



CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

**Research and Validation
at LanguageCert**

Volume 2

Edited by Peter Falvey and David Coniam

Language
Cert

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

Research and Validation at LanguageCert Volume 2

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: June 2023

ISBN: 978-9925-34-309-6



CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

**Research and Validation
at LanguageCert**

Volume 2

Edited by Peter Falvey and David Coniam



Foreword

Byron Nicolaides / CEO, PeopleCert Group

When I launched PeopleCert's language testing programme in 2015, my aim was to provide well-researched world class language tests developed in line with PeopleCert's core values of quality, innovation, passion and integrity. Everything we do reflects our core values, and I am now proud to introduce this second volume in our research series 'Certifying Quality in Assessment and Learning' which brings together more of the important work we have done in this context. I encourage and support our team of experts in the development of their professional skills, and you will see many of the chapters in this volume are written by members of our team.

Progress in the development of the LanguageCert suite has evolved quickly ever since our acquisition of the City and Guilds Suite in 2015. We now offer a well-designed, reliable and valid set of exams that meets a wide range of stakeholder requirements. Our focus on continuous improvement, our attention to the real world needs of our users and our ability to innovate at speed means that we are constantly improving what we do and how we operate. When the pandemic significantly disrupted international assessment for example, we were able to move most of our assessment on-line very quickly using our own proprietary on-line proctoring (OLP) system. This allowed those who work with us a largely uninterrupted service. Two chapters in this volume focus on aspects of OLP delivery looking at perceptions and equivalence with conventional delivery.

With the development of the LanguageCert Test of English (LTE) we have focused on providing an adaptive and linear testing system in all 4 language skills. The system is easy to use and measures across all CEFR levels. PeopleCert is a leading global provider of business and IT qualifications with over 50,000 corporates, numerous governments and half a million individuals using our qualifications every year. LTE is designed to service the English language assessment requirements of PeopleCert's business and IT customers and the world of work in general. Two chapters in this volume explore the stability of the item banks that feed LTE as well as other LanguageCert tests. Quality and consistency of measurement are vital aspects of any assessment system and are key features of our approach.

The LanguageCert IESOL suite of level-based exams, ranging from CEFR level pre-A1 to C2, has evolved from the examinations we acquired from City and Guilds and are being used in an education context by schools in many countries to provide meaningful learning attainment targets for learners in a curriculum-oriented context. However, variants of these exams are also being used by the UK Home Office for visas and immigration (UKVI) as well as for international study purposes in a secure English language testing (SELT) context. Research into the stability and reliability of these SELT exams is reported in this volume. It is this research which lays the foundations for one of the key LanguageCert innovations in 2022-3 – the launch of LanguageCert Academic and General exams in September 2022. Following consultation with users of our SELT system we recognized that some changes would be welcomed. As a result, we revised our IESOL B2 and C1 exams making them even more fit for purpose. This evolution is described in detail in chapter 1 - An Exercise in Evolution: Refocusing LanguageCert IESOL C1 to the Academic Context.

Volume 3 of our research is already underway, and I look forward to its publication in due course.

Finally, I would like to thank the contributors to this volume for their hard work and commitment to LanguageCert. I would also like to thank the organisation's staff and the greatly valued network of partners around the world who make the 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 making a contribution to the academic arena of assessment and learning.

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 in 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. LanguageCert has also developed a number of language qualifications – not only for English, but also for other languages. Whereas 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 UK visa applicants.

The current volume is Volume 2 of a series that includes a range of papers produced by LanguageCert's Research Team. Volume 1 (Falvey and Coniam, 2022) contains chapters describing the certification of quality in assessment. One of LanguageCert's missions is to produce tests of the highest quality and its research agendas are designed to support this mission. Readers will note the rigour of the statistical tools that have been developed and used in order to better analyse the data created in the administration of examinations in such areas as calibration, comparison and matching results to different assessment frameworks. These actions have contributed significantly to the provision and accumulation of research evidence about our examinations. A significant contribution to this volume is Jones' Chapter 1 which describes the evolution of two of the LanguageCert SELT examinations into a LanguageCert General and Academic offering. The chapter's content is indicative of the care taken to employ the best research methodology and statistical tools in order to develop and improve the quality of LanguageCert examinations.

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 in 2019 by the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire. Since then, an in-house research programme has been established, headed by Professor David Coniam. This programme builds on and extends – as a sequel to Volume 1 – some of our early research but the LanguageCert

Research Team's brief also 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 a consistently high quality, embodying the best principles of high-stakes assessment. In addition, in order to ensure independent, external validation of its high-stake public examinations, the new, evolved LanguageCert Academic examination has been submitted to ECCTIS (Education Counseling and Credit Transfer Information Service) for validation purposes.

High stakes examinations have to be rigorously developed, trialled, piloted, analysed and constantly improved because of the nature of their use. The language examinations that LanguageCert offers include those that require language proficiency certification for visa and immigration purposes. These examinations are of vital importance to candidates and open to public scrutiny. Therefore, they must meet the highest standards of validity and reliability. That is why the research team is constantly analysing the growing database of evidence that is produced in their administration, and researching ways of ensuring rigorous oversight of the examinations. The visas that are sought by potential students, especially in higher education, can fundamentally affect people's lives and, of course, contribute to the receiving country's overall strategies for aspiring students of higher education. That is why they must not only be but be seen to be exemplary and the research that underpins them of the highest quality as we hope the volume's chapters attest.

References

Falvey, P., & Coniam, D. (eds.) (2022). *Certifying quality in assessment: Research and validation at LanguageCert*. Volume 1. London, UK: LanguageCert.

Contents

Foreword - Byron Nicolaides	5
Preface - Marios Molfetas.....	7
Contributors	11
Common Abbreviations	13
Introduction - Peter Falvey.....	15
SECTION 1: ASSESSMENT AND CURRICULUM	23
Chapter 1: An Exercise in Evolution: Refocusing LanguageCert IESOL C1 to the Academic Context - Cathy Jones.....	25
SECTION 2: CALIBRATION/VALIDATION STUDIES	55
Chapter 2: Externally-Referenced Anchoring of LanguageCert SELT Tests - Michael Milanovic, Tony Lee, David Coniam and Yiannis Papargyris	57
Chapter 3: Aligning LanguageCert SELT Tests to the LanguageCert Item Difficulty Scale - Tony Lee, Yiannis Papargyris, Michael Milanovic, Nigel Pike and David Coniam	71
Chapter 4: LanguageCert SELT Writing Test Quality - David Coniam, Irene Stoukou, Tony Lee and Michael Milanovic.....	81
Chapter 5: Exploring Item Bank Stability Through Live and Simulated Datasets - Tony Lee, David Coniam and Michael Milanovic.....	95
Chapter 6: Exploring Item Bank Stability in the Creation of Multiple Test Forms - David Coniam, Tony Lee and Michael Milanovic.....	107
SECTION 3: ORIGINAL RESEARCH	115
Chapter 7: The Role of Expert Judgement in Language Test Validation - David Coniam, Tony Lee, Michael Milanovic, Nigel Pike and Wen Zhao	117
Chapter 8: Using Self-Assessments to Investigate Comparability of the CEFR and CSE: An Exploratory Study Using the LanguageCert Test of English - Wen Zhao and David Coniam.....	135
Chapter 9: Online Invigilation of English Language Examinations: A Survey of Past China Candidates' Attituded and Perceptions - David Coniam	155
Chapter 10: The Delivery of Speaking Tests in Traditional or Online Proctored Mode: A Comparability Study - Michael Milanovic, Tony Lee and David Coniam.....	169
Chapter 11: Interactional Competence and the Role Roleplay Plays: The LanguageCert Perspective - Leda Lampropoulou	183

Chapter 12: Validating Communicative Tests of Reading and Language Use of Classical Greek - David Coniam, Polyxeni Poupounaki-Lappa and Tzortzina Peristeri	203
Glossary of Statistical Techniques Used in the Volume - Peter Falvey.....	215

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 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.

Angeliki Cheilari 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 also holds an M.Ed. 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.

Cathy Jones is an Academic Associate at LanguageCert, working on qualification development. She has worked in qualification development for almost twenty years with particular expertise in developing multi-level language curricula, tests and teaching materials. Cathy holds a BA in French and History of Art from University College London.

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.

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.

Michael Milanovic, previously CEO of Cambridge Assessment English, has been working extensively with PeopleCert since 2015. He 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 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.

Irene Stoukou is Chief Examiner Team Leader at LanguageCert. She is responsible for the implementation of fit-for-purpose marking and assessment processes, as well as the training and continuous monitoring of examiners. She holds a BA in English Language and Literature from Aristotle University of Thessaloniki, an MA in Modern and Contemporary Literature, Culture, and Thought from Sussex University, and is currently a PhD candidate at Aristotle University of Thessaloniki.

Wen Zhao is Dean of the School of Foreign Studies at Jinan University, Guangzhou. Her main publication and research interests are in corpus linguistics, English curriculum and instruction, and EFL writing. She has been working and researching in English language teaching and learning, and has been involved in national English curriculum development for senior secondary vocational education and College English education.

Common Abbreviations used in the Book

ALTE	Association of Language Testers in Europe
ANOVA	Analysis of Variance
APA	American Psychological Association
AT	Adaptive Test
CEFR	Common European Framework of Reference
CET	College English Test
CoE	Council of Europe
CRELLA	Centre for Research in English Language Learning and Assessment at the University of Bedfordshire
CSE	Chinese Standards of English
CTS	Classical Test Statistics
CTT	Classical Test Theory
DIF	Differential Item Functioning
EFL	English as a Foreign Language
ELT	English Language Teaching
ESOL	English to Speakers of Other Languages
FOR	Frame of Reference
GMAT	English Language Teaching
IC	Interactional Competence
IELTS	International English Language Testing System
IESOL	International English for Speakers of Other Languages
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
NARIC	National Recognition Information Centre
Ofqual	Office of Qualifications and Examinations Regulation
OLP	Online Proctoring
OPI	Oral Proficiency Interviews
PB	Paper-Based
PBC	Point Biserial Correlation
PBM	Paper-Based Marking
PPM	Pearson Product-Moment Correlation
PTME	Point Measure Correlation
QCA	Qualifications and Curriculum Authority
RQ	Research Question
SA	Self Assessment
SD	Standard Deviation
SEM	Standard Error of Measurement
TCC	Test Characteristic Curve
TLU	Target Language Use
TM	Traditional Mode
TOEFL	Test of English as a Foreign Language
UKVI	UK Visas and Immigration
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 follows the publication of Volume 1 (2022) with a further compilation of research and assessment curriculum studies aimed at supporting LanguageCert's research-led, quality-oriented approach to its language assessments. The volume is divided into three sections comprising twelve separate chapters that address a variety of assessment topics.

Readers should note that, as all the chapters have been written to be read on their own, there is a certain amount of repetition especially when the structure and makeup of the various examinations are described and discussed. In addition, each chapter contains its own reference section.

Section 1: Assessment and curriculum

Section 1 consists of one chapter.

Chapter 1, *An Exercise in Evolution: Refocusing LanguageCert IESOL C1 to the Academic Context*, describes how an examination evolution occurs. It provides the rationale for the evolution, its purpose and the needs it meets, the curricular factors in play, the development of the examination, and its pretesting, piloting and eventual offering to the public. The evolution that is described in this chapter arises from the redevelopment of an existing examination, the LanguageCert IESOL C1 from a more general proficiency test of English skills to an academic context. As reported in the Introduction to this volume, many of the LanguageCert examinations are taken for the purpose of fulfilling English Language proficiency requirements for the granting of visas for immigration, study and work purposes.

The development of the IESOL Academic and General tests, described here, focuses on academic language requirements, developed by LanguageCert personnel and pre-tested and piloted internationally, at LanguageCert approved test centres under secure test-taking conditions, with pretesting populations which are representative of each test's intended candidature.

The chapter describes the evolution and refocusing of the LanguageCert IESOL C1 to the academic context. The chapter's content is indicative of the care taken to employ the best research findings, methodology, and statistical tools in order to develop and improve the quality of all LanguageCert examinations. Readers will

note that detailed descriptions and demonstrations of the use of these tools are provided in many of the other chapters in the volume.

Section 2: Calibration/Validation studies

Section 2 consists of five chapters, each of which helps to describe in detail the methods used to calibrate and validate LanguageCert tests. Many of these methods and tools have been used to better develop the LanguageCert IESOL C1 for Academic Purposes. The two elements of validation and calibration are important, especially in high stakes examinations. Validity is the concept that a test really tests what it is supposed to test – and not something else. Many of the studies in this section focus on this element. Calibration is the process by which tests are aligned reliably to a measurement scale or external frame of reference such as the CEFR (the Common European Framework of Reference of Languages) or the CSE (the Chinese Standards of English).

The CEFR has become one of the most commonly referenced, if not the main, framework that examinations, examinations systems, curricula and learning materials are aligned to. This allows efficient and authentic judgements to be made by immigration authorities, educational institutions, admission tutors and vocational employers about language proficiency qualifications proffered to meet language proficiency requirements in different countries. The CSE (Chinese Standards of English) is a Chinese framework against which other frameworks are often matched in order to examine how closely candidates who have been examined under the CSE, measure up against other assessment frameworks in an international context such as the CEFR.

