009) comprised 300 scripts from the Hong Kong Year 11 (age 17) HKCE (Hong Kong Certificate of Education) Writing examination from 2005. The topic on which candidates had to write approximately 300 words in 70 minutes was "Write an article for the school magazine describing your experience during a 'Working Week' campaign you participated in and describing how you felt about the job you chose."

The 300 scripts were drawn from a representative cross-section of the Year 11 English language population. All 300 scripts had been marked by two 'good' markers (i.e., with good statistics); the study also had access to candidates' overall subject score (i.e., the total for all four papers) which comprised the HKCE English language examination.

The HKCE Writing was marked against four subscales, each comprising six levels. These are provided in Table 3, along with the marking scales used by *LanguageCert* and the CS AWA system.

Table 3: Marking schemes

	1	2	3	4	5	Levels
LC	Task fulfilment	Accuracy and range of grammar	Accuracy and range of vocabulary	Organisation		0-3
IELTS	Task achievement	Grammatical range and accuracy	Lexical resource	Coherence and cohesion		1-9
Carne-gie	Task Response	Grammatical Range and Accuracy	Lexical Resource	Coherence & Cohesion		1-9
HKCE	Relevance & adequacy of Content for purpose	Accuracy & appropriacy of punctuation, vocabulary, language patterns		Planning & Organisation	Appropriacy of tone, style & register; appropriacy of features for Genre	(0) 1-6

The HKCE examination was externally benchmarked against the IGCSE. In turn, the IGCSE has been benchmarked against the CEFR. Table 4 below plots the rough fit of correspondences between the grades across the different systems, and how these broadly translate to CEFR and IELTS levels (against which the CS AWA system returns its scores).

Table 4: Correspondences between grades

Exam	Levels			
HKCE	5* ('6')	5	4	3
IGCSE	A*	A	B	C
CEFR	C1	B2	B2	B2
IELTS	7	6.5		5.5

The AWA program used to rate the HKCE scripts was BETSY (Rudner & Liang, 2002). Despite BETSY being something of a black box in terms of how it analysed, it achieved correlations with the mean of the two raters of 0.82, virtually identical to the two markers' inter-marker correlation.

The Current Study

To issues are explored in the current study, one analytic and one methodological.

The *analytic focus* concerns the reliability of the CS AWA system. For the moment, the fact that the CS AWA system produces output which mirrors LC's own marking domains (see Table 3) is taken as a starting point for

construct validity. Given that a good inter-rater correlation is taken as 0.8 (Hatch & Lazaraton, 1991), the same figure will be the target in the following:

1. Inter-marker agreement between the CS AWA system and the HKCE markers
2. Correlation between the CS AWA and the BETSY grades
3. Correlation between the CS AWA system and the HKCE subject grade score

The *methodological issue* concerns how the output produced by the CS system on a nine-point scale may be rescaled to fit *LanguageCert's* six-point scale based on the CEFR.

Reliability Analysis

CS were first sent three 'good' scripts for training purposes, after which they were sent 300 scripts for analysis. Results returned consisted of a score for each candidate against each of the four domains laid out in Table 3 and an overall average score.

As an external point of reference, inter-marker correlations are first presented in Table 5, along with the correlation of the Writing test scores with the candidates' overall subject score for English. Spearman correlations are used since the scales are ordinal in measurement terms.

Table 5: Correlations

	HK Mkr 1	HK Mkr 2	CS AWA	BETSY
HK Mkr 1	—			
HK Mkr 2	0.83	—		
CS AWE system	0.69	0.69	—	
BETSY	0.80	0.78	0.66	—
HKCE subject score	0.81	0.83	0.65	0.82

The Hong Kong markers had a strong inter-marker correlation of 0.83; the CS AWA system has a moderate-to-high correlation of 0.69 with both markers. The Hong Kong markers had a very high correlation with the overall subject grade – partly inflated by the fact that the Writing test grade itself formed part of the subject mark. The CS AWA system correlated at a moderate-to-high 0.65 with the HKCE subject score. BETSY's scores correlated highly with the human markers and correlated at a moderate-to-high 0.66 with the CS AWA system. All correlations were significant at the 1% level.

Removing scripts below HKCE level 3 (approximately IELTS 5) did not improve the correlations.

A series of contingency tables were then calculated between all scale levels for all the mean HKCE ratings against the CS ratings. CS reports grades to the nearest 0.25. For ease of readability, however, means have been rounded to 0.5 in Table 6 below.

Table 6: Comparison of grades

HK mean scores	CS mean scores							Total
	5	5.5	6	6.5	7	7.5	8	
0	1	4	2	0	0	0	0	7
0.5	3	4	6	1	0	0	0	14
1	1	7	6	2	0	0	0	16
1.5	0	6	3	3	2	0	0	14
2	0	2	7	7	2	0	0	18
2.5	0	2	7	12	2	0	0	23
3	0	0	7	16	7	2	0	32
3.5	0	0	2	8	16	4	0	30
4	1	1	2	6	25	2	1	38
4.5	0	0	1	6	16	8	0	31
5	0	0	2	8	11	7	0	28
5.5	0	0	1	5	10	4	0	20
6	0	0	0	5	13	11	0	29
Total	6	26	46	79	104	38	1	300

As can be seen, as one moves from low to high on either scale, grades move in a similar fashion. This is supported by the chi square figure [$\chi^2 (72, 300) = 267.41, p < .001$] which is significant at the 1% level.

Given the approximate fit of HKCE level 5 to IELTS 6.5 (see Table 4), the fact that the CS AWA system returns a number of 7s and 7.5s for the current Hong Kong Year 11 dataset makes sense. In general, the fit between the two set of ratings scales appears to be quite close.

Although the current study has not intended to examine levels in depth, there is some possible cause for concern with the CS AWA ratings – toward the bottom end of the scale. While the HKCE grades ranged across the whole scale – from 0 to 6 – the CS grades were in a much narrower range: from 4.75 to 7.75. Some scripts graded 0 (indicative of “no or minimal performance”) by the HK raters were awarded a 5 by the CS AWA system [Note 1]. Figure 2 below presents the script of one candidate (there were two such instances in the output, graded 0, by the Hong Kong markers).

Figure 2: Mismatching candidate example

Script 2110001: HK mean 0.25; CS AWA mean 5.75

My school recently conducted a 'Working Week' scheme during many students could choose to work as one suggers.

Number one is a reporter, reporter help f.1-f.2 students read. Number two is a teacher is a teacher, teacher help F.4-F.5 students at class read book. Number three is a restaurant cook, if went mother eat, go to resaurant cook, it help you cook. Number five a photographer help you happy to gril.

My took part in the scheme. Now I careers advisor has asked my to write an article for the school magazine describing my experience during the Working Week and how me felt about the job you chose.

The script in Figure 2 is also very short – at 109 words. CS does not penalise for lack of words. It is anticipated that 'essay length' can be introduced as a variant in the CS system.

Methodological Issue: Equating Score Levels

This section explores the methodological issue of how CS AWA nine-level system may be eventually equated with the LC six-level system. In the discussion below Multi-Faceted Rasch Analysis (MFRA) using the computer program FACETS (Linacre, 2020) has been conducted on the current data to exemplify.

In the analysis below, the two sets of ratings have been regarded as different 'items', i.e., the two sets of rating scales. There are three markers: the two Hong Kong markers, and the CS AWA system. The linking elements in the calibration are the scripts, which enable calibration using the two different rating scales. The calibration results are all measured by a single measurement ruler, permitting the two scales to be matched up. The scale structures in the calibration were set as cardinal or nominal so that '0' can also be calibrated.

Multi-Faceted Rasch Analysis

The analyses below first present a picture of general fit. Following this, data is presented to illustrate where level cut scores might be drawn for the HKCE scales in relation to the data obtained from the CS analysis.

Given that the main statistical procedures used in this study involve Rasch, the reader is referred to the outline of the Rasch measurement model provided in the Glossary of statistical terms at the end of the volume.

In Rasch analysis, acceptable ranges of Infit and Outfit are 0.5 to 1.5. Table 6 presents the general picture of fit. For current purposes the logit scale has been rescaled to a mean of 50 and an SD of 10.

Table 7: Scale fit to the Rasch model

RATER	Fair Average	Measure	S.E.	Infit MnSq	Outfit MnSq
HK-2	3.54	52.28	0.37	0.65	0.64
HK-1	3.55	52.17	0.37	0.72	0.71
Carnegie	6.32	45.55	0.57	1.6	1.63
Mean (Count: 3)	4.47	50	0.44	0.99	1
S.D. (Population)	1.31	3.14	0.09	0.43	0.45
S.D. (Sample)	1.6	3.85	0.11	0.53	0.55

Mean severity levels are similar in all three raters – around the scale mean of 50. The two Hong Kong raters have both Infit and Outfit Mean squares near the lower threshold of 0.5, indicating narrower ratings. The CS AWA rater has both Infit and Outfit near the upper threshold of 0.5, indicating more varied ratings.