It is, thus, important for ongoing studies to analyse and improve these two fundamental elements in assessment are reported on frequently both to authenticate the examinations being studied, the analyses carried out on them and to ensure that recipients of the results can be confident of the quality of those examinations. In the same way, the purpose of calibration studies is to provide reassurance that test forms measure in a consistent and reliable manner and where appropriate, assessment frameworks can be matched meaningfully and satisfactorily against each other.

Chapter 2, *Externally-Referenced Anchoring of LanguageCert SELT Tests*, describes a statistical methodology that enables assessment professionals to address the long-standing problem of how to compare test forms that have no items in common. In high stakes examinations, security is of paramount importance, so the use of common items is at times eschewed when tests are administered at different times in different geographical areas in order to thwart systematic attempts to cheat. The chapter provides a detailed description of the process in operation.

Chapter 3, *Aligning LanguageCert SELT Tests to the LanguageCert Item Difficulty Scale*, presents a study designed to investigate how LanguageCert SELT tests are aligned to the LanguageCert Item Difficulty (LID) Scale. The chapter builds on the study, reported in Chapter 2, which established, through the use of externally-referenced anchoring, that the LanguageCert SELT B1–C1 tests are robust. The LID scale represents an important backdrop to the development of the LanguageCert IESOL C1 for Academic Purposes described in Chapter 1.

Any serious examination body needs to ensure that, where it uses level specific tests (these being tests aimed at specific CEFR level such as B2 or C1 for example), that such tests accurately reflect the underlying measurement scale and measure at the desired level. In this chapter, the alignment of LanguageCert SELT tests to specific CEFR levels (in relation to the two objectively marked components of Listening and Reading) is investigated.

As the chapter illustrates, the LanguageCert SELT tests, in general, assess at their designated CEFR level but, by design, also contain items which allow them to assess across levels. At the C1 level, there are items which assess above C1 and, at the other end, below C1. Likewise, at the B2 level, there are items which assess both above and below B2. The findings provide evidence that when LanguageCert SELT B2 and C1 were redeveloped as LanguageCert General and Academic respectively, their capacity to measure across multiple CEFR levels was confirmed.

Chapter 4, LanguageCert SELT Writing Test Quality, investigates the quality of a test – the LanguageCert SELT Writing Test – not covered in Chapter 3. The chapter focuses on a vital aspect of writing tests – test quality. The large sample where over 11,000 candidates were analysed, comes from the LanguageCert SELT Writing Tests administered at CEFR levels B1 and B2 during the period 2021-2022. 60 examiners and 18 different tasks were involved. Using Many-Facet Rasch Analysis (MFRA), the study explores the consistency of marking in terms of examiner, task, and rating scale fit and severity. The use of MFRA provides analysts with an understanding of test quality, superior to Classical Test Statistics alone. Indeed, the LanguageCert General and Academic tests (LCG and LCA) are very much reliant on the research that is described in this section of the Volume. It is clear that linkage and calibration to the CEFR is vital for the LCG and LCA so that their results can be seen to be aligned with the scales of the CEFR.

Results from the study indicate that, for the different test facets, fit to the Rasch model was generally good. The task and rating scale severity ranges were generally within acceptable limits. Crucially, examiner fit was good, with only a small number of examiners exhibiting misfit. Against the backdrop of the analysis reported, the study concludes that the SELT Writing Tests pitched at CEFR levels B1 and B2 are robust and fit for purpose.

Chapter 5, Exploring Item Bank Stability Through Live and Simulated Datasets, explores the robustness of large item banks. LanguageCert manages the construction of its tests, exams and assessments using a sophisticated item banking system which contains large amounts of test material with content characteristics such as macroskills, grammatical and lexical features; and measurement characteristics such as Rasch difficulty estimates and fit statistics. In order to produce content and difficulty equivalent test forms, it is vital that the items in any LanguageCert bank manifest stable measurement characteristics. These large, stable and robust item banks are essential for the evolution of the LanguageCert Academic and General examinations because they provide a reliable basis for measuring the effectiveness of the items that make up the tests and the overall scores that determine the levels of proficiency that the candidates have achieved.

The chapter describes the first of two linked studies exploring the stability of one of the item banks developed by LanguageCert. This particular bank has been used as an adaptive test bank and comprises 827 calibrated items. It has been administered to over 13,000 test takers, each of whom have taken approximately 60 items. The purpose of these two exploratory studies is to examine the stability of this adaptive test item bank from

both statistical and operational perspectives in order to facilitate the reliable placing of students both efficiently and accurately, rather than exposing them to the multiple items that they would normally face in other testing situations.

Results pointed to item bank stability, indicating that items comprising the adaptive item bank are of high quality both in terms of content and statistical stability – an important consideration when item banks are used for multiple purposes. Potential future stability was confirmed by results obtained from a Bayesian ANOVA. The aim of the study was to lay the groundwork for a subsequent follow-up study where the utility of this adaptive test item bank is verified by the construction, administration and analysis of a number of linear tests.

Chapter 6, Exploring Item Bank Stability in the Creation of Multiple Test Forms, describes the complex item banking system that facilitates the construction of LanguageCert tests. The system's item banks contain large amounts of test material covering a wide range of content and construct characteristics. They are calibrated on the basis of Rasch difficulty estimates, fit statistics, and classical test statistics analysis and described in terms of content such as topic, grammatical and lexical focus.

Effectively constructed and managed, item banks allow for the creation of test forms which are consistent and comparable both in terms of content and difficulty. This is relevant not only when creating tests intended to measure at a particular level such as CEFR level B1 but also when developing tests which measure across multiple levels from A1 to C2.

The ability of LanguageCert to create stable and reliable item banks ensures that when a test is developed, one such as the LanguageCert Academic examination, the items that are selected to make up the various tests are consistent and, in the case of multiple test forms, reliable and accurate. This feature is always important, but in the case of high-stakes examinations such as the LanguageCert Academic test (LCA) and the LanguageCert General test (LCG), it is a vital component of such a high-quality examination.

The chapter concludes by claiming the items that comprise the item bank have been well set, and provide strong support for the robustness of the item bank as a clearing house from which many different tests may be constructed.

SUMMARY OF SECTION 2

The five chapters in Section 2 are linked together by one purpose: to illustrate the various ways in which LanguageCert studies combine to demonstrate the effectiveness of methods used to ensure the robustness of the tests themselves and how those tests can be compared, calibrated, and matched both to each other and to levels in different assessment frameworks such as the CEFR and the CSE (see Weir, 2005). These methods have enabled the developers of the LanguageCert Academic and General tests to move smoothly and efficiently in their construction, knowing that they can rely on the stability of the item banks that are essential components of their creation to produce high quality examinations.

Section 3: Original research

Section 3 consists of six diverse chapters. Each chapter refers to different aspects of LanguageCert's examinations that have been investigated.

Chapter 7, *The Role of Expert Judgement in Language Test Validation*, focuses on one specific aspect of language test validation, namely the role of expert judgement. Given that the calibration of test materials generally involves the interaction between empirical analysis and expert judgement, the chapter explores the extent to which scale familiarity might affect expert judgement as a component of test validation in the calibration process.

In order to ascertain whether expert judges were equally comfortable placing test items on two separate scales (the CSE or CEFR), experts from a prestigious university in China who set the (College English) CET-based test, were asked to expert judge the CET items against the nine CSE levels with which they were very familiar. They were then asked to judge the LanguageCert LTE items against the six CEFR levels, with which they were less familiar. Both sets of expert ratings and the test taker responses on both tests were then calibrated within a single frame of reference and located on a single scale – the scale developed and used by LanguageCert.

In the analysis of the expert ratings, the CSE-familiar raters exhibited higher levels of agreement with the empirically-derived score levels for the CET items than they did with the equivalent LTE items. This supports the proposition that expert judgement may be used in the calibration process where the experts in question have a strong knowledge of both the test material and the standards against which the test material is to be judged. However, when the judges were asked to place LTE items on the CEFR scale, results suggested that expert judgement may be less reliable in the evaluation of unfamiliar items against less familiar standards. The study supports that proposition that expert judgement is a useful dimension of item calibration but highlights the importance of familiarity in improving the accuracy of such judgements.

Chapter 8, *Using Self-Assessments to Investigate Comparability of the CEFR and CSE: An Exploratory Study Using the LanguageCert Test of English*, reports on the use of self-assessments for triangulation purposes in comparability studies between the Common European Framework of Reference for Languages (CEFR) and the China Standards of English (CSE). This study helps to set the groundwork for determining the correspondence between LanguageCert Tests which are aligned to the CEFR and the CSE.

The study helps to establish such equivalences by using self-assessments in two tests: the LanguageCert Test of English of reading and language use for the CEFR; and a comparable test of reading and language use produced by a top-tier China university.

Chapter 9, *Online Invigilation of English Language Examinations: A Survey of Past China Candidates' Attitudes and Perceptions*, moves to the topic of online invigilation. Drawing on a previous large-scale study examining the reactions of past candidates to the use of online invigilation – or online proctoring (OLP) – in the delivery of high-stakes English language examinations (Coniam et al., 2021), this chapter investigates the topic further by reporting on the responses of the subset of China candidates in the larger-scale sample.

The chapter explores the challenges and benefits that both modes offer in terms of accessibility, fairness, security and cheating. A strong endorsement by the China cohort 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. The results are indicative of a broad acceptance of OLP, pointing to strong future uptake of the OLP mode of test delivery. These are reassuring findings when, post-Covid, OLP needs to be secure, reliable and student-friendly when the candidates are facing high stake examinations such as the LanguageCert Academic examination.

Chapter 10, *The Delivery of Speaking Tests in Traditional or Online Proctored Mode: A Comparability Study*, moves from a general examination of the benefits and challenges of online proctoring to the more specific and challenging task of assessing speaking skills. The chapter compares high-stakes English language Speaking Tests administered face-to-face in either a traditional centre-based mode (TM) or in an online proctored mode (OLP) in order to determine whether different modes (TM/OLP) produce different scores. This is an important investigation when LanguageCert are faced with a growing number of candidates who, because of such factors as the pandemic and their remote location, need to access a mode of assessment which is not face to face but which has to be both reliable and secure.

The data examines the results of a large sample of test takers taking English language Speaking Tests at four CEFR levels – B1 to C2 – via TM or OLP. The data were analysed using descriptive statistics, effect size differences and equivalence tests. While a small degree of difference in scores obtained between modes was apparent at C2 level, the differences were not found to be statistically significant.

The study finds that whether Speaking Tests are delivered in online proctored mode or in traditional face-to-face mode, test takers receive comparable scores. The study confirms the contention that mode of test delivery does not significantly affect test taker scores. These findings help stakeholders to be confident that LanguageCert examinations that use OLP can be relied on. This is particularly the case in the development of a high stakes examinations such as the LanguageCert Academic test. However, it should be noted that work on test security remains a challenge to testing agencies and OLP offers its own specific challenges that are being addressed constantly.

Chapter 11, *Interactional Competence and the Role Roleplay Plays: The LanguageCert Perspective*, continues the theme of assessing spoken English skills but examines the competence known as interactional competence and the use of roleplay as an assessment tool. The co-construction of meaning and the shared nature of the interaction are seen to be operationalised in an optimal manner using the roleplay task. The effect of the task is explored through the perspective of the LanguageCert IESOL Speaking exams, which are used as examples to demonstrate the issues of scalability, discriminability, score separability, and the so-called interlocutor effect.

Chapter 12, *Validating Communicative Tests of Reading and Language Use of Classical Greek*, moves away from tests of English proficiency to the realm of testing Classical Greek. It builds on earlier work by Poupounaki-Lappa et al. (2021), which described the development of a communicative test of Reading and Language Use of Classical Greek, calibrated to the CEFR at levels A1 and A2.

In this chapter, analysis moves further by outlining how the two tests of Classical Greek were calibrated both together and to the CEFR. The chapter is also designed to be of interest to educators of other classical languages not only by facilitating robust test design, but also by demonstrating the methods by which tests can be linked together on a common scale (as with the CEFR) or by linking tests one to another (e.g., different end-of-year tests, at different points in time).

SUMMARY OF SECTION 3

While Section 3 is more diverse than Section 2, all the chapters nonetheless contribute to research about LanguageCert's examinations and to the ways in which research is conducted to constantly provide data not only for the institution itself but also for all the stakeholders and other assessment professionals worldwide who wish, in varying ways, to draw on the research. Many of the chapters in Section 3 also exemplify the research that fed into the development of the LanguageCert Academic and General examinations.





SECTION 1: ASSESSMENT AND CURRICULUM



Chapter 1: An Exercise in Evolution: Refocusing LanguageCert IESOL C1 to the Academic Context

Cathy Jones

Abstract

No quality test can be static. To ensure ongoing fitness for purpose, test developers need to respond dynamically to changing stakeholder expectations and requirements. This chapter discusses the methodology for refocusing LanguageCert IESOL C1 to operate more effectively in the measurement of English language skills needed for academic study at undergraduate, post-graduate or professional level. It describes how LanguageCert Academic – a four-skill, multi-level test, aligned to a common underlying measurement scale – derives from a bank of pretested and calibrated assessment material and associated validation research based on an established candidature. This chapter highlights underpinning research, evidence and best practice which have informed the development and definition of a high-stakes relevant, reliable and secure test. It covers test purpose and construct, proficiency levels, task selection, test content, assessment criteria, test delivery and results and an integrated learning ecosystem.

Keywords: test design, test purpose, test content, washback, integrated learning ecosystem.

Introduction

LanguageCert, a part of the *PeopleCert* group, is a leading international assessment company. It administers a This chapter is based on Chapter 1 of Falvey and Coniam (eds.), Volume 2 of the LanguageCert series *Certifying Quality in Assessment and Learning*.

It is often said that ‘qualifications open doors’ in the sense that candidates take high-stakes exams to access life-changing opportunities such as university admission or migration for work. In the same way that a door

provides access to a new room or space, a qualification can provide access to higher education, career opportunities, experiences and communities. Assessment of all kinds can have a transformative impact on the life chances of individuals and as such there is an ethical and moral responsibility to ensure that, as powerful gateways to new learning experiences, personal growth and professional development, exams are reliable, secure and fit for purpose.

The metaphorical door's function must be checked regularly to make sure it is well-oiled and remains a good fit, with minimal shrinkage or expansion over time, and that it is well kept, up to date and in keeping with its surroundings.

Imagine achieving the task of opening the door but without any proof that it was you who had successfully managed to prise it open, or that the parameters had changed and the door you had opened was in fact no longer in use and you had missed a small notice reading 'Please use other door'. In other words, if we apply this analogy to a qualification, the qualification needs to be valid and reliable; it needs to test what it purports to test reliably and consistently over time.

A faulty door that doesn't fit properly or function as it should, just like a qualification which is incomplete, outdated or irrelevant, will lead to frustration and disappointment and thwarted potential. The intention of this chapter is to convey the breadth and depth of considerations in developing a test which is reliable, secure and fit for purpose as well as sufficiently innovative, user friendly and recognised in a competitive market.

English language proficiency is increasingly becoming a requirement for many academic and professional endeavours. Accurate and reliable English language assessment is vital in determining an individual's level of proficiency in English. Assessment of English language proficiency ensures that individuals can communicate in English effectively whether it be for academic, employment or social purposes. To do this, assessments need to measure an individual's communicative competence - their ability to understand and use language accurately, appropriately and fluently.

This chapter describes how LanguageCert IESOL C1, a proficiency test of more general English skills has been refocused for a more academic context. The chapter covers different aspects of test design and development. It includes the rationale and underpinning research for the evolution. It discusses the test construct and how this definition extends beyond the test to inform the design of learning materials and makes a positive impact by design. It covers the test measurement scale, scoring and reporting. The chapter also describes the ongoing programme of internal and external research and validation as the LanguageCert Academic test is pre-tested, piloted and launched to the public.

Background

As a leading provider of language exams and qualifications recognised by universities, employers and governments around the world, LanguageCert exams are designed to assess language skills in a real-world context, using tasks and materials that are relevant to candidates' specific needs and goals. LanguageCert ensures that the CEFR is embedded into the test development cycle and the quality and level of test materials reflect this – providing an international standard for assessing language proficiency.

The LanguageCert English language portfolio includes a range of established, recognised, successful, high-stakes qualifications, including: LanguageCert International English for Speakers of other Languages (IESOL), a level-specific exams, ranging from A1 to C2 for both occupational and personal use. The portfolio also includes the LanguageCert Test of English, a multi-level adaptive test of English in the workplace, as well as a suite of as well as a suite of secure level-specific IESOL SELT (Secure English Language Test) qualifications, using ESOL exam structures, tasks, and items. The IESOL SELT qualifications meet the specific requirements of the UK Home Office as proof of English language competence for visas and immigration for life, work or study visa types.

In 2020, development of Language Cert Academic (LCA) was conceived as a dynamic response to changing markets and stakeholder expectations. As a result, work began to extend the portfolio with a high-stakes test for the academic sector, LanguageCert Academic, together with a counterpart qualification, LanguageCert General (LCG) for those wanting to migrate for work or study in an English-speaking context. Both tests derive from the same item bank and report scores across relevant levels on the same measurement scale for the four skills, Listening, Reading, Writing and Speaking. The focus of LanguageCert Academic is fine tuned for an explicit academic purpose in terms of contexts, tasks and levels and is the main focus of this chapter. LanguageCert General (LCG) will be the focus of a subsequent paper later in 2023. One of the main outcomes of the evolution of the existing IESOL B2 and C1 tests into the LanguageCert Academic and LanguageCert General exams is that it enables measurement and certification across a broader range of language attainment levels. This meets growing demand from test takers and recognising institutions for more breadth in how single level examinations assess.

A phased roll out of LCA and LCG began in 2022 to ensure that all issues related to the effective delivery of the exams could be addressed. A gradual roll-out (Phase 1) was planned deliberately to ensure not only a smooth introduction of the revised exams but also to avoid confusion with existing IESOL SELT exams used for UK visas and immigration (UKVI). LCA and LCG have been designed to replace four single level tests, already in use by UKVI in 2023. Phase 2 of the rollout took place from June 2023 when LanguageCert General and Academic were made more widely available in a large number of test centres managed by Prometric and PeopleCert.

Purpose

This chapter describes the development of LanguageCert Academic as an exercise in responsive test development and test evolution as part of a continuous review cycle. It also exemplifies for test users how ongoing research informs best practice and how it can be applied to test development where a different if related context or purpose is required.

An Evidence-informed Approach

The LanguageCert Academic test development was built on a portfolio of research and validation covering three main areas:

1. Wider underpinning research into assessment, learning and teaching
2. Research and validation on the wider portfolio of LanguageCert qualifications carried out both by the LanguageCert research team and external research (e.g., conducted by CRELLA, UK NARIC (now UK ENIC), etc.

3. Research undertaken by the LanguageCert research team with specific reference to LCA

Figure 1 below shows how these different bodies of research draw on and feed back into each other in an ongoing reciprocal cycle. Qualification development draws on research undertaken by LanguageCert, as well as the underpinning body of wider assessment research. The qualification-specific research generated for LCA feeds back in turn to the wider assessment landscape, and informs future LanguageCert products as well as wider development of how assessment of this kind can be used to develop products to support international progression and mobility.

Figure 1: Use of assessment research in test development at LanguageCert



Underpinning Evidence

The LCA test assesses the English language abilities needed for students to participate in higher education, and to participate in campus life in English-speaking contexts. There is extensive evidence for the nature of the language tasks that international students need to engage in when studying at tertiary level in English-speaking countries. Much of this is summarised in Xi and Norris (2021). The design of the LanguageCert Academic test was informed by such evidence, and consideration was given to research into representative tasks and features of language use from a wide range of sources (Appendix 1).

The TOEFL 2000 Listening Framework, developed by the Educational Testing Service (Bejar et al., 2000) flags the importance of a number of cognitive processes involved in listening. The framework identifies a range of listening materials and discourse encountered in academic contexts and the skills and strategies required for success. The value of designing tasks that test higher-order cognitive skills such as analysis and evaluation is also discussed by Field (2012), who found that lecture-based questions are cognitively valid because they test real-world academic skills and processes.

Similarly, the TOEFL 2000 Reading Framework (Enright et al., 2000) details the skills required for reading a range of materials in different genres in academic contexts. As for the Listening Framework, this framework also calls attention to the importance of defining the underlying cognitive processes in reading. The academic reading construct was examined in a key study by Weir et al (2009) which looked at the relationship between IELTS and the reading experiences of students during their first year at university. The study found a positive link between IELTS test scores and students' subsequent experience of the academic reading demands at university, including understanding vocabulary, textual features, organisation and discourse coherence.

Writing skills and strategies required for success in an academic context and the cognitive processes involved in academic writing are highlighted in the TOEFL 2000 Writing Framework (Cumming et al., 2000). Nesi and Gardner (2018), used the British Academic Written English (BAWE) corpus, which includes just under 3000 good-standard university-level student writing responses across four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) and four levels of study (undergraduate and taught Masters level), to explore characteristics of student writing in tertiary education. As part of its findings, the study established that certain genres, namely essays and reports, were common across all disciplines.

The skills and strategies for academic speaking success are provided in the TOEFL 2000 Speaking Framework (Butler et al., 2000). Brown and Ducasse (2019) investigated differences between performance in TOEFL iBT speaking tasks with performances on academic oral assessment tasks in first-year students across three faculties. The study found that the TOEFL iBT tasks were largely represented in the academic tasks, but with some difference across the two contexts in terms of complexity and cognitive demand.

LanguageCert has made a significant contribution to the research landscape, and LanguageCert research and validation research has provided detailed evidence of direct relevance to the evolution of LCA and LCG. Extensive and frequent calibration and validation of the LanguageCert suite, is presented by Milanovic et al. (2023a). They describe how the anchoring of IESOL SELT tests can be externally referenced to provide an evidence-informed statistical methodology. This methodology can then be used to ensure comparability and robust equivalence of test forms on an underpinning scale. Work to align tests to the LanguageCert Item Difficulty (LID) scale reported by Lee et al (Chapter 3, this volume) is a prerequisite when extending tests from single to multi-level and adding to the existing suite. This study established how LanguageCert IESOL pass/fail level-specific SELT tests not only assess at their designated level but also include items which assess above and below the designated level of the test. This corroboration of a nascent multi-level linear test has informed the extension of the testing scale to allow LCG and LCA test takers to be placed across the four target CEFR levels of proficiency most pertinent to each domain.

LanguageCert research has also been instrumental in evaluating the stability and robustness of large-scale item banking for high-stakes qualifications. LanguageCert uses a proprietary secure item bank (IB) to manage all tests with strict access protocols and workflows for process compliance. Reports can be run to interrogate the volume of materials at different stages of production in the item bank, by test and task type. Reports can also be run for more detailed information, such as the amounts of materials at certain difficulty levels by test part. Items and tasks are commissioned into the IB with the intention of re-using the material over time and across test versions. A variety of re-use parameters are in place for different types of products and different skills. LanguageCert have explored item bank stability in several research pieces, Lee et al (Chapter 5, this volume) through live and simulated datasets and Coniam et al (Chapter 6, this volume), in reference to the

creation of multiple test forms. Both these pieces have directly supported the development of LanguageCert Academic and LanguageCert General, providing necessary evidence of the integrity and stability of the item banks.

What is Academic English and Why is it Important?

Tests of general English are designed to assess an individual's overall language proficiency in a variety of domains, including professional, social, occupational, personal as well as educational. Tests of general English can be useful in determining a candidate's overall language proficiency but they may not be sufficient for academic purposes. Knoch (2015) defines the concept of academic literacies as a "set of social practices and conventions that surround academic writing and discourse". Knoch argues that academic English involves a set of academic specific skills and competencies, including analysis, evaluation, synthesis and academic writing. By definition, general English tests do not intentionally focus on the language, skills, expectations, conventions and styles that students will encounter in academic contexts. Turner's (2002, 2012) Assessment of English for Academic Purposes (AEAP) framework, details productive and receptive language skills required for academic success. For example, in academic reading and writing, students are expected to read and write more complex and lengthy texts than candidates of general language proficiency, because a higher level of comprehension and analysis is required. Critical thinking is essential for understanding complex ideas, evaluating information, identifying bias and for developing original and evidence-based arguments. Test takers have an opportunity to develop and demonstrate these skills in the LCA test by completing tasks such as presenting an argumentative essay on a topical subject and participating in a discussion based on evaluation of an academic source.

For LCA, general academic English refers to the type of language that students need for university and college programmes. This includes generic academic vocabulary and expression relevant to most domains (i.e., not subject or discipline specific), and competences used across common academic tasks (e.g. writing essays, giving presentations).