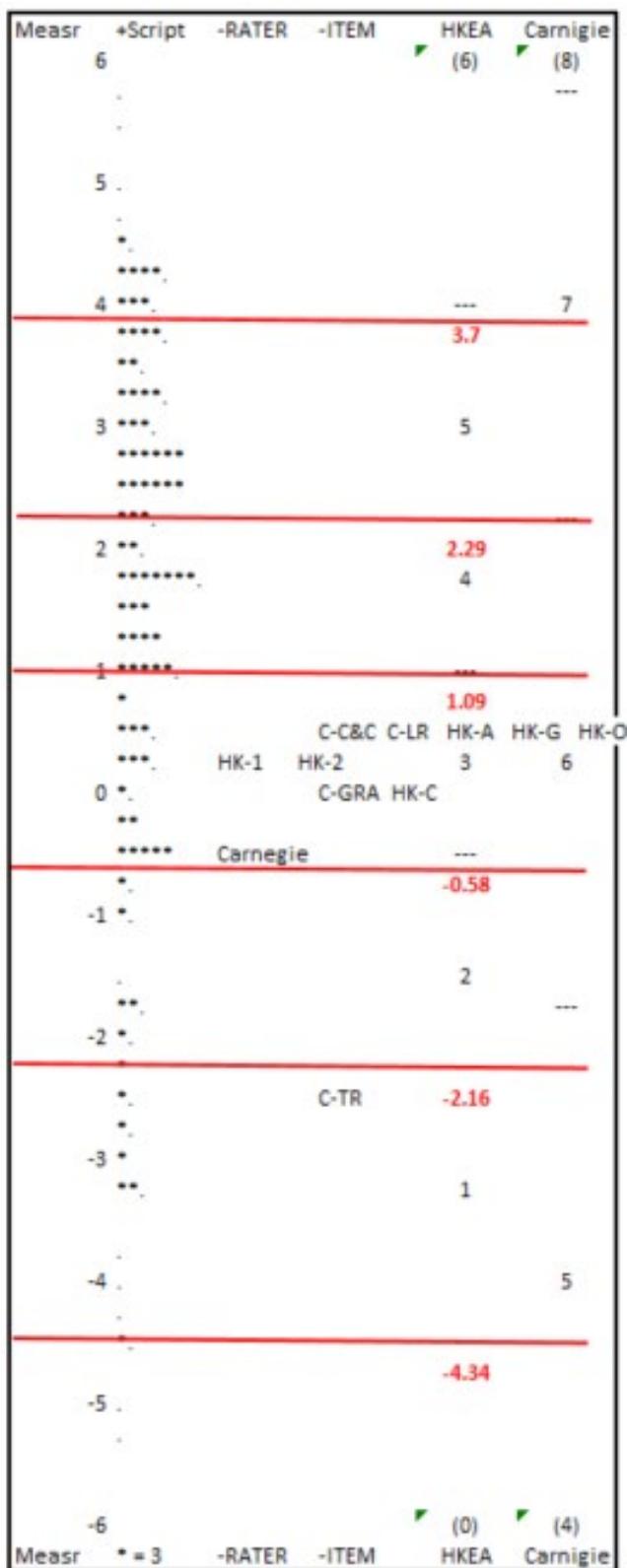
Table 7 below illustrates where the cut score for each HKCE level might be drawn. The *Rasch-Andrich Threshold* values in the final column are the logit measures which define the cut scores.

Table 7: Rasch-Andrich Threshold values

HKCE scales	DATA			QUALITY CONTROL			RASCH-ANDRICH Thresholds	
	Count		Cum.	Average	Expected	OUTFIT		
Score	Used	%	%	Measure	Measure	MnSq	Measure	S.E.
6	226	9%	100%	3.54	3.39	0.8	3.7	0.08
5	455	19%	91%	2.7	2.59	0.7	2.29	0.06
4	523	22%	72%	1.68	1.67	0.7	1.09	0.06
3	504	21%	50%	0.51	0.54	0.6	-0.58	0.08
2	326	14%	29%	-0.89	-0.91	0.6	-2.16	0.1
1	243	10%	15%	-3.08	-2.84	0.5	-4.34	0.13
0	123	5%	5%	-4.52	-4.37	0.9		

Figure 3 below presents the Subscale / Marker map. The red lines indicate where the *Rasch-Andrich Threshold* values occur, and where level cut scores might be drawn for the HKCE scales in relation to the data obtained from the Carnegie analysis.

Figure 3: Subscale / Marker map



It will be seen from the map above that the subscales for all three markers fall in a comparatively restricted range. The outlier is the CS AWA Task Response subscale, which appears to be very lenient in its awarding of grades. This leniency echoes the discussion of mismatching candidates above, with the candidate presented as an illustration receiving a level '6' grade for Task Response despite having only written 109 words. This is an issue that might need to be revisited at a later date.

To conclude this section, an extrapolation is presented from Table 7 and Figure 3 above. Table 8 below presents the lowest CS value for each HKCE scale level.

Table 8: Match between HKCE scales and Carnegie IELTS levels

HKCE scale	Lowest CS values	Rasch-Andrich Threshold
6	7.5	3.7
5	6.95	2.29
4	6.65	1.09
3	6.35	-0.58
2	5.8	-2.16
1	4.8	-4.34

Taking the lowest CS value as a cut score, Table 8 illustrates how the two scales would match. A CS / IELTS-type score of between 4.8 to 5.79 would equate to a Level 1 in HKCE terms. Between 6.95 and 7.49 would mean an HKCE grade of 5. A CS score of 7.5 or above would be an HKCE 6.

Given that LC has six levels of test, consideration needs to be given as to how output from the CS AWA system may be recalibrated to match LC levels. There are two possible ways forward.

The first is that there are accepted levels of correspondence between the nine IELTS levels, the CEFR, and hence by extension, to LC's interpretation of the CEFR levels in its own tests. Since the CS AWA correlates generally highly with established human scores, for expediency's sake, it would be acceptable to initially make use the IELTS > CEFR > LC test level correspondence.

A longer term and possibly a more defensible methodological approach involves using a recalibration of LC's Writing tests against the common scale being established for all LC tests. Calibration of the LTE listening and reading item bank has been conducted. The next step will be to calibrate the Writing test against this scale. Once this has been conducted, a picture will be available of the length of the LC A1 to C2 Writing test scale. At this point, the methodology presented above may be used to map *Rasch-Andrich Threshold* values from the CS AWA system to LC's Writing scale.

Conclusion

From the analysis above, it can be seen that the scores produced by the two sets of markers – the human raters and the CS AWA system – are reliably close. Further, the MFRA methodology has illustrated how the scores produced on the two different datasets – the six-point HKCE scale and the nine-point CS AWA / IELTS scale – may eventually be aligned.

Notes

1. There have clearly been errors in some of the transcriptions in the Hong Kong dataset, it should be noted, but then that should have perhaps have resulted in an even lower grade being returned by the CS AWA system.

References

- Alderson, J.C. (2000). Technology in testing: The present and the future. *SYSTEM*, 28, 593-603.
- Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). New York: Routledge.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3). Available at: <http://escholarship.bc.edu/jtla/vol4/3/>.
- Bennett, R.E., & M. Zhang. 2016. Validity and automated scoring. In *Technology and Testing*. In F. Drasgow, (ed.) Routledge: New York.
- Burstein, J. (2003) The e-rater scoring engine: Automated essay scoring with natural language processing. In: Shermis, M. D. and Burstein, J. (eds.), *Automated essay scoring: A crossdisciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 113–122.
- Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, 19, 273-299.
- Chapelle, C. A. & Douglas, D. (2006). *Assessing Language through Computer Technology*. Cambridge, UK: Cambridge University Press
- Chung, G.K., & O'Neil, H. F., Jr. (1997). Methodological Approaches to Online Scoring of Essays. ERIC Document Reproduction Service No. ED 418 101.
- Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics* 20, 147–161.
- Coniam, D. 2009. Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCALL Journal*, 21(2): 259-279.
- Drechsel, J. (1999) Writing into Silence: Losing Voice with Writing Assessment Technology. *Teaching English in the Two-Year College*, 26(4): 380–387.
- Evelyn, H., & Lazaraton, A. 1991. *The Research Manual: Design and Statistics for Applied Linguistics*. New York: Newbury House.\
- Foltz, P.W., Kintsch, W. & Landauer, T.K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, , 25(2&3), 285-307.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.
- Lee, Y-W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3), 391–417.
- Linacre, J. M. (2010). *FACETS, XX* [Computer program]. Chicago, IL: Winsteps.com.
- Linacre, J. M. (2012). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.
- Lu, X. 2017, Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, 34(4): 493-511.

- Rudner, L. M. & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment*, 1(2).
- Shermis, M.D. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20: 53-76.
- Shermis, M.D.,& Burstein, J.(2003). Automated essay scoring: A cross-disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, 31, 57-71.
- Yannakoudakis, H., Andersen, O., Geranpayeh, A., Briscoe, T., & Nicholls, D. (2018) Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3): 251-267.

Chapter 14: Towards a Communicative Test of Reading and Language Use for Classical Greek

Polyxeni Poupounaki-Lappa, Tzortzina Peristeri
and David Coniam

[*Poupounaki-Lappa, P., Peristeri, T., & Coniam, D. (2021). Towards a communicative test of reading and language use for Classical Greek. Journal of Classics Teaching, 22 (44).*]

Abstract

This paper describes the development of a communicative test of Reading and Language Use for Classical Greek, aimed at students at CEFR (Common European Framework of Reference for Languages) levels A1 and A2. A discussion is first provided of traditional pedagogical approaches which have for many decades dominated the teaching of classical languages, followed by suggestions why these may be supplanted with more modern communicative approaches. Focus then moves to assessment, where, it is suggested, methods are equally rooted in traditional, form-focused methods. If teaching is to become more communicative, it is argued, so should assessment. Against this backdrop, the development of a test of Reading and Language Use for students of Classical Greek at CEFR levels A1 and A2 is described.

Key words: Classical Greek, communicative approach, assessment, reading and language use, CEFR levels

Introduction

Calls are increasingly being voiced by educators and teachers of classical languages for a more communicative, and humanistic approach to the teaching of classical languages – see e.g., Hunt (2021), Lloyd (2021), Manning (2021). Against this background of proposed changes in pedagogy, this paper proposes parallel changes in assessment, because assessment impacts pedagogy (see Rind *et al*, 2019, Cheng, 2005), arguing the case for

a communicative test of reading and language use for learners of Classical Greek. Discussion first explores the extent to which Classical Greek currently features in European schools and the traditional – i.e., analytical rather than communicative – pedagogies by which it is predominantly taught.

For a number of reasons, such as teacher interest in communicative approaches to language learning, learner interest and motivation, as well as efforts to improve uptake by schools, a considerable number of academics and practitioners have been advocating that Classical Greek be taught in a similar manner to modern languages: that is, based more around communicative approaches to language teaching – see e.g., Richards & Rodgers (2014). Following an exploration of a communicative approach in teaching, the current paper then moves to an examination of assessment in classical languages, where there would appear to be even less use of modern methods and where traditional, grammar-translation approaches to assessment predominate.

Following this background, discussion shifts to the development of a communicative test of reading and language use for Classical Greek, to be used with international students. An outline of tests of Reading and Language use of Classical Greek at CEFR levels A1 and A2 is provided, illustrating how curriculum development adheres to communicative CEFR principles. The two tests have been developed so that, following trialling, they may be calibrated together to a single CEFR-linked scale.

Uptake of Classical Greek as a Subject in the School Curriculum

Classical Greek (or 'Ancient Greek') remains a key feature of many education systems. Besides being a compulsory secondary education subject in Greece, it is quite widely taught in many countries in Europe and around the world. This section briefly explores its reach.

In Italy, Classical Greek is a compulsory subject for the 'Liceo Classico' high schools (ages 14–19) and an estimated 6.7 percent of Italian students study Classical Greek over a period of five years (statista, 2021).

In the Netherlands, all classical secondary school 'gymnasia' students study Latin and Classical Greek for three years (ages 12-15) with an average of two to three teaching hours a week per language. Students can then opt to focus on one of the two languages for another three years, with an average of four to six teaching hours a week (van Bommel, 2016).

In Germany, Greek is taught in many 'gymnasia' (grammar schools). Additionally, students aged 14 and over in 200 high schools are able to opt to learn Classical Greek as a third foreign language (Chrysopoulos, 2016; fonien, 2016).

In the UK, Classical Greek is taught at elite secondary schools, with a small, if consistent, number of candidates each year sitting Year 11 GCSEs and Year 13 A levels – see Taylor (2003) for a discussion.

Classical Greek is also taught in Uruguay, where there is a substantial Greek community (ellines, 2018).

While the study of Classical Greek is viewed in some quarters as having little relevance to modern society, other scholars are more positive, discussing how the uptake of Classical Greek may be furthered (Foster, 2015). Gibbs (2003), for example, writes about the outlook for the teaching of classics being 'more auspicious' than in

previous decades. Bracke (2015) discusses the teaching of the subject in primary schools in Wales. Hunt (2018) describes movements to increase – with some success – the uptake of the classics in UK schools in the period 2010–15. Holmes-Henderson *et al.* (2018) review ways of increasing the uptake and reach of the classics.

Teaching Methods and Materials

Teaching methods for classical languages such as Latin and Greek have been typically traditional – largely *grammar translation* (Nielson, 2018). Under such a methodology, the majority of the instruction is conducted through the medium of the students' first language, where the focus is on reading texts, or 'more typically, translating them' (Gruber-Miller, 2006).

In contrast, the essence of a 'communicative' approach to language teaching (e.g., Richards & Rodgers, 2014) may be summarised as a focus on communication, rather than simply on form, and on the communicative needs of learners, rather than solely focusing on linguistic systems such as grammar.

Given that an accepted tenet of the teaching of modern languages is that communication and language use should be at the forefront of any pedagogical approach, it is instructive to consider why grammar translation as a methodology predominates in the field of classics education. In the UK, where classical languages are taught in a comparatively small number of elite schools, subjects such as Latin and Classical Greek have been considered essentially tests of intellectual ability. In many instances, the teaching of Classical Greek is limited to analysing the original text in terms of grammar, syntax and form – with the focus being on translating the text in students' L1 and engaging in a philological analysis. In this context, grammar translation as a toolkit for language study and analysis has been viewed as the vehicle to access meaning and to establish understanding of the written message. Against this backdrop – and with the aim of the teaching of modern foreign languages being *communication* – it is perhaps not surprising that teaching methods underpinning classical and modern languages have diverged. Nonetheless, a strong call is currently being made that, if the teaching of classical languages is to have a degree of relevance in the modern world, and for students to want to study them, more communicative-focused approaches to language learning and teaching need to be adopted. This argument is exemplified in the volume by Lloyd & Hunt (2021), where the case for communicative approaches to the teaching of classical languages is clearly made, and which resonates strongly through the work of many educators and practitioners, a number of whom are cited below.

The lack of a focus on communicative competence has contributed to making the teaching of the classics cognitively rather demanding and this has led to a preponderance of dull, dry teaching materials and examinations, which bear little relevance to any form of communication and so have limited relevance to the learner; see e.g., Taylor (2003). Gruber-Miller (2006), on the issue of communicative competence, argues that if grammar is taught to the exclusion of communication skills, the result is students who have 'a limited ability to comprehend and produce complex discourse fluently and accurately'.

Major (2018) argues in a similar vein in his discussion of the *Standards for Classical Language Learning* (2017) in the USA, a document promoting the teaching of classical languages for all age groups and educational levels. Given that a major aim of the *Standards* is explicitly stated as *communication*, Major (2018), summarising much that is deficient with relation to the dominant approaches and pedagogical resources for teaching Classical Greek, states:

Greek language teaching and pedagogical support materials are woefully out of sync with interest in the language outside the academy (Major, 2018, p.55)

Major (2018) views the *Standards for Classical Language Learning* (2017) as being a force for helping to reorient the priorities and focuses of Greek language classes in the USA in ways that 'both correspond to broader interest and result in improved language comprehension' (Major, 2018, p.55).

With reference to the 'broader interest' issue alluded to above, it is interesting to note that while modern authors rarely write in Classical Greek, there are instances of modern-day attempts to further the reach of Classical Greek: Jan Křesadlo has written poetry and prose in a Classical Greek style, and versions of Harry Potter and Asterix are also available in Classical Greek. This resonates with Pettersson & Rosengren's (2021) work in attempting to increase learner interest in Latin by creating modern resources in Latin.

There has been, it should nonetheless be noted, some innovation in the teaching of Classical Greek. Moore (2013), for example, outlines the use of song in the Greek classroom and how this may be used to motivate student interest. Bayerle (2013) describes the use of team-based learning to generate interest and activity in his Classical Greek classes. Dvorsky-Rohner (2008) describes the use of gaming exercises and strategies to engage and motivate learners of Classical Greek. Hill (2021) outlines the operation of a 'conventiculum' – all-day immersion activities – in classical Greek.

Assessment

In the field of modern languages, mirroring the changes in the approaches to teaching, assessment has seen similar shifts: with moves to more communicative and relevant forms of assessment. Paltridge (1992) frames one of the key aims of communicative language testing as measuring how well a test taker is able to perform 'real life' language tasks, or activities. As Bachman and Palmer (2010) outline, the validity of a language test is predicated on how much test scores reflect what test takers can do in a language. The corollary of this is that assessment tasks will bear some relevance to the real world, and not be solely tests of grammar.

As outlined above, the teaching of Classical Greek is gaining some momentum towards a more communicative outlook. In contrast, assessment remains very traditional. An examination of the UK's A level Classical Greek Examinations, for example, reveals a heavy focus on translation and the explication of grammatical rules. Assessment at college level also reflects a heavily analytical focus. Watanabe (2010), for example, describes a Greek standardised multiple-choice examination taken by students from 24 American colleges and universities. The college Greek examination outlined by Watanabe does not assess any 'communicative constructs' (Harding, 2014), with all test items being discrete-point tests of the forms of noun cases, verb mood and tense, followed by similar questions on a short reading passage. The focus is purely on form (Mahoney, 2004); there is very little in the test which addresses any of the communicative forces outlined by Gruber-Miller (2006), or which have a place in the *Standards for Classical Language Learning*.

The Development of a Communicative Test of Reading and Language Use in Classical Greek

The test outlined below attempts to address some of the issues created by the lack of ‘communicativeness’ in existing Classical Greek tests. To be truly communicative, a test should ideally assess all language skills (see Mitchell *et al.*, 2019). As a first, albeit limited, step in this direction, the test developed assesses Reading and Language Use, with the test seen as a qualification suitable for young people or adults who intend to apply for higher education or professional employment. The test is intended to be calibrated to the Common European Framework of Reference (CEFR) at levels A1 and A2 [Note 1]. Development has taken place based on the CEFR manual (Council of Europe, 2001) with its detailed specifications guiding the development of tests, and addressing what is assessed, how performance is interpreted, and how comparisons in achievement may be made. The test has been developed by LanguageCert, an Ofqual-regulated awarding organisation offering globally-recognised language qualifications, whose purview is the assessment of modern languages in communicative contexts.