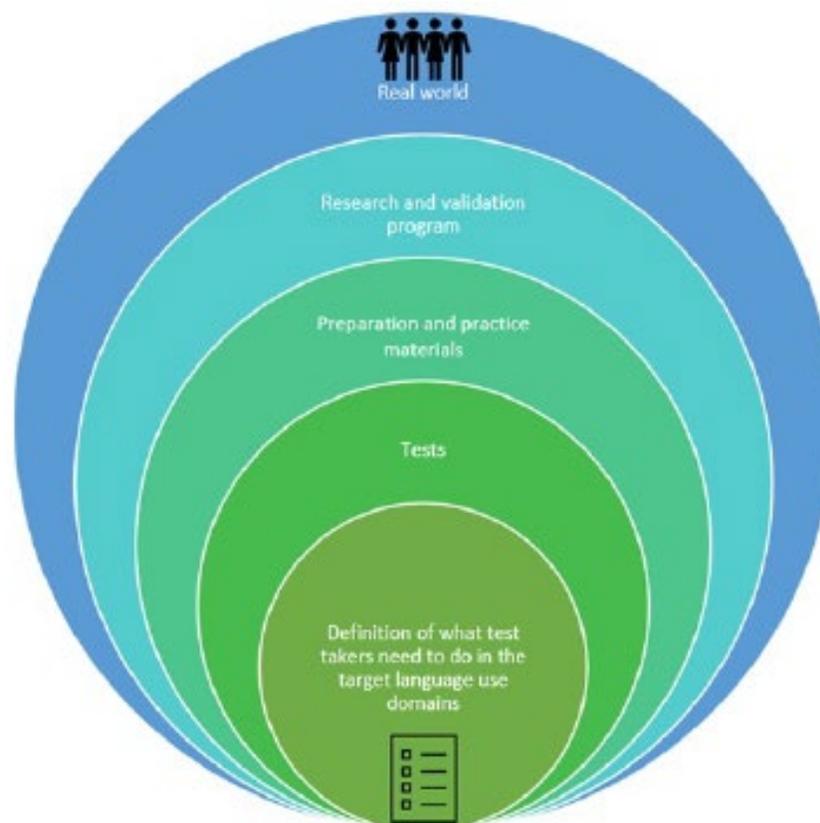
In evolving the IESOL SELT test and populating the item banks with materials appropriate for an academic context, it was essential to understand the skills, competences and cognitive processes that are specific to general academic English. Refocusing the IESOL SELT test involved more than simply recasting a bank of items and changing the scenario of given tasks. For example, a listening item such as "You will hear two friends talking at an art gallery" could simply be reframed as "You will hear two art students talking at an art gallery". However, reviewing existing materials had to take into account a range of detailed considerations, including the subject of the conversation, and the specific skills, vocabulary and cognitive processes that are the focus of the item. Potentially, if the dialogue between the gallery visitors was about the art on display it could be suitable for recasting in the way outlined above, but if the dialogue was more about lost property, the location of the museum shop or the entry costs, it would be less appropriate to be included in a test of academic English and the reframing as two students in an art gallery would serve a face-validity purpose only. There is a place for some 'academic-related' content, but it can only constitute a small fraction of the total content. This necessitated a large-scale commission of items and task content appropriate for testing the academic target language use domain.

Defining the Target Language Use Domain

The focus on domains, and the target language use within them, permeates all aspects of test design, development, and delivery. This includes how LanguageCert ensure candidates are supported with domain-specific practice tests and learning materials. LanguageCert do this 'by design', with all aspects of each qualification being fully integrated and aligned.

The conceptual model in Figure 2 below illustrates the connections that shape LanguageCert's approach to language assessment, and the position of learning and preparation materials within these connections.

Figure 2: Approach to assessment, learning and preparation in the real world



At the core of the concept is the definition of what test takers need to do in the target language use (TLU) domain of the test.

This definition of what test takers need to do in the real world is critical; it is based on knowledge and experience, informed by close engagement with key stakeholders (including the LanguageCert Advisory Council, the LanguageCert Academic Panel, and the CRELLA Concordancing Studies Review Panel) and is validated through ongoing research. The definition is monitored, and refined in line with shifts in real-world requirements as well as new research, and validation findings. This foundational definition shapes the design of LanguageCert's

tests. Test specifications and assessment criteria ensure appropriate depth and breadth of coverage of the test construct.

Preparation and practice materials support the tests. Such materials connect what is learnt and practiced prior to the test, with the skills defined and tested in the exam. Detailed explanations of test structure and requirements combine with practice questions, mock tests, and related activities to help learners understand and build the defined skills and their confidence.

An associated research and validation programme, including stakeholder review and consultation, and impact analysis, has the foundational definition of LanguageCert's test construct at its heart and encompasses both the tests and their related learning materials. Formal structures such as that of the Academic Panel create a clear sounding board for LanguageCert's research and validation studies.

The real world wraps around every aspect of the conceptual model, including the definition of the domain's TLU, the test construct, the tests, their learning materials, and the research and validation programme. That the real world permeates all aspects of the assessment work is vital; it ensures the accuracy and relevance of the TLU for specific domains. The intention is that practising and developing these skills and competences will enable learners to succeed as candidates in the test and then, beyond that, to succeed as individuals in the real-world domains the tests are designed to represent.

Washback by Design

Washback by design refers to the intentional and systematic incorporation of the potentially positive impact of an assessment on teaching and learning into the test development process. Green (2007) has examined the effects of high-stakes qualifications such as IELTS on teaching and learning, exploring the effect of assessment and evaluation criteria on development of test-taking strategies and development of critical thinking and analytical skills alongside communicative language competence. Cheng and Sultana (2022) provide a comprehensive review of washback research in language testing and the potential for assessment to promote positive washback in teaching and learning. They highlight a need for continuing research and assessment policies that promote positive washback and support teaching and learning.

Designing assessments that promote positive washback and measuring their intended impact is complex and challenging and yet, emphatically, non-negotiable. To deliver an assessment without attempting to understand or measure its intended (and unintended) consequences and its impact on the lives and life chances of test takers would be morally and ethically questionable.

The area of washback by design is one in which LanguageCert is poised to make a contribution, adding to the corpus of work already undertaken by Cheng, Green and others in the field.

Washback by design is explicit in LanguageCert assessment services and processes and is a fundamental consideration in developing tests and preparatory learning materials. LanguageCert supply learning and preparation materials to encourage test takers and their tutors not to prepare for the tests blind to the language skills necessary to succeed, and unclear on how they will be tested. 'By design' means the recognition and response to the need for positive washback in all processes for developing tests and their related learning materials. This approach ensures alignment between what language learners experience as they prepare for

LanguageCert tests, and what they experience in the exams. It also ensures that the skills learners practice for the tests have real-world validity and maximise learners' opportunities for success in their studies.

An overarching intention is to contribute to understanding how assessment might be used to improve educational outcomes. If the test is not fit for purpose, it is understandable that teaching (or learning) to the test can constitute negative washback in terms of a narrowing of the curriculum or a reliance on skills or knowledge which are irrelevant – nothing more than hurdles to clear in an exam scenario. However, in terms of educational outcomes, if the test is designed consultatively to meet the specific needs of stakeholders – including students, teachers, employers and policy makers – then LCA may be viewed as a test which accurately encapsulates curriculum objectives and as such reflects practical language use and therefore exerts positive impact. By promoting the honing and development of relevant skills in the realm of teaching and learning, assessment can be seen as the portal to opportunities to use the same skills in the real world as enablers of success, progression and transformation.

LanguageCert's ongoing research programmes assess and assure that washback is effective. Professor Tony Green, Director of CRELLA is leading this research. Together with Professor Liying Cheng, Director of the Assessment and Evaluation Group at Queen's University, Kingston, Ontario, he is a member of LanguageCert's Concordancing Studies Review Panel.

Domain Relevance

Domain specificity reflects ongoing research, benchmarking studies, and reviews to ensure that LanguageCert tests are relevant to, and representative of, the targeted domains. The approach draws on the wider language assessment literature and work specifically undertaken by LanguageCert is outlined below.

LanguageCert Academic derives from the established, regulated and internationally recognised, LanguageCert IESOL SELT C1. The IESOL SELT qualifications already reflect commonly accepted, best practice principles of language assessment, as well as meeting many requirements of the domain-specific stakeholders.

To ensure that these principles were being upheld, the IESOL qualifications were subject to independent evaluation in 2019 by UK NARIC against the relevant Common European Frame of Reference (CEFR) descriptors. Key considerations included linguistic complexity in terms of vocabulary, grammar and syntax; text domain and topic(s); authenticity; discourse type; text length; structure and presentation. UK NARIC identified a range of appropriate and relevant domains covered in the assessments, including personal, occupational, professional, educational, and public, with a good representation of input and output text types, including articles, adverts, diary entries, within personal, professional/ occupational, educational, and public domains. In the same way as for the IESOL qualifications, the evolved LanguageCert Academic test (together with LanguageCert General) was submitted to the UK's ECCTIS (Education Counselling and Credit Transfer Information Service) for external review.

LanguageCert has an Academic Panel to embed domain-specific expertise and experience into qualification design and ongoing review and development. The members of the panel provide a breadth of domain expertise spanning international education, academic admissions, English language teaching and accreditation, career readiness, and employability. Through regular reviews and consultation, the Panel supplies invaluable insight into the demands and expectations of each domain or sector, and how the tests can and do perform in those areas. This group also provides access to a wider network of specialists who are used to inform test

design and domain tailoring. The outputs of this insight and consultation are integrated with LanguageCert's test development processes, covering construction, rating, and grading.

Designing Tests that Measure Language Competence

The LanguageCert System of examinations test a range of different English language skills, sub-skills, and competences. The theoretical underpinning for how to achieve this comes from the works of Bachman and Palmer (2010), Canale and Swain (1980), and Weir (2005), amongst others. The internationally accepted CEFR model, which applies to language use and language learning, is also used.

The CEFR divides a learner's competences into General Competences and Communicative language competences. Communicative language competences are then further subdivided into three: Linguistic, Sociolinguistic and Pragmatic competences. These involve consideration not only of the communication, but also of the strategies used by learners, and hence the functional language skills learners demonstrate when they communicate. In a thought-provoking contribution to this area, Lampropoulou (Chapter 11, this volume), discusses a subset of Pragmatic Competence, namely Interactional Competence (IC). IC is discussed and described within the context of speaking skills where, it is proposed, IC can be assessed most usefully through the methodology of role-playing in a speaking skills task. The data were gathered from LanguageCert tests. This promising development is an example of how the LanguageCert research team constantly seeks innovative, improved and effective methods of assessing language proficiency skills.

When considering how to operationalise such theoretical models of language use, two factors which influence how a test looks are investigated: the authenticity of items, and the 'directness' with which competences are tested. Two important aspects of authenticity are situational and interactional authenticity. Situational authenticity refers to the closeness with which tasks and items represent language activities from real life; interactional authenticity refers to the naturalness of the interaction between test taker and task, and the mental processes required to carry out the task. The CEFR identifies a framework of six levels of communicative language ability as an aid to setting learning objectives and measuring learning progress or proficiency level. This conceptual framework contains a set of descriptor scales, expressed in the form of 'Can-Do' statements which give guidance to test developers.

Other important contextual features include characteristics of the test takers. When developing the structure and content of the LanguageCert tests, the target test population is considered. Examples are: typical age, cognitive development, and purpose of the learners in the process of language learning. This ensures that the materials are accessible, relevant, and interesting to engage with for the typical population for the test. The LanguageCert Academic exam aims toward learners wanting tertiary education study in an English-speaking environment (including where English is not the first language).

This approach enables LanguageCert to create tailored examinations which are set at an appropriate difficulty level for the intended candidature and desired outcomes, and that are relevant to the intended domain or context (e.g., English for academic purposes). These tests generate evidence in the form of results in each skill and overall, as well as define the ability level each individual test taker has shown in the test.

Tests in the LanguageCert System elicit samples of performance which are interpretable, based on a model of the test takers' competence. Test responses are scored to ensure the test taker's communicative ability in each skill measures against the LanguageCert Global Scale. This scale maps to the CEFR (through Can-Do statements and statistical analysis) and extrapolates both to the real world and equivalent language tests. The predictive validity of tests in the LanguageCert System allows receiving institutions and employers to assess how successful the test taker is likely to be in terms of coping with the language demands of a higher education course of study.

Testing the Domain Across the Four Skills

This section outlines domain relevance across the skills..

Developing Domain Relevance in the Listening Tests

The LCA Listening tests consist of 30 items across four parts. The range of content types in the IESOL Listening tests for C1 are appropriate for the targeted domain in terms of task types and robust statistical measurement and allow test takers to focus on content rather than familiarity with too many different activity requirements. Consequently, in the LCA test specifications, the main change to content is that all new tasks are focused on the target domain. For example, in Listening Task 3 test takers hear a lecture, rather than an informational talk and in Listening Task 4, test takers hear a multi-speaker discussion on an academic subject rather than a dialogue on a general topic. The test is designed to assess higher levels of comprehension, for example constructing meaning or making inferences when listening to a lecture or a conversation in a tutorial. The test comprises authentic listening materials including lectures, podcasts, interviews and discussions, on some abstract subjects, reflecting real-life demands of listening in an academic setting.

Range of Accents

Each Listening test uses a range of accents across the various parts of the examination, to ensure a test taker does not experience just one type of accent during their test.

The listening components use a range of accents drawn from the UK and other English-speaking countries, including North American, Australian, UK regional and national varieties, as well as other accents including Irish and South African.

The balance and proportion of accent representation also relates to the lengths of time different accents are heard during the tests.

The balance of accents also reflects the current markets for LanguageCert's test products. LanguageCert responds to target geographies where the test takers study or migrate to, and recognises where institutions reside. As the market is dynamic, this balance is continuously reviewed and integrated with the test development and maintenance programme.

There are checks and balances in LanguageCert's documented test creation procedures to ensure that an appropriate balance is achieved across test forms, and this is kept under review.

Developing Domain Relevance in the Reading Tests

The Reading tests consist of 30 items across four parts. The Reading test includes a range of content types, including multiple-choice questions, gap filling and multiple matching. The tasks include a range of source texts of different lengths relevant to the domains of the tests. Two of the IESOL SELT content types are unchanged and two new content types have been included to target level and domain more effectively.