The following section presents an overview of the A1 and A2 level tests, with reference to test specifications covering both grammatical and functional aspects of communicative language ability. These specifications were produced over an extended period by a group of Classical Greek academics and teachers. The detailed set of specifications and associated official practice material is available in the *LanguageCert Test of Classical Greek (LTCG) Qualification Handbook* for the examination (LanguageCert, 2021). The examples provided below are drawn from this *Qualification Handbook*.

There are three key components to the test specifications, as per the CEFR Manual: reading and vocabulary subskills; topics; and grammatical/syntactic structures.

Reading is specified in terms of three components:

1. Reading subskills
2. Vocabulary range features
3. Text structure subskills

Figure 1 below presents a snapshot of these elements. The exemplars in the figures below are drawn from the A1 level test.

Figure 1: Reading skills (from LTCG Qualification Handbook)

<p>Reading subskills</p> <ul style="list-style-type: none"> • understand very short simple narratives and descriptions • recognise the purposes of short texts where the purpose and intended audience is clear • understand viewpoints if made clearly and simply <p>Vocabulary range features</p> <ul style="list-style-type: none"> • understand very familiar words and phrases in simple short text • understand isolated words, short simple phrases and grammatical structures that link clauses and help identify time reference <p>Text structure subskills</p> <ul style="list-style-type: none"> • understand the organisational, lexical and grammatical features of short simple texts • recognise different purposes of simple texts
--

A range of topics are covered relevant to test takers in the modern world. Figure 2 presents a sample.

Figure 2: Topics (from LTCG Qualification Handbook)

<p>A1</p> <ul style="list-style-type: none"> • personal identification • house and home, environment • daily life • free time, entertainment • activities • travel • relations with other people • health and bodycare • food and drink • places • weather • measures and shapes • education 	<p>A2 (building on A1 specs)</p> <ul style="list-style-type: none"> • measures and shapes • health and bodycare • food and drink
--	--

The most extensive part of the syllabus, nonetheless, as is perhaps to be expected, relates to the specification of grammatical, lexical and syntactic features. This came as a requirement for beginner levels in order to guide test takers and educators in organising their study prior to taking the exam. Key elements covered in the specifications are laid out in Figure 3 below.

Figure 3: Grammatical and syntactic categories (from LTCG Qualification Handbook)

- alphabet
- syllables – accentuation
- parts of speech / syntax
- verb forms
- nouns
- pronouns
- prepositions
- articles
- adjectives
- infinitives
- participles

A fuller set of specification is provided in Appendices 1 and 2. These are presented in two columns, one for the A1 level test, and a second for the A2 level test (which subsumes the A1 component).

In line with Council of Europe principles and practice, CEFR-linked tests link with communicative ‘Can-do’ statements about language use. Figure 4 presents a sample.

Figure 4: Can-do statements (from LTCG Qualification Handbook)

- I can recognise familiar words and very basic phrases in short texts.
- I can understand simple information if there is visual support.
- I can understand the main points and locate specific information in short simple texts on familiar matters.
- I can work out the probable meaning of unknown words from the context.

The use of ‘Can-do’ statements relates to the functional use of linguistic resources, and functional competence (Council of Europe, 2001). Tests aimed at lower-competency levels – as with the current A1 and A2 level tests – show an emphasis on finite elements of language form. That said, it can be seen how test parts go beyond a mere form analysis and require a fuller understanding of meaning conveyed. Part 2 focuses on relating meaning to a given image; in Parts 3 and 4, the textual nature of the assessment goes beyond the assessment of grammar and calls for inferencing meaning utilising cohesion and coherence devices of the transmitted message.

Test Components

The A1 and A2 tests both consist of four parts, each using distinct task types to assess specific sub-skills. Figure 5 elaborates.

Figure 5: Classical Greek task types (from LTCG Qualification Handbook)

Part 1: 10 items: Multiple matching	(images and words)
Part 2: 10 items: True/False	(statements and visuals)
Part 3: 10 items: Multiple-choice cloze	(gapped text)
Part 4: 10 items: Multiple matching	(gapped text)

A sample paper for the A2 level Classical Greek test is provided in Appendix 4. Further detail and sample papers – official practice material – for both A1 and A2 levels are available at <https://www.languagecert.org/en/language-exams/classical-greek/languagecert-test-of-classical-greek/languagecert-test-of-classical-greek-a1-ltcg-reading-and-language-use>.

While the tests follow communicative principles and are a mixture of authentic and level-adapted materials, the tests at times draw on classical texts that embody the spirit of Greek writers and their works (see, for example, the University of Oxford's <https://www.classics.ox.ac.uk/preliminary-examination-classics-and-english#/>). This use of classical texts echoes the well-accepted practice of the use of literature in the language classroom (Daskalovska & Dimova, 2012).

Trialling

To ensure quality and credibility, in any test development, trialling is a key element in the process which has to be conducted before a test may be formally offered. The trialling of the A1 and A2 level tests is planned for mid-2021 with each test being administered to a group of test takers identified as being at the appropriate level. The two tests have a common section so that they may be calibrated together, with the materials used in the two tests placed on a common difficulty scale. Test takers will also complete a set of 'Can-do' statements (a sample of which was presented in Figure 4 above), self-assessing their ability in Classical Greek. This will provide triangulated feedback to enable greater validity to be attached to the results obtained from the analysis and the calibration of the two tests together.

Notes

1. The Council of Europe's Common European Framework of Reference (CEFR) has played a decisive role in the teaching and setting standard for initially European languages. The CEFR organises language proficiency in six levels, A1 to C2. These can be regrouped into three broad levels: Basic User, Independent User and Proficient User, with levels defined through 'can-do' descriptors. See <https://www.coe.int/en/web/common-european-framework-reference-languages/illustrations-of-levels>.

References

- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Bailey, J.S. (2021). Communication in all modes as efficient preparation for reading a text. In Lloyd, M.E., & Hunt, S. (eds.) *Communicative Approaches for Ancient Languages*. London: Bloomsbury.
- Bayerle, H. (2013). Team-based learning to promote the study of Greek. *Teaching Classical Languages*, 5(1), 1-17.
- Bracke, E. (2015). Bringing classical languages into a modern classroom: Some reflections. *Journal of Classics Teaching*, 16(32), 35-39.
- Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge: Cambridge University Press.
- Chrysopoulos, P. (2016). German Junior High School Students Study Ancient Greek. Available online: <https://greekreporter.com/2016/12/19/german-junior-high-school-students-study-ancient-greek/>
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg Cedex, France: Council of Europe.
- Daskalovska, N., & Dimova, V. (2012). Why should literature be used in the language classroom? *Procedia-social and Behavioral Sciences*, 46, 1182-1186.
- Dvorsky-Rohner, D. (2008). Gaming in beginning Greek: Taking advantage of the six weeks' opportunity. *Teaching Classical Languages*, 4(1), 15-29.
- ellines. (2018). Uruguay is a fan of Greek culture. Available online: (<https://www.ellines.com/en/good-news/39036-uruguay-is-a-fan-of-greek-culture/>).
- Fonien. (2016). Ancient Greek is "cool" in German schools. Available online: Ancient Greek is "cool" in German schools–fonien.gr.
- Foster, F. (2018). Attitudes to ancient Greek in three schools: A case study. *The Language Learning Journal*, 46(2), 159-172.
- Gibbs, M. (2003). The place of classics in the curriculum of the future. In Morwood, J. (ed.) *The teaching of classics*. Cambridge: Cambridge University Press.
- Gruber-Miller, J. (2006). Communication, context, and community. Integrating the standards in the Greek and Latin classroom. In Gruber-Miller, J. (Ed.) *When dead tongues speak: Teaching beginning Greek and Latin*. New York: Oxford University Press.
- Gruber-Miller, J. (2018). The Standards as integrative learning. *Teaching Classical Languages*, 9(1), 19-38.
- Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, 11(2), 186-197.
- Hill, R.S. (2021). A conveitculum for speakers of Ancient Greek: The Lexington Σύνοδος Ἑλληνική. In Lloyd, M. E., & Hunt, S. (eds.) *Communicative Approaches for Ancient Languages*. London: Bloomsbury.
- Holmes-Henderson, A., Hunt, S., & Musié, M. (eds.). (2018). *Forward with classics: Classical languages in schools and communities*. London: Bloomsbury.
- Hunt, S. (2018). Getting classics into schools? Classics and the social justice agenda of the coalition government 2010-2015. In Holmes-Henderson, A., Hunt, S., & Musié, M. (eds.). *Forward with classics: Classical languages in schools and communities*. London: Bloomsbury.

- Hunt, S. (2021). Active Latin teaching for the inclusive classroom. In Lloyd, M. E. & Hunt, S. (eds.) *Communicative approaches for ancient languages*. London: Bloomsbury.
- LanguageCert.org. (2021). Exam Information. Available at: <https://www.languagecert.org/en/language-exams/classical-greek>.
- Lloyd, M. E., & Hunt, S. (eds.) (2021). *Communicative approaches for ancient languages*. London: Bloomsbury.
- Lloyd, M. E. (2021). Exploring communicative approaches for beginners. In Lloyd, M. E., & Hunt, S. (eds.) *Communicative approaches for ancient languages*. London: Bloomsbury.
- Mahoney, A. (2004). The forms you really need to know. *Classical Outlook*, 81, 101-105.
- Major, W. (2018). Recontextualizing the teaching of ancient Greek within the new standards for classical language learning. *Teaching Classical Languages*, 9(1), 54-63.
- Manning, L. (2021) Active Latin in the classroom: Past, present and future. In Lloyd, M. E., & Hunt, S. (eds.). *Communicative approaches for ancient languages*. London: Bloomsbury.
- Mitchell, R., Myles, F., & Marsden, E. (eds.) (2019). *Second language learning theories*. New York: Routledge.
- Moore, T. (2013). Song in the Greek classroom. *Teaching Classical Languages*, 4, 66-85.
- Morwood, J. (ed.). (2003). *The teaching of classics*. Cambridge: Cambridge University Press.
- Nielson, M. (2018). *The grammar-translation method and the communicative approach: Combining second language acquisition approaches to teach Lucan and Statius in high school*. Doctoral dissertation. Arizona: The University of Arizona.
- Paltridge, B. (1992). EAP placement testing: An integrated approach. *English for Specific Purposes*, 11(3), 243-268.
- Pettersson, D., & Rosengren, A. (2021). The Latinitium Project. In Lloyd, M. E., & Hunt, S. (eds.). *Communicative Approaches for Ancient Languages*. London: Bloomsbury.
- Richards, J., & Rodgers, T. (2014). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.
- Rind, I., & Mari, M. (2019) Analysing the impact of external examination on teaching and learning of English at the secondary level education, *Cogent Education*, 6, 1.
- Statista (2021). Share of students enrolled at upper secondary schools in Italy for the academic year 2019/2020, by type of lyceum. Available online: <https://www.statista.com/statistics/572827/share-of-enrollment-in-upper-secondary-schools-italy-by-type-of-school/>
- Taylor, J. (2003). Learning Greek. In Morwood, J. (ed.) *The teaching of classics*, 95-105. Cambridge: Cambridge University Press.
- Van Bommel, B. (2016). Classics between prosperity and crisis: Greek and Latin education in 21st-century Holland. Available online: <https://www.addisco.nl/classics-between-prosperity-and-crisis-greek-and-latin-education-in-21st-century-holland/>.

Appendix 1: LTCG Reading Subskills and Topic Specifications (from *LTCG Qualification Handbook*)

A1	A2
<p>Reading subskills</p> <ul style="list-style-type: none"> understand very short simple narratives and descriptions find and obtain specific, predictable information in simple texts recognise the purpose of short texts where the purpose and intended audience is clear. understand viewpoints if made clearly and simply. <p>Vocabulary range features</p> <ul style="list-style-type: none"> understand very familiar words and phrases in simple short text understand isolated words, short simple phrases and grammatical structures that link clauses and help identify time reference understand the meanings conveyed by capital letters and full stops in very simple sentences. identify time reference (present-future) <p>Text structure</p> <ul style="list-style-type: none"> understand the organisational, lexical and grammatical features of short simple texts 	<p>Reading subskills</p> <ul style="list-style-type: none"> understand simple narratives and descriptions on familiar topics recognise the different purposes of text provided the purpose and intended audience is clear locate specific predictable information in everyday short texts on familiar matters understand a simple line of argument simply expressed understand the main ideas and gist of simple narratives <p>Vocabulary range features</p> <ul style="list-style-type: none"> recognise high frequency words and words with common spelling patterns in simple texts understand punctuation and capitalisation used in simple and compound sentences recognise key grammatical forms such as noun declensions, verb tenses, infinitives and participles identify time reference (past-present-future) in short simple narratives and descriptions <p>Text structure</p> <ul style="list-style-type: none"> recognise different purposes of simple texts

A1	A2
Topics <ul style="list-style-type: none">• daily life• education• food and drink• free time, entertainment• health and bodycare• house and home, environment• language• measures and shapes• personal identification• places• relations with other people• travel• weather	Topics (building on A1 specs) <ul style="list-style-type: none">• height• length• weight• capacity• personal comfort• fitness, nutrition• eating and drinking out

Appendix 2: LTCG Grammar and Syntax Specifications (from *LTCG Qualification Handbook*)

	A1	A2 (building on A1 specs)
Alphabet	<ul style="list-style-type: none"> • letters of Classical Greek 	
Syllables – accentuation	<ul style="list-style-type: none"> • syllables • accentuation 	
Parts of speech / syntax	<ul style="list-style-type: none"> • inflected parts of speech • uninflected parts of speech 	<ul style="list-style-type: none"> • subject – verb – object • predicate
Verb forms	<ul style="list-style-type: none"> • present reference: thematic, athematic verbs • present tense, active/middle voice, indicative of Εἰπί, Λύω and common regular verbs • future reference • future tense, active/middle voice, indicative of Εἰπί, Λύω and common regular verbs • future tense of athematic verbs 	<ul style="list-style-type: none"> • past reference • imperfect tense: syllabic, temporal augment • imperfect tense, active/middle voice, indicative of Εἰπί and common verbs • aorist tense: syllabic, temporal augment • aorist tense, active/middle voice, indicative of Εἰπί and common verbs • perfect tense: reduplication, temporal augment • perfect tense, active/middle voice, indicative of common verbs
Nouns	<ul style="list-style-type: none"> • gender • number • case • 1st, 2nd and specific 3rd declension nouns • irregular nouns 	<ul style="list-style-type: none"> • irregular nouns
Pronouns	<ul style="list-style-type: none"> • personal pronouns – verb subject 	<ul style="list-style-type: none"> • interrogative pronouns • indefinite pronouns
Prepositions	<ul style="list-style-type: none"> • prepositions 	
Articles	<ul style="list-style-type: none"> • definite article 	
Adjectives	<ul style="list-style-type: none"> • 2nd declension: two-ending, three-ending adjectives • 3rd declension: two-ending, three-ending adjectives 	<ul style="list-style-type: none"> • 3rd declension: nasal and liquid adjectives
Infinitives	<ul style="list-style-type: none"> • active/ middle voice of present tense of infinitives 	<ul style="list-style-type: none"> • active/ middle voice of future tense and first aorist of infinitives
Participles	<ul style="list-style-type: none"> • active/ middle voice of present tense of participles 	<ul style="list-style-type: none"> • active/ middle voice of future and first aorist tense of participles

Appendix 3: Sample Can-do Statements for CEFR Levels A1 and A2 (from *LTCG Qualification Handbook*)

A1	A2 (building on A1 specs)
<ul style="list-style-type: none"> I can recognise and read familiar words. I can read short simple texts and understand simple information. I can get an idea of the content of simpler informational material and short simple descriptions. I can control a few simple grammatical structures and syntax patterns in a learnt repertoire. 	<ul style="list-style-type: none"> I can understand the main points and locate specific information in short simple texts on familiar matters. I can recognise high-frequency words and words with common spelling patterns on familiar topics. I can work out the probable meaning of unknown words from the context. I can understand texts describing people, places, everyday life, and culture, etc., provided that they are written in simple language.

Appendix 4: A2 Classical Greek Sample Paper

<p>LanguageCert A2 Test of Classical Greek Practice Test A2</p> <p>Candidate's name (block letters please) _____</p> <p>Centre no _____ Date _____</p> <p>Time allowed: • Reading and Language Use 80 minutes</p> <p>Instructions to Candidates • An Answer Sheet will be provided. • All answers must be transferred to the Answer Sheet. • Please use a soft pencil (2B, HB).</p>	<p style="text-align: center;">Language Cert</p> <p style="text-align: right;">Practice Test A2</p> <p>Part 1 Look at the images below. Match the images (A–J) to the words.</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  A </td> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  B </td> </tr> <tr> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  C </td> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  D </td> </tr> <tr> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  E </td> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  F </td> </tr> <tr> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  G </td> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  H </td> </tr> <tr> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  I </td> <td style="width: 50%; text-align: center; padding-bottom: 10px;">  J </td> </tr> </table> <p>LanguageCert Test of Classical Greek Reading and Language Use (A2) Copyright © 2020 2</p>	 A	 B	 C	 D	 E	 F	 G	 H	 I	 J
 A	 B										
 C	 D										
 E	 F										
 G	 H										
 I	 J										
<p>Practice Test A2</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; vertical-align: top; padding-right: 10px;"> <p>1. λέων</p> <p>2. κοχλίας</p> <p>3. παις</p> <p>4. ὄρκτος</p> <p>5. ὄρος</p> <p>6. δάσος</p> <p>7. ὄπε</p> <p>8. θύμω</p> <p>9. σφύριχος</p> <p>10. σάλπιξ</p> </td> <td style="width: 50%; vertical-align: top; padding-left: 10px;"> <input type="checkbox"/> <input type="checkbox"/> </td> </tr> </table>	<p>1. λέων</p> <p>2. κοχλίας</p> <p>3. παις</p> <p>4. ὄρκτος</p> <p>5. ὄρος</p> <p>6. δάσος</p> <p>7. ὄπε</p> <p>8. θύμω</p> <p>9. σφύριχος</p> <p>10. σάλπιξ</p>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<p>Practice Test A2</p> <p>Part 2 Look at the image and then read the statements. Choose the correct answer True or False for each statement according to the image.</p> 								
<p>1. λέων</p> <p>2. κοχλίας</p> <p>3. παις</p> <p>4. ὄρκτος</p> <p>5. ὄρος</p> <p>6. δάσος</p> <p>7. ὄπε</p> <p>8. θύμω</p> <p>9. σφύριχος</p> <p>10. σάλπιξ</p>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>										

Practice Test A2

1. Μα γυνή στάση τήν γῆν.