Analyses of test efficacy indicated that the true/false task in the IESOL SELT specification would not have measured or discriminated sufficiently in an academic context. A short answer task in the IESOL SELT specification was also replaced. Instead, LCA includes a new Part 1, divided into Part 1a and Part 1b, both of which are vocabulary tasks. Part 1a is a multiple-choice task in which test takers read six sentences and replace a highlighted word in each sentence without changing the meaning. There are four options to replace each word. Part 1b is a multiple-choice cloze task in which test takers select the correct word or phrase to fill gaps in a short text. The focus of the new Part 1 tasks is on lexico-grammatical awareness of vocabulary and structures. For use in an academic context, sentences and texts are taken from academic documents, and so feature the language and structures used in the academic domain.

Language and context have been refined to increase relevant target language use in the different academic domains. Reading Part 2 (a multiple-matching task in which test takers select the correct sentences to complete gaps in a text) exemplifies this. Test takers must show understanding of how meaning is built up in discourse; thereby demonstrating their awareness of text organisation and discourse features. In Halliday (1994), emphasis is on the importance of analysing not just individual sentences, but also the relationships between them in order to understand how meaning is created in discourse. In this Reading task, the candidate needs to show awareness of how cohesive devices function to link sentences and paragraphs as well as understanding of the overall coherence, unity and continuity of the text. The two distractor sentences are written in the same style and on the same theme as the text. Together, the two distractors must fit in most of the gaps and can only be discounted by careful reading. Preparation for, and success in, this task type supports test takers' ability to tackle authentic academic texts. Successful test takers will be equipped with a strategic and analytical approach to understanding the organisation of ideas in discourse of this kind; knowing how meaning is structured in logical chunks and identifying the linguistic markers which will unlock the meaning of what they are reading.

Developing Domain Relevance in the Writing Tests

LCA contains two writing tasks. The focus of the first task is on the type of short report writing based on some data input (such as a table or graph) that a student in higher education will need to produce. The emphasis is on reporting on the data presented, explaining trends, and explaining likelihood and probability. The piece of writing needs to be succinct and may also include recommendations for future action. The second task focuses on the development of a longer piece of writing on an academic and/or topical matter. The test taker needs to produce a coherent piece of writing where they argue a position and draw a conclusion, requiring the candidate to show critical analysis, evaluation, communicate ideas effectively, support arguments and drawing on existing literature/frameworks for context.

Writing test quality was the focus of a study conducted by Coniam et al (Chapter 4, this volume) in which many facet Rasch analysis was used to explore consistency in marking and linkage and calibration to the CEFR. The

study found the two extended writing tasks writing tests from which LCA and LCG are derived, robust and fit for purpose. Indeed, the two extended writing task formats are well established in tests of English for academic purposes, including TOEFL and IELTS (Cumming, 2013), as they reflect a range of expository and descriptive task types encountered in academic contexts across disciplines.

Developing Domain Relevance in the Speaking Tests

Two changes have been made from the LanguageCert IESOL SELT C1 (Academic) in the LCA speaking test. The first is the introduction of a read-aloud task followed by a discussion of the topic. In the LCA test, this task centres on appropriate subjects which facilitate a tutorial type of discussion between the test taker and interlocutor. The second change is the amendment of the 'long turn' task at the end of the test. This is now more relevant to the academic domain by consistently featuring text types found in academia and by the introduction of a more 'formal' presentation. The opportunity to listen and respond to follow-up questions in real time in both these tasks also introduces an important feature of academic seminars, tutorials and other opportunities for academic discussion.

These changes expand upon a central component of the tests, which is the use of domain-specific role play to simulate and assess language competence in specific scenarios. Role play tasks are used in most of the LanguageCert Speaking suite of qualifications, as research has shown that they can imitate aspects of spoken language discourse in an authentic and realistic manner, and can be useful in measuring conversational competence as exhibited in the test takers' performance (Kormos, 1999). Okada (2010) discusses roleplay in Oral Proficiency Interviews (OPIs) in terms of its construct validity, and he describes the competencies displayed in performing a role play activity as highly resembling those observed in real-life conversations. He concludes by recognising roleplay as a valid assessment instrument. Lampropoulou (2023), demonstrates the value and efficacy of role-play in assessing Interactional Competence in LanguageCert examinations of speaking skills.

In the Speaking tests there are dedicated role-playing activities in Part 2. During these activities the interlocutor sets the context by informing the test taker of the scenario and the roles to be assumed. In the LCA test, role play tasks have the test taker interact with tutors concerning assignments, with a university accommodation officer about their accommodation options, or present them with a situation where they discuss student council matters with other college students. Scenarios also include arranging an outing with another student or discussing a journal article's recommendations.

These scenarios enable a high degree of domain authenticity, as the test tasks resemble the TLU domain. In the interactions described above, which can either be brief or develop unscripted for a longer period depending on the test taker's ability, a wider range of functions can be elicited than the interlocutor-structured interaction allows, such as asking for information, expressing regret, complaining, and offering and either accepting or rejecting an invitation for example (LanguageCert, 2020).

Developing Domain Relevance in the Marking Criteria

Domain-specific mark schemes are employed for LanguageCert Academic.

There are four separate criteria used for the marking of Writing:

1. Task achievement (and, for Academic only, Argumentation)
2. Organisation and coherence
3. Accuracy and range of grammar
4. Accuracy and range of vocabulary

In the marking of Speaking, the five separate criteria are:

1. Task Fulfilment and Communicative Effect
2. Coherence
3. Accuracy and range of grammar
4. Accuracy and range of vocabulary
5. Fluency, intonation, and pronunciation

While the criteria above may be seen to be universal, it is their application to each respective domain that differs. That application reflects the nature of the domain-specific tasks designed in the exams and outlined in this chapter. For example, under task fulfilment in the LCA test, the writing tasks require the ability to present relevant information, as well as expand upon and support key points, using a different style and tone. This approach flows across to the organisation, grammar, and vocabulary criteria, where a marking premium is placed upon the ability to create and sustain a logical flow, to convey meaning effectively, and use correct punctuation. This difference in focus is operationalised through the training of examiners using sample test taker scripts which illustrate the features referred to above, and in the mark schemes.

In high stakes exams such as LCA, it is essential for examiners to make informed and reliable expert judgements. The role of expert judgement in language test validation was examined by Coniam et al (Chapter 7, this volume), that established how examiner familiarity with items, standards and scales affects the accuracy of their judgement. Examiner training and standardisation documentation has been produced for LCA with these key findings in mind.

Reliability and Scoring

LCA reports performance across a wider range of levels than IESOL C1. This responds to demand from test takers and recognising institutions. LCA is focused on the B2 and C1 tests but also measures at B1 and C2. The test has an increased number of items from 26 to 30 in order to facilitate a greater spread of item difficulty and improve the ability to report with confidence across a range of CEFR skill levels.

Results are reported against the CEFR levels and on the LanguageCert Global Scale (Milanovic et al, 2023b). The Global Scale score (which is provided by language skill and overall result) gives finer gradations of performance within the CEFR levels but is also a standalone measure that can be aligned with any relevant external scale.

The Global Scale for reporting results has been established through the pretesting and live calibration of test materials at LanguageCert, and through the mapping of the Academic and General tests against other examinations in the same domains (for example IELTS) via the CEFR. The accuracy of these measures is determined and verified by the concordance study, currently in progress. The study examines the extent of overlap in content and performance between LCA and LCG and IELTS Academic and General Training tests. This study confirms a strong direct correlation between the tests of around 0.9.

The LCA test is a multi-level assessment, unlike the level-specific IESOL tests. Level-based tests however, can also typically measure across multiple levels. For instance, Lee et al (Chapter 3, this volume), has shown that the IESOL SELT level-based tests assess at their target CEFR levels, but also contain an appropriate number of items to allow assessment across levels. Specifically, the IESOL SELT C1 examination has items which assess above and below C1. Likewise, at the B2 level, there are items in the IESOL SELT B2 examination which assess both above and below B2. This feature is extremely useful for stakeholders who have to make decisions about candidates based on their results.

These findings are contained in a study by Lee et al., (Chapter 3, this volume) on aligning LanguageCert SELT examinations to the LanguageCert Item Difficulty scale in which the alignment of LanguageCert IESOL SELTs is explored in relation to the two objectively marked components of Listening and Reading. The use of externally referenced anchoring demonstrated the robustness of the four CEFR test levels B1–C2. For example, in the case of LanguageCert IESOL SELT C1 test, most accurate measurement was observed across two CEFR levels (B2 and C1) and reasonable measurement at the lower end of C2 and upper end of B1.

This ability to assess across multiple levels is enhanced in LCA (and LCG). Both tests' multi-level assessment capability has been enhanced by increasing the number of items in each test form. This has been done in the knowledge that the IESOL tests support accurate measurement across the two levels that each targeted, and reasonable measurement across four levels. By increasing the number of items in each of the General and Academic tests, accuracy has increased across levels. This enhancement also included refining the item types in the LCA Reading test; in particular the replacement of the True/False task. This refinement ensures that the full range of levels is tested effectively, and that all items discriminate well.

New materials target specific levels as defined in the Item Writer Guides (IWGs). The materials are created by experienced LanguageCert writers and reviewers. Used in combination with calibrated anchor items, Lan-

LanguageCert are confident that both tests assess across the stated ability range effectively. This is reinforced by ongoing research to locate all LanguageCert assessment products on its underpinning measurement scale, and aligning all LanguageCert products to the CEFR through which equivalence with other qualifications can be drawn.

LanguageCert estimates the standard error of measurement (SEM) for all tests, and uses it for each cut-score (the decision levels) in the Listening, Reading, Speaking and Writing skill tests.

Measurement Scale

The Global Scale is used to measure each test taker's performance. The Global Scale reports scores on a 0 to 100 scale. These levels of attainment can relate to overall performance in one examination, performance by skill or both these parameters. The Global Scale corresponds directly to LanguageCert's internal LID (LanguageCert Item Difficulty) scale.

The LID scale has been in use since 2016. It is a scale of item difficulty used for item banking and test construction purposes. Item difficulty values range from CEFR Pre-A1 through to high C2 level. The LID was developed using both expert judgement and statistical analyses. Eight expert consultants, each of whom have spent over 20 years writing, editing and vetting test materials to measure directly against the CEFR, completed a standards-setting exercise which generated anchor material to enhance and validate the scale. These anchor items then underwent trials and live tests, with all other items measured against them, thereby giving each item a difficulty value on the LID scale (See Lee et al, 2023a).

An in-depth analysis was conducted on all anchor items and a small number were eliminated from analysis and from further use as anchors, as they were not measuring as predicted. Rasch and Classical Statistical analyses were then carried out on all live and trial tests. By this method, many test items in the item bank are now considered fully calibrated. Research and validation studies in this area are contained in Coniam et al., (2021a) and Coniam et al., (2021b).

The Global Scale links to the LID scale and thereby the CEFR levels. In turn, this means that performance on LanguageCert tests is directly comparable to exams by other English language testing organisations, such as IELTS and Cambridge Advanced. Figure 3 illustrates how the Global Scale reports against the CEFR levels.

Figure 3: The LanguageCert Global Scale



In practice the LanguageCert Global Scale is operationalised in the test taker’s three-page test report (Appendix 2).

The Global Scale allows ease of interpretation for test users and a finely tuned results service across all language skills. As shown, performance can be separated in each skill and overall, so that a test taker is not only described as having ‘B2 ability’, but a more precise level of detail is provided on test taker’s performance. The

Test Report shows an overall score, the overall CEFR level of attainment reached, and the score for each of the skills using both the Global scale and the CEFR level of attainment.

The Global Scale, launched with the LanguageCert Test of English (LTE), measures from pre-A1 to high C2 (i.e., across the full 0 –100 range). The LTE has been successfully administered to tens of thousands of test takers

worldwide, and the Global Scale has received good customer feedback in terms of its simplicity, clarity, and ease of use.

Items in the Reading and Listening tests range in difficulty from CEFR level B1 to C2, with the vast majority of items focusing on the B2 and C1 levels (Vocational to Proficient). The difficulty of items is established through pre-testing and live test calibration using Rasch and Classical Statistical analysis. All Reading and Listening items are calibrated to the LID (LanguageCert Item Difficulty) scale (and hence the LanguageCert Global Scale) which runs from CEFR Pre-A1 to C2 levels. Examples of the ways in which items are calibrated using Rasch and Classical Statistical analysis are described in a large number of chapters in Falvey and Coniam (2023 - this volume), and reveal that this method of calibration is demonstrably more efficacious than Classical Statistical analysis on its own.

Each LCA Reading and Listening test is designed to cover a wide range of the B2/C1 CEFR 'syllabus' (i.e., those areas covered by the Can-Do statements in the CEFR). A broad range of Reading and Listening sub-skills are tested, as is a range of grammar, vocabulary, and awareness of functional language. Tasks are set in contexts that are appropriate for the nature of the candidature and the desired outcomes of the test. That is, the LCA test has items and tasks largely set in the academic domain (i.e., contexts that are relevant to test takers intending to study in higher education).

For the LCA Writing and Speaking tests, detailed mark schemes are used by examiners. In terms of Writing, test takers complete two writing tasks. Task 1 requires test takers to respond to a visual and textual input and then produce an extended piece of writing of 150 to 200 words describing the data and predicting future trends. In Task 2, the test taker must produce a longer piece of discursive writing of around 250 words to address a topical issue which has a general academic context, e.g., the use of alternative energy forms or methods of education. The test taker is expected to argue a position and strengthen their argumentation with examples and supporting ideas.