2. Πεις προσέρχεται κρατῶν ισχίδος.

3. Ο ἄγρος γέμει ἔλαιον.

4. Ιέραι διστασί επί τη δινδρον.

5. Γιαρακιμένη κρήνη δινει υδατός εστι.

6. Εν τῇ εἰσόδῳ τοῦ σπηλαίου κύων καὶ γολῆ κιντα.

7. Ο καρός αθηρός εστι.

8. Οι γεωργοι τὰς ἔλαιας ραβδίζουσιν.

9. Εἰς ἀνήρ κεθύνει ύπο πνος δινδρου.

10. Τιες ἐργάται ἀριστοποιοῦνται.

Practice Test A2

Part 3
Read the text. Choose the correct answers to complete the text.

Ο χειμών λίαν δεινή ὡρα ἔστι. Τούτη γάρ πολλὴ (1) τούς ὅριους καλύπτει. Εν τῷ χειμῶνι καὶ οἱ γεωργοὶ καὶ οἱ πομπέες ἴσχυραν ὄνται. Οὔτε τοὺς ἄγρους μετεπιφύονται διὸ την πυνηνήν γένον, (2) ἢ προδότα νέμουσιν, ὃν προφήτης ὑπέστη ἐπί τοῖς ἀγροῖς. Οἱ δὲ ναΐται τὰ πτολεῖα εἰς τοῖς (3) οὐ μόνον ἀγούσαι ἀλλα καὶ ενταῦθα (4) μεριχοῖ τοῦ ερεός. Οὐ γάρ εστιν ἥρεος η δάδατα καὶ ἄλλα ὡραῖα καὶ διὰ τοὺς χαμένας. Εν ταῖς οἰκίαις οἱ ἀνθρώποι μέν ἔνια κανουσι καὶ τὸ πῦρ μπονούν. ἔξω (5) τοῦ οἴκου πικνούς γῆτανος φέρουσι καὶ τάς κεφαλὰς καλύπτουσιν. Πλεκάδες δὲ αἱ κατ' ἔγραν δέσαι μεταὶ γιδῶν εἰσὶ καὶ οἱ ὑπέρβατοι οἱ δύνανται (6) Οἱ δὲ λάκαιοι (7) τὸν λημὸν ἐπὶ τὰς ἀγλαῖς τῶν προφήτων πίπουσι, οὓς οἱ πομπέες διώκουσιν καὶ πολλάκις πτονεύουσι. Οἱ οὐρανοί ἐν τῇ ὑδρίᾳ τοῦτη θρίβισαν (8) νεφῶν καὶ ὁ ἥλιος σπανιών λάμψει. Λόγιτε (9) οἱ ἀνθρώποι, νέοι καὶ γέροντες, τὸν χειμώνα βαρέωνς φέρουσι καὶ τό (10) μενοῦσι.

1. a) ψῦχος
-
- b) χιον
-
- c) λίθος

8. a) μενάλια
-
- b) μεγάλου
-
- c) μεγάλων

2. a) ἦ
-
- b) ἀλλά
-
- c) οὐτε

9. a) οὖν
-
- b) μάστε
-
- c) εἰ

3. a) λιμένα
-
- b) λιμένας
-
- c) λιμνός

10. a) θεος
-
- b) ἱερός
-
- c) τέρας

4. a) μένοντεν
-
- b) μένουσιν
-
- c) μένει

5. a) δέ
-
- b) οὐν
-
- c) γάρ

6. a) βαδίζειν
-
- b) σκάμπειν
-
- c) κτίζειν

7. a) ἐκ
-
- b) διδ
-
- c) ἐν

Practice Test A2

Part 4

Read the text. Choose the correct words to complete the text. There are two extra words you will not need.

Η χώρα τῆς Αττικῆς τιέψυκεν πλείστας προσδόους παρέχεσθαι. Όπισι δὲ γνωσθή ὅτι αἰληθῆς τούτο λέγω, πρώτον θιγητομοι τὸν φύσιν τῆς Αττικῆς. Αἱ μὲν ὥραι τοῦ ἑτού εἰσιν πρατόταται καὶ αὐτὰ τὰ γιγνόμενα μαρτυρεῖ τούτο: Πολλὴ φυτὰ σὲν βλαστάνειν (1) ἐνδέσ καὶ καρποφορεῖν. Ωστέρε δὲ ἡ γῆ, οὕτω καὶ ἡ θάλασσα περὶ τὴν χώραν (2) ἐστι. Καὶ δύσι οἱ θεοὶ ἀγαθὸς παρέχουσι, τοῦτα πάντα ἐνταῦθα πριψαίστα μὲν ἅρχεται, βραδέως δὲ (3) Οὐ μόνον δὲ κρατεῖ τοῖς ἐπὶ ἐνιαυτῶν θάλλουσι τε καὶ (4) , ἀλλὰ καὶ ἀΐδια ἄγαθα ἔχει ἡ χώρα. Λίθοι ἐν αὐτῇ (5) εἰσιν, ἔξιν κάλλιστα μὲν ναοι, κάλλιστα θέατρα (6) , εὐπρεπέστατα δὲ θεῶν ἀγάλματα: πολλοὶ ἔξι ἀλλιών χωρῶν θαυμάζουσιν (7) Ἐστι δὲ καὶ γῆ ἡ σπειρούμενη μὲν οὐ φέρει καρπόν, ὀρυπομένη δὲ πολλαπλασίους τρέψει ἢ εἰ σπον ἔφερε. Καὶ ὑπάρχυρός ἐστι σαφῶς θεῖα μοιρά. Οὐδεμία τῶν ἔγγυς (8) ἔχει τοσαῦτα ἀγαθά. Εὐλόγως τις ἀνένθησε οἰκείσθιοι τὴν πόλιν ἀμοι τὰ μέσα τῆς Ἑλλάδος καὶ πάσις δὲ τῆς οἰκαιμένης. Τσω γάρ ἄν τινες πλέον (9) αὐτῆς, τοσούτων χαλεπωτέρας ἡ μύχεσιν ἡ θάλαττας ἐντυγχάνουσιν. Καὶ περίρρυτος καὶ ἀμφιθάλαττος γαρ ἔστι, ὥσπερ (10)

A ὄρυπτωσιν

B μεγίστη

C ἀφθονοί

D ἀπέξχωσιν

E δύνανται

F νήσος

G λήγε

H γίγνονται

I γηράσκουσιν

J πλουσία

K πόλεων

L ταῦτα

Chapter 15: Recapping and Looking Ahead

Peter Falvey

Introduction

This chapter briefly recaps the LanguageCert approach to research, the research methodologies used and the products of that research. It also provides a flavour of on-going research and how this builds on the research presented in the current volume.

Recap of Current Volume

Four areas of research are covered in the current volume.

Section 1, *Background*, consisted of two chapters.

Chapter 1, *LanguageCert – A Multilingual Language Testing System*, provides a background introduction to LanguageCert's testing system, how it originated and developed, and its overall orientation towards assessment. Chapter 2, *External Validation of LanguageCert's English Language Examinations*, focuses on the all-important concept of validity and in this case, the part of validity that is essential for testing organisations—external validity.

Section 2, *Explorations Into Test Quality*, comprises four chapters, with each dealing with a facet of research into test quality.

Chapter 3, *Gauging Quality in the IESOL LanguageCert Listening and Reading Tests*, examines the reliability of LanguageCert's Listening and Reading tests. Chapter 4, *Examiner Quality and Consistency across LanguageCert Writing Tests* examine qualitative tests of writing in the context of examiner standardisation and training for quality, reliability and consistency. Chapter 5, *Potential Bias in LanguageCert IESOL Items: A Differential Item Functioning Analysis (DIF)*, describes the LanguageCert approach to identifying potential bias. Chapter 6, *Task Equivalence in LanguageCert IESOL Writing Tests*, examines the equivalence of test forms.

Section 3, *Calibration Studies*, consists of three chapters.

Chapter 7, *Validating the LanguageCert Test of English Scale: The Paper-based Tests*, describes the first stage of measurement scale development for the LanguageCert Test of English (LTE) through the validation of the initial LanguageCert Item Difficulty (LID) scale. Chapter 8, *Calibrating the LanguageCert Test of English Adaptive Test*, moved on from the developments described in Chapter 8, of the enhanced LID scale, to adapt that scale to the LTE computer adaptive test. Chapter 9, *Externally-Referenced Anchoring: Equating Expert Judgement and Rasch Measurement Values in LanguageCert IESOL English Language Tests*, explores test validation and the maintenance of integrity by the judicious use of Rasch Analysis in association with expert judgement.

Section 4, *Original Research*, consists of five chapters, describing a wide range of research carried out by the LanguageCert research team.