In the marking of Writing, candidates are assessed against four criteria. These are:

1. Task Achievement on Task 1 and Task Achievement and Argumentation on Task 2
2. Accuracy and Range of Grammar used
3. Accuracy and Range of Vocabulary used
4. Organisation

The use of separate criteria to measure different aspects of Writing performance allows the LCA test to deliver rich feedback to both test takers and receiving organisations, and provides indications as to where further development is needed by the test taker. The marking criteria have been adapted from the LanguageCert IESOL C1 examination Writing marking criteria. At the outset, the criteria were based on the descriptors for Writing in the CEFR in conjunction with the nature of the task. These original criteria have been developed over many years, with active consideration of their relevance and applicability. Feedback has been collected from trainers, examiners, and examiner-monitors (senior examiners) to finetune the wording of the criteria so that examiners find them easy to use, so that they reflect test taker output, and so that the key features expected from test takers in the exam at each CEFR level are considered.

The evolved and current IESOL C1 Writing marking criteria were then adapted to better suit the academic context. For example, argumentation has been added to the Task 2 'Task Achievement and Argumentation' criteria to reflect the nature of academic writing.

The criteria have also been extended to measure performance across a broader range of ability (from A2 to C2) to report reliably across an extended range of CEFR levels.

Writing scripts are marked by two human examiners. If there is a significant difference in mark awarded, the script is passed to a third (more senior) examiner whose marks are final. It is intended that, in the medium to longer-term, auto-marking by computer will be introduced as part of a hybrid scoring solution.

For Speaking, the test is split into four parts. Part 1 involves responding to questions across a range of topics. In Part 2, the test taker takes part in two role-plays which are set in an academic setting. In Part 3, the test taker reads aloud a short piece of writing of around 100 words in length. The extract is the type of primary source or reading that a student may be asked to read out in a tutorial, for example. In Part 4, the test taker is provided with some visual and textual input and asked to provide a two-minute talk relating to the information

In the marking of Speaking, test takers are assessed against five criteria. These are:

1. Task Fulfilment and Communicative Effect
2. Coherence; Accuracy and Range of Grammar
3. Accuracy and Range of Vocabulary
4. Fluency, Intonation
5. Pronunciation

Just as for Writing, the use of separate criteria to measure different aspects of Speaking performance allows the LanguageCert Academic test to deliver rich feedback to both test takers and receiving organisations and provides indications as to where further development is required on the part of the test taker.

The marking criteria have been adapted from the IESOL C1 Speaking test marking criteria. At the outset, the criteria were based on the descriptors for Speaking in the CEFR, in conjunction with the nature of the tasks. These original criteria have been developed over many years of use, with active consideration of their relevance and applicability. Feedback has been taken from trainers, examiners, and examiner-monitors (senior examiners) to fine-tune the wording of the criteria so that examiners find them easy to use, so that they reflect test taker output, and so that the key features expected from test takers at each CEFR level are considered.

The evolved IESOL C1 Speaking marking criteria were then adapted to better suit the academic context. For example, greater emphasis has been placed on coherence and fluency which are important features in a higher educational setting where students need to provide well-structured talks and responses to questions in a tutorial. The criteria have also been extended to measure performance across a broader range of ability (from A2 to C2).

Currently, test taker output in the Speaking test is marked by two human examiners; by the interlocutor immediately after the test and by a second examiner who awards marks subsequently by accessing the video record-

ing. The first criteria 'Task Fulfilment and Communicative Effect' is marked by the interlocutor and provides more of a 'general impression' score, while the second examiner marks the other criteria. The interlocutor general impression mark is then double-weighted. If there is a significant difference in marks awarded, then the recording goes to a third (more senior) examiner whose marks are final.

In the medium to longer-term, auto-marking by computer is being planned to be introduced as part of a hybrid scoring solution. A hybrid assessment model will garner the proven benefits of both human and machine marking.

Methodology

LanguageCert's Assessment Development department contains academics as well as professional linguists and assessors, who publish research on all aspects of our language qualifications. An Advisory Council supports this team and helps it to meet regulatory obligations to bodies such as Ofqual.

All tests and test items are constructed and assured using high-calibre writers operating to clear guidelines, workflows, and quality assurance protocols which include layers of reviews, editing, statistical analyses, and vetting. The proprietary item bank is used to manage all LanguageCert's tests, with strict access protocols, and robust workflows for process compliance. LanguageCert's team of markers includes expert Chief Examiners as well as Markers and their Team Leaders. All undergo stringent training before marking live papers. A defined marking process operates within the proprietary marking application, which standardises, and quality assures the process and its outputs. All test taker digital, audio and video interactions during tests are recorded and securely stored so that there is a verifiable evidence base for all results. In addition, robust quality assurance protocols are applied to secure integrity and fairness for the test and the test taker.

Bias

LanguageCert uses Differential Item Functioning (DIF) analyses to explore whether any subgroup of test takers sitting a test is being unfairly disadvantaged. Investigating DIF is key to understanding and dealing with test bias (Coniam and Lee, 2021).

In Coniam and Lee (2021), DIF analysis took place on IESOL exams delivered from 2018 to 2021. This population contained IESOL exams delivered for the UK Government's UKVI scheme. For each CEFR level four variables were explored: native language, age, gender, and test centre. The DIF analysis used Rasch measurement, with DIF strength reported in line with Zwick et al. (1999).

For gender – typically a key variable in the exploration of DIF – there was a very low incidence of 3% DIF. An examination of Reading or Listening items indicated that there was no significant DIF in either skill. With the findings confirming that the LanguageCert tests analysed exhibit low levels of gender bias, a methodology is in place for the ongoing monitoring of DIF on all LanguageCert exams. Native language and age showed mod-

erate-to-large DIF. This, however, is likely to be due to these two categories being diverse with only very few entries from small sub-test populations.

As an international organisation, LanguageCert strives to ensure its tests are valid, reliable and have a positive impact on learners. An important part of ensuring fairness to test takers is to minimise any bias in the test materials. The process of eliminating bias begins with the formation of the test specifications. These are written with direct reference to the nature of the intended or anticipated candidature to ensure the tests are fully fit-for-purpose. This detail is checked at annual reviews and when the test formats are revised. LanguageCert makes sure writers understand who the target domain test users are, and that they consider aspects such as the level of cognitive processing of typical test takers, and their cultural contexts.

LanguageCert's Item Writer Guides and the training process stress bias awareness, and the requirement to produce materials which will not favour or discriminate against certain test takers. This entails ensuring test materials are as free from specific regional or national cultures as possible, and that topics are universal. Item writers have a list of taboo topics to aid in this. These taboo topics include areas which may cause distress or distraction to test takers, or relate to unfortunate experiences they have suffered (e.g., war or drugs), through to specific aspects of local cultures (e.g., milkmen in Britain) which may be alien to the local culture of the test taker or beyond their life experience. The LanguageCert team also take care to not introduce test material which may test general knowledge or specific technical knowledge, rather than language ability.

Ongoing Development, Monitoring and Evaluation

Ongoing stakeholder engagement is crucial in the continuous development of LCA. The LanguageCert Academic Panel, which sits under the LanguageCert Advisory Council, convenes quarterly, bringing together experts from across the higher education sector and a range of geographical regions to provide guidance, critiques and feedback on the development and delivery of the qualification. Panel members share feedback derived from their experience and expertise in the international higher education sector and provide insights into key challenges and opportunities relating to career-readiness and employability.

In addition to the input of the Academic Panel, feedback is provided by way of regular webinars, presented by development staff to stakeholders such as institutional administrators, admissions tutors and other key personnel involved in the admission, tutoring and mentoring of successful candidates coming to the UK for education purposes. LanguageCert disseminate findings of their research and invite comment and participation via a quarterly update from the assessment research and validation team, Research Insights. This publication also has a role in communicating and inviting dialogue with our stakeholders and Language Cert Academic and LanguageCert General research will become a regular feature in this publication as the roll-out is widened which may test general knowledge or specific technical knowledge, rather than language ability.

Conclusion

This chapter describes how an examination evolution occurs. It provides the rationale for the evolution, its purpose and the needs it meets, the curricular factors in play, the development of the examination, and its pretesting, piloting and eventual offering to the public. LCA is closely based on an existing examination, the LanguageCert IESOL C1. Its revision from a general English test to one that is more targeted to an academic context is described here in some detail as is a significant body of research that has informed and guided the redevelopment.

The development of the IESOL Academic and General tests, described here, focuses on academic language requirements, developed by LanguageCert personnel and pre-tested and piloted internationally, at LanguageCert-approved test centres under secure test-taking conditions, with pretesting populations which are representative of each test's intended candidature. The chapter's content is indicative of the care taken to employ the best research findings, methodology, and statistical tools in order to develop and improve the quality of all LanguageCert examinations.

References

- Bachman, L. F., & Palmer, A.S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.
- Cheng, L. & Sultana, N. (2022). Washback: Looking backward and forward. In Fulcher, G. & Harding, L. (Eds.). *Routledge Handbook of Language Testing*.
- Coniam, D., & Lampropoulou, L. (2020). A review of LanguageCert IESOL Listening and Reading test reliabilities 2018-2020. London, UK: LanguageCert.
- Coniam, D., & Lee, T. (2021). Potential bias in LanguageCert IESOL items: A Differential Item Functioning analysis. London, UK: LanguageCert.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021a). Validating the LanguageCert Test of English scale: The paper-based tests. London, UK: LanguageCert.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021b). Validating the LanguageCert Test of English scale: The adaptive test. London, UK: LanguageCert.
- Coniam, D., Lee, A., & Milanovic, M. (2023a). Exploring item bank stability in the creation of multiple test forms. In Falvey, P & Coniam D. (eds.). *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK.
- Coniam, D., Lee, A., Milanovic, M., Pike, N., & Wen Zhao. (2023b). The role of expert judgement in language test validation. In Falvey, P & Coniam D. (eds.). *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK
- Coniam, D., Stoukou, I., Lee, A., & Milanovic, M (2023c). LanguageCert SELT Writing Test quality. In Falvey, P & Coniam D. (eds.). *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol .2. LanguageCert: London: UK

- Cumming, A. (2013) *The cognitive validity of the IELTS Academic Writing task*. IELTS collected papers. Cambridge: Cambridge University Press.
- Falvey, P., & Coniam, D. (eds.) (2022). *Certifying quality in assessment: Research and validation at LanguageCert*, Vol. 1. London, UK: LanguageCert.
- Falvey, P., & Coniam, D. (eds.) (2023). *Certifying quality in assessment: Research and validation at LanguageCert*, Vol. 2. London, UK: LanguageCert.
- Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education: 25*. Studies in Language Testing, Series Number 25 Cambridge: Cambridge University Press.
- Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing*, 28, 39-52.
- Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing*, 16(2), 163–188.
- Lampropoulou, L. (2023). Interactional competence and the role roleplay plays: The LanguageCert perspective. In Falvey, P. and Coniam C. *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK.
- Lee, A., Papargyris, Y., Milanovic, M., Pike, N., & Coniam, D. (2023a). Aligning LanguageCert SELT tests to the LanguageCert Item Difficulty scale. In Falvey, P. and Coniam C. *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK.
- Lee, A., Coniam, D., & Milanovic, M. (2023b). Exploring item bank stability through live and simulated datasets. In Falvey, P. and Coniam C. *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK.
- Milanovic, M., Lee, A., Coniam, D., Papargyris, Y. (2023a). Externally-referenced anchoring of LanguageCert SELT tests. In Falvey, P & Coniam D. (eds.). *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert Vol 2*. LanguageCert: London: UK.
- Milanovic, M., Pike, N., Lee, T., & Coniam, D. (2023b). *The LanguageCert Global Scale*. London, UK: LanguageCert.
- Okada, Y. (2010). Role play in oral proficiency interviews: Interactive footing and interactional competencies. *Journal of Pragmatics*, 42, 1647–1668.
- Turner, J. (2004). Language as academic purpose. *Journal of English for Academic Purposes*, 3(2), 95-109.
- Turner, J. (2012). Providing a space for the socio-political dynamics of EAP. *Journal of English for Academic Purposes*, 11(1), 17.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndmills, UK: Palgrave Macmillan.
- Xi, X., & Norris, J. M. (Eds.). (2021). *Assessing academic English for higher education admissions*. Routledge.
- 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.

Appendix 1: Reference List of Specific Skill-based Studies

Listening

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). TOEFL 2000 listening framework. Educational Testing Service.

Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In *IELTS Collected Papers, 2*. Cambridge University Press.

Reading

Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). TOEFL 2000 reading framework. Educational Testing Service.

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. In *IELTS Research Reports 9*. British Council and IELTS Australia.

Writing

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). TOEFL 2000 writing framework. Educational Testing Service.

Nesi, H., & Gardner, S. (2018). The BAWE corpus and genre families classification of assessed student writing. *Assessing Writing*, 38, 51-55.

Speaking

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). TOEFL 2000 speaking framework. Educational Testing Service.

Brown, A., & Ducasse, A. M. (2019). An equal challenge? Comparing TOEFL iBT™ speaking tasks with academic speaking tasks. *Language Assessment Quarterly*, 16(2), 253-270.

Appendix 2: Sample Candidate Test Report



LanguageCert Academic (Listening, Reading, Writing, Speaking)

Test Report

Candidate Information

Last Name:	Candidate's Last Name		
First Name:	Candidate's First Name		
Date of birth:	xx Month xxxx		
Candidate Number:	99800...		
Candidate URN:	PPC/...		
ID Type:			
ID Number:		Nationality:	

Test Centre Information

Date of Test:	xx Month xxxx	Date Test Results issued:	xx Month xxxx
Test Centre number:		Test Centre country:	
Mode of Delivery:			

Candidate Results (out of 100 on the LanguageCert Global Scale)

Listening		Writing	
Reading		Speaking	
Total Score			
CEFR Level			

Marios Molfetas
LanguageCert
Responsible Officer

THIS IS NOT A CERTIFICATE

LanguageCert is a business name of PeopleCert Qualifications Ltd, UK company number 09620926
LanguageCert reserves the right to amend the information given before issuing certificates to successful candidates.

info@languagecert.org

Candidate Performance Feedback (Writing Part 1)

Task Fulfilment	
Accuracy and Range of Grammar	
Accuracy and Range of Vocabulary	
Organisation and Coherence	

Candidate Performance Feedback (Writing Part 2)

Task Fulfilment	
Accuracy and Range of Grammar	
Accuracy and Range of Vocabulary	
Organisation and Coherence	

Candidate Performance Feedback (Speaking)

Task Fulfilment and Communicative Effect	
Coherence	
Accuracy and Range of Grammar	
Accuracy and Range of Vocabulary	
Pronunciation, Intonation and Fluency	

CEFR Level	Scaled Score	Performance Descriptors (Listening, Reading, Speaking, Writing)
C2	90 - 100	<ul style="list-style-type: none"> • Can understand with ease any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided there is a familiarity with the accent. • Can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works. • Can write clear, smoothly flowing complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points. • Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
C1	75 - 89	<ul style="list-style-type: none"> • Can understand extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly. • Can understand long and complex factual and literary texts, appreciating distinctions of style. • Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices. • Can express him/herself fluently and spontaneously without much obvious searching for expressions.
B2	60 - 74	<ul style="list-style-type: none"> • Can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. • Can read articles and reports concerned with temporary problems in which the writers adopt particular attitudes or viewpoints. • Can use the language fluently, accurately and effectively on a wide range of general, academic, vocational or leisure topics, marking clearly the relationships between ideas. • Can communicate spontaneously with good grammatical control without much sign of having to restrict what he/she wants to say, adopting a level of formality appropriate to the circumstances.
B1	40 - 59	<ul style="list-style-type: none"> • Can understand the main points of clear standard speech on familiar matters regularly encountered in education, work and leisure, etc. • Can understand texts that consist mainly of high frequency everyday or job-related language. • Can produce simple connected text on topics which are familiar or of personal interest. • Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
A2	20 - 39	<ul style="list-style-type: none"> • Can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance • Can read and understand very short, simple texts such as personal letters • Can give a simple description of people, daily routines, likes/dislikes etc. as a short series of simple phrases and sentences linked into a list. • Can write a series of simple phrases and sentences linked with simple connectors like 'and,' 'but' and 'because'.
A1	10 - 19	<ul style="list-style-type: none"> • Can recognise very familiar words and phrases when people speak slowly. • Can read and understand very simple sentences on familiar topics. • Can produce simple mainly isolated phrases about people and places. • Can write simple isolated phrases and sentences.
<p>The above descriptors are adapted from the Common European Framework of Reference for Languages (2018). Text from these is reproduced by kind permission of the Council of Europe.</p>		



LanguageCert Global scale	CEFR	LanguageCert General	LanguageCert Academic	LanguageCert Global scale
100				100
99				99
98				98
97				97
96				96
95				95
94				94
93				93
92				92
91				91
90			90	90
89				89
88				88
87				87
86				86
85				85
84				84
83				83
82				82
81				81
80				80
79				79
78				78
77				77
76				76
75		75	75	75
74				74
73				73
72				72
71				71
70				70
69				69
68				68
67				67
66				66
65				65
64				64
63				63
62				62
61				61
60		60	60	60
59				59
58				58
57				57
56				56
55				55
54				54
53				53
52				52
51				51
50				50
49				49
48				48
47				47
46				46
45				45
44				44
43				43
42				42
41				41
40		40	40	40
39				39
38				38
37				37
36				36
35				35
34				34
33				33
32				32
31				31
30				30
29				29
28				28
27				27
26				26
25				25
24				24
23				23
22				22
21				21
20		20		20
19				19
18				18
17				17
16				16
15				15
14				14
13				13
12				12
11				11
10				10
9				9
8				8
7				7
6				6
5				5
4				4
3				3
2				2
1				1
0				0

THIS IS NOT A CERTIFICATE

LanguageCert is a business name of PeopleCert Qualifications Ltd, UK company number 09620926
 LanguageCert reserves the right to amend the information given before issuing certificates to successful candidates.

info@languagecert.org





SECTION 2: CALIBRATION/ VALIDATION STUDIES



Chapter 2: Externally-Referenced Anchoring of LanguageCert SELT Tests

Michael Milanovic, Tony Lee, David Coniam and Yiannis Papargyris

Abstract

This chapter reports on the use of externally-referenced anchoring by LanguageCert as a methodology for vertically aligning test forms: i.e., aligning test forms to a calibrated midpoint. External anchoring is the method that has been developed to fill an analytical gap. The standard way of carrying out test form comparison is by using common items that are set in the tests to be compared. The advantage of using externally-referenced anchoring is that it allows analysts to align test forms that have no common items.

This study presents the analysis of a sample of the Listening and Reading test forms. The test forms comprise those LanguageCert SELT tests that assess at CEFR levels B1–C1. Using Rasch measurement to vertically align tests on the basis of prior expert judgement (Lee et al., 2022), the robustness of the LanguageCert SELT B1–C1 tests is illustrated. An analysis of the test forms reveals three findings of close matches:

1. between the items in the different test forms
2. between the test forms and the LanguageCert Item Difficulty (LID) scale
3. And, as a consequence, between the test forms and the respective CEFR levels

The results provide support for the claim that LanguageCert SELT tests are well set, with each test appropriately positioned at its respective CEFR level. The results also demonstrate how high-stakes tests that have no common items can be aligned with each other, the LID scale and respective CEFR levels.

Keywords: externally-referenced anchoring, SELT, IESOL, listening tests, reading tests, Rasch

Introduction

In Chapter 1, Jones describes the development of a dynamic response to changing stakeholder expectations and requirements by refocusing LanguageCert IESOL C1 for a new purpose: academic study at undergraduate, post-graduate or professional level. For purposes of security in such high-stakes examinations, separate test forms are developed. The issue of comparability is, thus, of vital importance. In order to address this issue, the current chapter describes, a procedure to ensure the reliability of different test forms for the same test. The procedure has been developed to ensure the reliability and comparability of tests even when different test forms share neither common items nor common candidates. As explained above, the reason for the absence of common test items and candidates in high stakes tests is to ensure that strict test security of the tests is maintained.

The chapter extends LanguageCert's exploration of quality in its examinations (see e.g., Coniam et al., 2021a; 2021b). Considerable importance is now attached to English language qualifications for work and study; this is reflected by the UK Visas & Immigration (UKVI) department establishing Secure English Language Tests (SELT) for candidates who wish to move to and/or work to the UK. LanguageCert was approved in 2020 as a provider of UK Home Office approved SELT tests and offers LanguageCert SELT (LST) four-skills tests at a range of levels. These levels are mapped to the Common European Framework of Reference (CEFR) for UK Visas & Immigration (UKVI) worldwide. They cover all visa requirements for living, working or studying in the UK.

In line with the type of visa being applied for to the UKVI, a language test exhibiting proof of competency in English at a particular level must be passed. Against this backdrop, this chapter examines the statistical quality of the LST B1–C1 Listening and Reading Tests, approved for UKVI language certification purposes, which were produced over the period 2020–2021. All test forms comprise 52 items.

Against the key test qualities of validity and reliability (Bachman and Palmer, 2010), central validity issues include how well the different parts of a test illustrate what a test taker can do – i.e., communicate – in English, and how well test scores provide an indication of test taker ability in relation to communicative language competence (Messick, 1989; Bachman and Palmer, 2010). The LST tests assess the communicative skills that test takers will be expected to control at particular levels of ability (i.e., in relation to the CEFR). Test content is designed to match target test takers – in terms of grammar, functions, vocabulary, topics etc., and the tasks have correspondingly relevant 'communicative' contexts.

If tests are to be of high validity and reliability, they need to be well constructed (Hughes, 2003). In this regard, LanguageCert test item writers are of the highest international standard and have extensive expertise in, and knowledge and understanding of, the different CEFR levels (see Papargyris and Yan, 2022). Test items are linked to the CEFR by expert judgement, a methodology which has proven – as long as adequate training and standardisation are in place – to be robust (Coniam et al., 2022).

The LST B1–C1 test forms analysed in this study constitute a sample of the test forms delivered by LanguageCert in the 18-month period from mid-2020 to late 2021. For security purposes, all LST Listening and Reading tests are currently constructed as standalone tests. Since test security mandates that test forms are separate from one another, there are no linking items or test takers by which direct cross-calibrating may be conducted. Fortunately, to address this issue, the externally-referenced anchoring methodology pioneered

by Lee et al. (2022) permits tests which have no common linking items to be vertically linked against the test's midpoint using item values, previously-established by expert judgement. It is therefore this methodology – externally-referenced anchoring – which is used in the current study to explore and illustrate how accurately the different LST B1–C1 test forms are anchored onto the LanguageCert Item Difficulty (LID) scale, and hence to the CEFR.

The key to establishing appropriate points on the LID scale involves the use of expert setters and their concomitant expert judgement. Such 'expert judgement' in language assessment is therefore a key factor in 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. While there has been debate over the use of expert judgement in standard setting (e.g., Alderson and Kremmel, 2013), generally, the use of expert judgement has been accepted as having a valid role in the field of language assessment for test validation and standard setting – see Lumley, 1993; Bachman et al, 1995. More recent validation studies involving expert judgement include VanderVeen et al. (2007), Song (2008), Gao and Rogers (2011), and van Steensel et al. (2013). In these studies, judges were reported to have reached high levels of agreement. The positive use of expert judgement is reflected in Lee et al.'s (2021) study utilising externally-referenced anchoring with other LanguageCert CEFR-related tests – the IESOL suite of tests (see also Coniam et al., 2022).

The use of expert judges is an important feature of the LanguageCert Academic test, particularly in the case of its speaking task.

The LanguageCert SELT Tests

The LST suite comprises tests at CEFR levels B1 to C2. Examination specifications reflect the requirements of the CEFR. These requirements demand that test materials writers have extensive expertise in, and knowledge and understanding of, the CEFR.

Each LST test has a designated CEFR level, with, as mentioned, all test forms carefully set using expert judgement and reviewed by other expert staff in the LanguageCert Assessment Team. The LanguageCert Item Difficulty (LID) scale referred to above 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 both 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 may be found in Table 2 below.

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

The four-skills LST tests are located on the LanguageCert Global Scale [Note 1] along with other LanguageCert test products: the LanguageCert Test of English, and the International IESOL suite of English language tests.

The methodology surrounding externally-referenced anchoring relates to the use of Rasch measurement. An overview of Rasch can be found in the Glossary of statistical terms and techniques (the final chapter of the current volume), along with an outline of the infit and outfit mean square statistics which are key to the interpretation of Rasch results in the context of data ‘fit’.

Externally-Referenced Anchoring, CEFR Levels and Test Forms

As mentioned previously, the methodology used in the current study is based on externally-referenced anchoring (ERA) (Lee et al., 2022). In ERA, test forms, which have no common items but comprise items which have been set at predefined and well-accepted CEFR levels, are anchored using the calibrated midpoints of a test form against the LID scale and against the CEFR. For each test level, the frame of reference (see Humphry, 2006) constitutes the respective CEFR scale locations calibrated through the test forms and items for that level.

Table 1 below first provides detail on the number of test forms and their candidatures analysed.

Table 1: LST test forms and candidatures

CEFR level	Test forms	Candidates
B1	9	10,808
B2	6	2,732
C1	6	581

The focus in the current study is B1 to C1. Due to a comparatively small candidature, the C2 test forms do not form part of the current analysis.

The analysis in the study examines nine test forms at LST B1 level, six at B2 and six at C1. There are, as mentioned, for reasons of security, no linking items or test takers by which cross-calibrating may be conducted within or across test forms or levels. In the current study, ERA uses the calibrated midpoints of B1–C1 on the LID scale to explore the anchoring of these LST levels on the LID scale, and against CEFR levels. LID scale ranges and midpoints for the three CEFR levels explored are presented in Table 2.

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

On the basis of vertical midpoint anchoring, ERA:

- enables an effective calibration of the items in each test form – given that no other restrictions are imposed on the items.
- reveals the items' goodness of fit between expertly-assigned values and calibrated item distributions.