Chapter 10, *Identifying Guessing in English Language Tests via Rasch Fit Statistics: An Exploratory Study*, discusses how analytical measurement techniques can be used to investigate the perennial problem of guessing in multiple-choice tests. Chapter 11, *The Development and Delivery of Online-Proctored Speaking Exams: The Case of LanguageCert International ESOL*, describes the development and operation of LanguageCert's online proctoring of language examinations. Chapter 12, *Online Proctoring of High-Stakes Examinations: A Survey of Past Candidates' Attitudes and Perceptions*, presents the results of a survey which analysed the perceptions of a large sample of past candidates of LanguageCert's proctored online examinations. Chapter 13, *Automated Writing Assessment (AWA) and the Carnegie Speech Writing Assessment System*, describes an exploration into validating the use by LanguageCert of the Carnegie Speech Automated Writing Assessment system with LanguageCert examinations. Chapter 14, *Towards a Communicative Test of Reading and Language Use for Classical Greek*, describes the development by LanguageCert of a communicative test of Classical Greek at CEFR levels A1 and A2.

Looking Ahead

The research outlined above has helped to establish a number of research priorities which are currently being pursued and will be reported on in subsequent volumes.

Extending Current Agendas

Research confirming the reliability of LanguageCert Listening and Reading tests was outlined in Chapter 3. Such research will be conducted and reported on an ongoing basis. Analyses of Speaking and Writing will be incorporated into the reports.

The investigation of differential item functioning, reported on in Chapter 5, is taking place across a wide range of examinations. While important in themselves in relation to fairness they are also key to analyses aimed at identifying malpractice.

Externally-referenced anchoring discussed in Chapter 9 is an innovative technique that has been introduced to determine comparability of test forms across LanguageCert suite. The technique shows promise and will be

used as one of the approaches to confirm the comparability of tests in the LanguageCert suite. Further work in this area will be reported in subsequent volumes.

The communicative tests of Reading and Language Use for Classical Greek, outlined in Chapter 14 are being analysed with a view to calibrating them to a common scale and subsequently to the CEFR.

New Agendas

A number of comparative studies are underway, one of which is a large-scale project investigating the relationship between the LID scale and the China Standards of English (CSE). Both scales are referenced to the Common European Framework of Reference for Languages (CEFR) using data derived from the LanguageCert Test of English and a number of other measures. This important study aims, in due course, to align all LanguageCert tests to the CSE.

Over half of LanguageCert's language examinations are conducted using its secure online proctoring systems. Initial investigations suggest that these delivery modes are comparable but further studies are planned to further the effect, if any, of taking a pencil-and-paper test as opposed to one delivered through on line proctoring.

Large sample comparability studies are also underway with the LanguageCert Test of English to monitor any effect on candidates taking either the paper-based or adaptive test versions of the LTE.

The use of expert judgement is fundamental to the development of test materials. A number of studies are underway to investigate the reliability of expert rater judgements across tests in relation to the determination of item difficulty. This work is complemented by an investigation of the extent to which perception of difficulty may vary between native and non-native speaker expert raters.

Future volumes of this series will also contain more pedagogically-oriented, practitioner-focused papers. Topics investigated will include process writing in English language teaching; the value of conducting needs analyses with students at the start of the school year; and the use and value of adaptive tests in English language teaching.



Glossary: Statistical Techniques Used in The Volume

Peter Falvey

Much of this section is adapted from Coniam and Falvey (2018), Chapter 8. Its purpose is to provide an overview of the statistical terms and methods used throughout the volume. The chapter is designed to assist the reader who, otherwise, would encounter a large amount of duplicate explanation throughout the following fourteen chapters, all of which use a variety of statistical analytical tools as part of their research methodology.

Statistical Tools used in the analyses

This glossary describes the use made of Classical Test Statistics, Rasch measurement, Rasch models and quantitative and qualitative data analysis. It also discusses the concept of Frame of Reference.

Certain studies described in this book use Classical Test Theory (CTT) to analyse data – specifically survey data. While the use of CTT enables statistical significance to be examined, there are inherent weaknesses with CTT statistics. First, analytical techniques in CTT require linear, interval scale data input (Wright, 1997). Raw data collected through Likert-type scales, however, are usually ordinal since the categories of Likert-type scales indicate only ordering without any proportional levels of meaning. Applying conventional analysis on ordinal raw data can therefore lead to potentially misleading results (Bond and Fox, 2007; Wright, 1997). Second, CTT uses total score to indicate respondent ability levels. This results in person ability estimates being item-dependent; i.e., although person abilities may be the same, person ability estimates are high when items are easy but low when items are difficult. Similarly, item difficulty estimates are similarly sample-dependent; i.e., even though item difficulties themselves are invariant, item difficulty estimates appear high when respondents' competence is low but low when respondents' competence is high.

Classical Test Theory (CTT) – often called the “true score model” – assumes that every test taker has a true score on an item if it is possible to measure that score directly without error. CTT analyses assume, therefore, that a test taker's test score is comprised of a test taker's “true” score plus a degree of measurement error.

An overview of the CTT statistics used in the current set of studies will be briefly presented below. These can be grouped broadly into **Descriptive Statistics** (statistics that simply describe the group that a set of persons or objects belong to) and **Inferential Statistics** (statistics that may be used to draw conclusions about a group of persons or objects).

Descriptive statistics used in the studies are the **mean** (the arithmetical average), the **standard deviation** (the measure of variability in the dataset), and the **variance** (the average of the squared differences from the mean; the standard deviation squared, in effect.).

Inferential tests may be conceived of as either **parametric** or **non-parametric**. **Parametric data** has an underlying normal distribution – which allows for greater conclusions to be drawn since the shape can be described in a more mathematical manner. Other types of data are all **non-parametric**.

Parametric and Non-Parametric Tests

Parametric Tests

Parametric inferential statistical tests used in the case study have been the t-test, ANOVA and Pearson correlations. These will now be briefly described.

The T-Test

The t-test is used to compare two population means, with a view to determining if there is a significant difference between the means. There are two types of t-tests, **unpaired t-tests** (where the samples are independent of one another) and **paired t-tests** (where the samples are related to each other). A t-test is commonly used when the variances of two normal distributions are unknown and when an experiment uses a small sample size (a sample size of 30 subjects is used in the studies as being the threshold for conducting statistical analysis [Ramsey, 1980]).

ANOVA (Analysis of Variance)

ANOVA is used to compare differences of means among more than two groups. This is achieved by looking at variation in the data and computing where in the data that variation occurs (giving rise to the name 'ANOVA'). Specifically, ANOVA compares the amount of variation between groups against the amount of variation within groups.

The Pearson Product-Moment Correlation (PPM)

The Pearson correlation is an estimate of the degree of the relationship between two variables. The scale runs from -1 through 0 to +1, where +1 shows a total positive correlation, 0 indicates no correlation, and -1 shows a total negative correlation.

The inter-rater correlation is one application of the PPM, indicating the measure of agreement between raters of scale-based assessment. Interpretations of correlation magnitude differ. Friedrich (1999), for example, suggests that a correlation of 0.5 indicates a "moderate to strong tendency". Hatch and Lazaraton (1991, p. 441) suggest that a "strong" correlation, as regards inter-rater reliability, should be taken as 0.8. Following the example of Friedrich (1999) and Hatch and Lazaraton (1991), a correlation of 0.5 has been adopted in these studies to indicate a moderate correlation, one between 0.5 to 0.8 as moderate to strong, and a correlation above 0.8 as strong.

Non-Parametric Tests

The non-parametric inferential statistical test used in the case study has been the Chi-squared test.

The Chi-Squared Test

The Chi-squared test is used with *nominal* data (where the data fall into 'categories'; for example, male/female, or Likert scales in the current studies). The Chi-squared tests compare the counts of responses between two or more independent groups, and determine whether there is a significant difference between expected and observed frequencies in one or more category.

Significance

All the statistical tests described above – both parametric and non-parametric – provide a figure regarding the level of significance (the p-value) which emerged on the test. The p-value is the probability of the result occurring by chance or by random error. The lower the p-value, the lower is the probability that the event being measured can be explained by chance. A p value lower than 5% ($p<0.05$) is generally accepted as the threshold of statistical significance, although in many cases the 1% level ($p<0.01$) indicates a stronger case for arguing for significance (see Whitehead, 1986, p. 59). A p-value > 0.05 therefore suggests no significant difference between the means of the populations in the sample, indicating that the experimental hypothesis should be rejected. Over the past few decades there have been a number of controversies about the use/over-use of significance in data analysis. A useful overview is provided in Glaser (1999, p. 291-296) and Schneider (<https://arxiv.org/ftp/arxiv/papers/1402/1402.1089.pdf>–accessed July 2017).

Test and Test Item Statistics

Facility Index

The range for an item with acceptable facility is taken as being in the range of 0.3 to 0.8. (see Falvey et al., 1994, p. 119ff)

Discrimination Index

An item discrimination (the point biserial correlation) of above 0.3 is considered 'good'. A discrimination of 0.2 to 0.3 is considered 'workable' while a discrimination of below 0.2 is considered unacceptable. (See Falvey et al, 1994, p. 126ff)

Test Reliability

Cronbach's alpha is a test reliability statistic which is generally the starting point for determining a test's worth, with the desirable level (for longer tests, i.e., 80 or more items) usually taken as 0.8 (see Ebel, 1965, p. 337). With shorter tests, lower reliability figures are cited; Ebel (1965, p. 337), for example, states 0.6 for 30 items.

Test Mean

An ideal mean for a 'final achievement' test (Hughes, 2003, p. 13) should be in the region of 0.5. Such a mean suggests – as Gronlund (1985) comments – that the test is generally appropriate to the level of a 'typical' or 'average' student in the class or group. A low mean can suggest that the test is too difficult, with a high mean suggesting that it is too easy (Zimmerman et al., 1990, p. 10). A mean in the region of 0.5 in general indicates that most students managed to finish the test; i.e., that they did their best, and did not simply guess. Further, a mean of 0.5-0.6 indicates that student scores are spread out, and maximises a test's discriminating power (Gronlund, 1985, p. 103).

Standard Error of Measurement

The standard error of measurement (SEM) indicates the extent to which test scores match 'true' scores because all tests will contain a degree of error. As a general rule, an SEM below 10% might be considered desirable. On the controversial Massachusetts Teacher Tests quite a large SEM (17%) was reported – see Haney et al., (1999) for a discussion of the problems associated with the administration of the Massachusetts Teacher Tests – which may be why opponents of the test felt that its reliability was questionable.

Effect Size

While statistical differences are discussed in terms of statistical significance, standard deviation units (SDUs) are also provided in certain instances so that the size of the differences between the two groups may be appreciated. Following Cohen (1988, p. 477-478), an SDU of 0.2 indicates a small effect, 0.5 a medium effect and 0.8 a large effect.

The Rasch Model and Many-Facet Rasch Analysis

In contrast to CTT, the use of the Rasch model enables different facets (e.g., person ability and item difficulty) to be modelled together. First, in the standard Rasch model, the aim is to obtain a unified and interval metric

for measurement. The Rasch model converts ordinal raw data into interval measures which have a constant interval meaning and provide objective and linear measurement from ordered category responses (Linacre, 2006). This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred to as 'logits') evenly spaced along the ruler. Second, once a common metric is established for measuring different phenomena (test takers and test items being the most obvious), person ability estimates are independent from the items used, with item difficulty estimates being independent from the sample recruited because the estimates are calibrated against a common metric rather than against a single test situation (for person ability estimates) or a particular sample of test takers (for item difficulty estimates). Third, Rasch analysis prevails over CTT by calibrating persons and items onto a single unidimensional latent trait scale – also known as the one-parameter IRT (Item Response Theory) model, (Bond & Fox, 2007; Wright, 1992). Latent Trait Analysis (LTA), a form of latent structure analysis (Lazarsfeld and Henry, 1968), is used for the analysis of categorical data. Person measures and item difficulties are placed on an ordered trait continuum by which direct comparisons between person measures and item difficulties can be easily conducted. Consequently, results can be interpreted with a more general meaning. Further, as the Rasch model provides a great deal of information about each item in a scale, its use enables the researcher to better evaluate individual items and how these items function in a scale (Törmäkangas, 2011).

The Rasch model has been widely applied in educational research, especially in the field of large-scale assessment (Schulz & Fraillon, 2011; Wendt et al., 2011). It helps to provide better assessments of performance, enhances the quality of measurement instruments, and provides a clearer understanding of the nature of the latent trait (Bos et al., 2011).

Model Fit

All measurements have expected outcomes: the measurement of a straight line requires, for example, that the object being measured has straight line edges. The one-parameter Rasch model, as a measurement model, expects assessment elements (persons and items) to conform to certain assessment properties in the model. Against this backdrop, the extent to which the assessment properties are adhered to by the assessment elements illustrate the concept of 'model fit' and how this is articulated through what might be termed *broad* and *more focused* criteria.

Broad criteria are the *Point Measure* correlation, and *Infit* and *Outfit* mean square statistics (i.e., estimates of population variance, or standard error). A more focused criterion involves *Standardised Infit* and *Outfit* (i.e., Z-score) statistics. These statistics are outlined briefly below.

Point Measure Correlation

The point measure correlation (PTME) in the Rasch model is comparable to the conventional point biserial correlation. Negative PTME values indicate a lack of model fit.

Infit

Infit may be seen as the 'big picture' in that it scrutinises the internal structure of an item or person. High infit mean square values indicate rather scattered information within the item or person, providing a confused picture about the placement of the item or person. Very small infit values indicate only very small variation and, provide therefore, little information to articulate clear and meaningful judgments about an item or person.

A perfect fit of 1.0 indicates that obtained values match expected values 100%. While acceptable ranges of tolerance for fit vary, acceptable ranges are generally taken as from 0.7 for the lower limit (30% below expectations) to for the upper limit 1.3 (30% above expectations) (see Bond et al., 2020).

Outfit

Outfit gives a picture of 'outliers', that is responses from persons or items that appear to be considerably out of line with where a person or item would expect to be placed. High outfit mean square values would flag an item or person as being out of line with the rest in the pool – hence an 'outlier'.

Standardised Z-Scores

The standardised Z-score for infit and outfit is a more refined model fit criterion, and an extension of the interpretation of mean square values. This is a t-test exploring how well the data fit the model; figures above 2.0 indicate distortion in the measurement system (Linacre, 2006).

Overall data-model fit

Overall data-model fit in Rasch can be assessed by examining the responses that are unexpected given the assumptions of the model. According to Linacre (2006), satisfactory model fit is indicated when about 5% or less of (absolute) standardised residuals are equal or greater than 2, and about 1% or less of (absolute) standardised residuals are equal to or greater than 3.

Frame of Reference (FOR)

To put Rasch measurement further into perspective, it is also important to understand the concept of the frame of reference (FOR) for measurement, and the parameters under which different tests may operate. Humphry (2006) defines a frame of reference as "compris[ing] a class of persons responding to a class of items in a well-defined assessment context." The relevance of this in the current context is that each test has, in Rasch terms, its own "internal logic" (Goodman, 1990). This internal logic refers to the starting point for Rasch measurement models: the basis for Rasch measurement is the total score of the test, computed from a particular set of items, from which the measurement based on the theoretical probability of the particular test is extrapolated (Goodman, 1990). The theoretical probability estimated from a particular test is independent of the test (items, persons and any other relevant facets) but not separated from it. The theoretical measurement estimated is, therefore, an objective measurement albeit specific to the test measured. Rasch calls this "specific objectivity", and occurs, for example, when we measure a rectangle and a circle with the metric. The

two objects may be equal in reference to the metric system (the theoretical and objective measurement) yet different in reference to one being the measurement of four straight lines and the other that of a circumference. Thus, the Rasch measurement of a test has to be interpreted within a particular FOR.

Many-Facet Rasch Analysis (MFRA) and Data Analysis

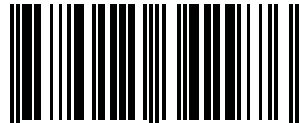
MFRA refers to a class of measurement models that extend the basic Rasch model by incorporating more variables (or facets) than the two that are typically included in a test (i.e., test takers and items). These other variables (or facets) may be markers, scoring criteria, or tasks.

References

- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J.: Erlbaum.
- Bos, W., Goy, M., Howie, S.J., Kupari, P. & Wendt, H. (2011). Rasch measurement in educational contexts Special issue 2: Applications of Rasch measurement in large-scale assessments. *Educational Research and Evaluation*, 17(6), 413-417.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd edition). Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Coniam, D. & Falvey, P. (2018). *High-stakes testing: The impact of the LPATE on English language teachers in Hong Kong*. Singapore: Springer.
- Ebel, R. L. (1965). *Measuring Educational Achievement*. Englewood Cliffs, NJ: Prentice-Hall.
- Falvey, P., Holbrook, J. & Coniam, D. (1994). *Assessing students*. Hong Kong: Longman.
- Friedrich, K. (1999). Interpreting correlation coefficients. <http://acad.cl.uh.edu/itc/educ6032/course/resources/unit2/index.htm>.
- Glaser, D. N. (1999). The controversy of significance testing: Misconceptions and alternatives. *American Journal of Critical Care*, 5(5), 291-296.
- Goodman, L. (1990). Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. *Sociological Methodology*, 20, 249-294.
- Gronlund, N.E. (1985). *Measurement and evaluation in teaching*. New York: Macmillan.
- Haney, W., Fowler, C., Wheelock, A., Bebell, D. & Malec, N. (1999). Less truth than error?: An independent study of the Massachusetts Teacher Tests. *Education Policy Analysis Archives*, 7(4). <http://epaa.asu.edu/epaa/v7n4/>.
- Hatch, E., & Lazaraton, A. (1991). *The research manual*. Boston, MA: Heinle and Heinle.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
- Humphry, S. (2006). *The impact of differential discrimination on vertical equating*. ARC report.
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
- Linacre, J. M. (2006). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.com.
- McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

- Ramsey, P. (1980). Exact type 1 error rates for robustness of student's t-test with unequal variances. *Journal of Educational Statistics*, 5(4), 337-349.
- Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, 102(1), 411-432.
- Schulz, W., & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation*, 17(6), 447-464.
- Törmäkangas, K. (2011). Advantages of the Rasch measurement model in analysing educational tests: an applicator's reflection. *Educational Research and Evaluation*, 17(5), 307-320.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing* 15 (2), 263-287.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndsills, UK: Palgrave Macmillan.
- Wendt, H., Bos, W., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation*, 17, 419-446.
- Whitehead, Paul. 1986. *Statistics 2*. London: Pitman.
- Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, 16(4), 33-45.
- Zimmerman, B. B., Sudweeks, R. R., Shelley, M. F., & Wood, B. (1990). *How to prepare better tests: Guidelines for university faculty*. Brigham Young University Testing Services website: <http://testing. byu. edu/info/handbooks/bettertests. pdf>.

ISBN: 978-9925-34-297-6



9 789925 342976