The anchoring goodness of fit is then evaluated by two metrics:

- 1) The extent to which a test's midpoint corresponds to the LID scale level.
- 2) The fit in terms of the extent to which the item distribution around a test's midpoint includes most of the items in a given test. Such fit is determined by a broadly bell-shaped distribution of item measures with the majority of item measures being clustered around the mean and falling between the 25th to 75th percentiles.

Research Questions

The research questions pursued in the current study are:

1. Do good Rasch infit and outfit statistics emerge from the externally-referenced anchoring of the LST B1–C1 test forms?
2. Do broadly bell-shaped item measure distributions emerge on the LST B1–C1 test forms?

Background Statistical Analysis

The reader is referred to the outline of the Rasch measurement model provided in the Glossary of statistical terms and techniques at the end of the volume.

Item Infit and Outfit

Analysis in the current study has been conducted via the Rasch analysis software Winsteps (Linacre, 2018). Appendices 1, 2 and 3 provide details of fit statistics. The majority of the items in all LST B1–C1 test forms had infit and outfit fit statistics 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, less than 5% of the items on any one test.

Reliability

Test reliability, for a 50-item test, is proposed as being 0.7 or above (Ebel, 1965). The equivalent of classical test reliability in Rasch is person reliability (Anselmi et al., 2019). As Appendices 1–3 illustrate, 0.8 or better was achieved by all LST B1–C1 test forms. This indicates that satisfactory test reliability has occurred in the data available for this study.

These two sets of background statistics are indicative of a set of robust, well-constructed tests. This means that the picture of test robustness confirms that the externally-referenced anchoring is being conducted against a backdrop of reliable tests.

Externally-referenced Anchoring Results

Test means and measures that emerged after the externally-referenced anchoring procedure are now examined, in particular means recorded at the 25th and 75th percentiles. Since in Rasch measurement the starting point of measurement is the mid-point (the 50th percentile), the CEFR level of a test should start in the middle of the item distribution. The central 50% of the item distribution (25% to 75% of the items) is therefore the range most precisely measuring the CEFR level. Such a distribution shows that 50% of the items are well targeted at the CEFR level intended by the test, and means that half of the items in the test are around the CEFR level presumed by the test.

Ideally, the 25th percentile will be located half a logit (10 LID scale points) below and the 75th percentile half a logit above the test midpoint (Lee et al., 2022).

Two sets of linked analyses are presented below. The first set provides a summary of percentile distribution values; the second provides a more visual impression in the form of item difficulty distribution graphs.

Percentile Distribution Values

Summary analyses of the LST B1–C1 test forms in table form are presented in Tables 3–5 below. Acceptable values are in green font; values which are greater than five LID scale points (a quarter of a logit) away from the established range are in red font.

Table 3 provides the relevant detail for the B1 level test forms.

Table 3: Percentile distributions in LST B1 test forms

	T206	T207	T208	T209	T384	T409	T414	T446	T593
Mean	100.00	100.01	100.00	100.00	99.99	100.00	100.00	100.00	100.00
SD	20.72	19.95	20.14	19.59	20.57	25.26	24.64	20.88	21.03
Maximum	159.34	145.40	139.98	141.43	150.09	157.75	175.02	138.25	158.75
75th percentile	117.08	111.89	116.59	113.48	116.51	115.78	118.29	115.14	112.91
50th percentile	98.92	101.33	100.66	99.60	97.07	103.78	97.17	97.71	100.54
25th percentile	87.72	90.97	83.65	85.17	86.95	82.36	82.51	86.32	85.99
Minimum	56.24	54.60	62.72	48.86	63.40	40.67	48.48	47.06	41.20

As can be seen, at the 25th percentile, all nine test forms are acceptably close to the lower scale range of 91. At the 75th percentile, there is some divergence, with six test forms showing a diverge of more than 5 LID scale points above the top of the LID scale range of 110 – in particular Tests T206 and T414. Nonetheless, the divergence seen is within half a logit (10 LID scale points) (Zwick et al., 1999), which means that the divergence is within acceptable bounds.

Table 4 provides the relevant detail for the B2 level test forms.

Table 4: Percentile distributions in LST B2 test forms (LID scale range: 111-130; midpoint: 120)

	T211	T219	T220	T363	T385	T421
Mean	120.00	120.00	120.00	120.00	120.00	120.00
SD	23.13	23.60	20.91	19.94	20.21	17.53
Maximum	183.97	172.19	186.28	189.18	156.26	153.73
75th percentile	134.75	134.11	130.88	131.22	138.34	132.54
50th percentile	118.92	120.46	117.59	118.83	120.15	117.87
25th percentile	103.95	102.19	109.34	107.21	102.35	107.80
Minimum	84.77	69.00	82.48	78.75	80.70	84.38

At the 75th percentile, all six test forms are close to the upper scale range of 130. At the 25th percentile, there is more divergence, with three test forms showing a diverge of more than 5 LID scale points – in particular Tests T219 and T385. Such divergence is, however, within half a logit of difference, despite some items being slightly easier than intended in three of the tests. Thus, the divergence can be accepted as being within acceptable bounds.

Table 5 provides the detail on C1 level test forms.

Table 5: Percentile distributions in LST C1 test forms (LID scale range: 131-150; midpoint: 140)

	T210	T222	T356	T364	T386	T588
Mean	140.00	140.00	140.00	140.00	140.00	140.00
SD	16.26	21.97	19.59	18.35	18.78	21.29
Maximum	175.56	196.41	190.32	179.01	186.88	190.73
75th percentile	152.56	151.16	152.73	152.08	155.40	148.38
50th percentile	140.40	140.04	136.16	142.24	140.20	140.71
25th percentile	127.75	127.75	125.85	125.16	126.98	126.79
Minimum	106.72	73.50	104.07	102.35	102.05	100.32

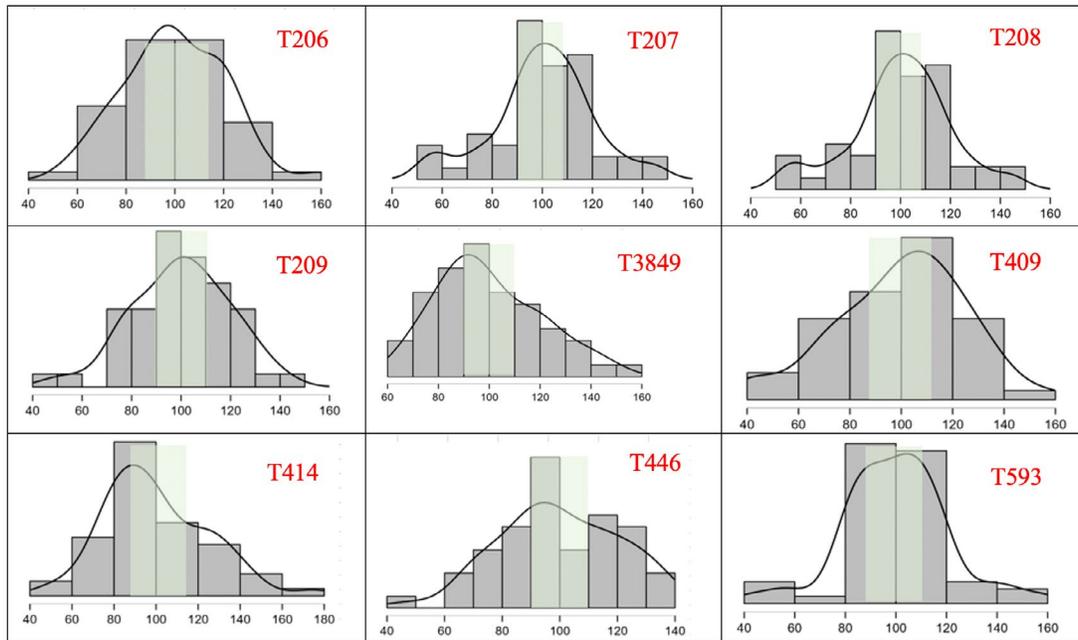
The C1 test forms show a close match with their LID scale ranges. At both 25th and 75th percentiles, all six test forms are close to the upper and lower scale ranges of 150 and 131. This means that all six tests have been well targeted at the C1 level.

Item Difficulty Distribution Graphs

To provide an accessible visual impression, item difficulty distributions are now presented in graph form in Figures 1–3. The green shading denotes the LID scale range for each test form. Frequency trend lines included across the scale for each test form provide a visual indication of the general shape of the distributions.

Figure 1 presents the item difficulty distributions for LST B1.

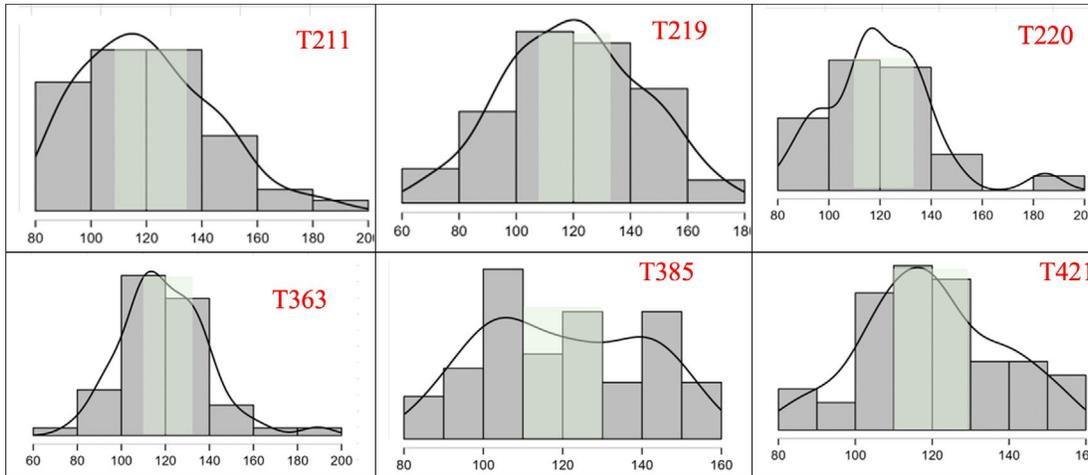
Figure 1: IESOL SELT B1: Item difficulty distributions (LID scale range: 91-110)



With the B1 test forms, there is a range of distributions. T414 is skewed slightly to the easy side; T446 has a comparatively wide distribution; T593 bulges around the midpoint. Nonetheless, in general, the green zones (the LID scale range) in the centre of the item distributions include a substantial number of the items in the B1 test forms. While not uniformly bell-shaped, the frequency trend lines do nonetheless indicate a regularity of shape.

Figure 2 presents the item difficulty distributions for LST B2.

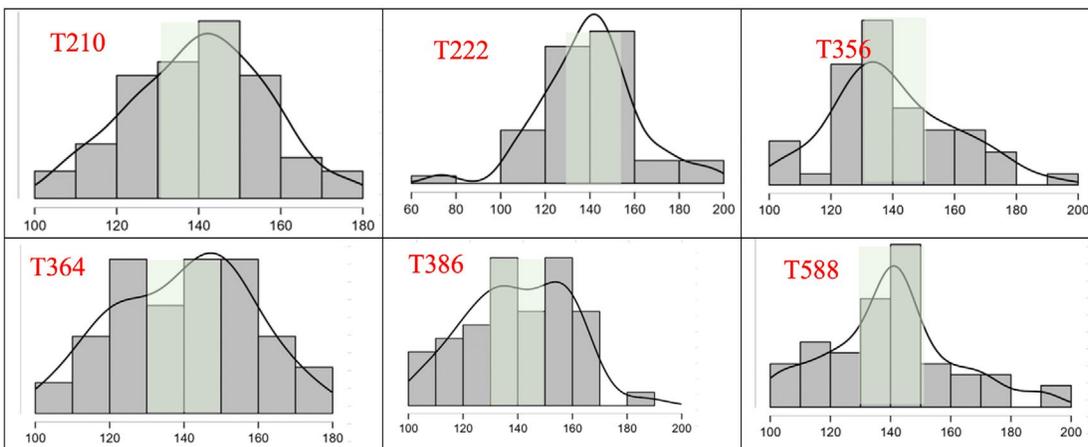
Figure 2: IESOL SELT B2: Item difficulty distributions (LID scale range: 111-130)



With the B2 test forms, distributions again show some divergence in their patterning. T211 is skewed slightly to the easy side; T220 has some outlying difficult items at the top end; T385 has a fairly flat distribution. Nonetheless, in general, the green zones (the supposed LID scale range) in the centre of the item distributions include a substantial number of the items in the B2 test forms. The frequency trend lines indicate a general regularity of shape, however, in general approaching a bell shape.

Figure 3 presents the item difficulty distributions for LST C1.

Figure 3: IESOL SELT C1: Item difficulty distributions (LID scale range: 131-150)



The C1 test form item distributions can be seen to be slightly more regular and bell-shaped than those for B2. T386 and T588 have some outlying difficult items at the top end of the scale, but the LID scale range (the green zones) again occupy a key section of the curve. The frequency trend lines again indicate a regularity of shape, approaching a bell shape.

In summary then, it can be seen that the expert-set items for the LST B1–C1 test forms match well with calibrated LID scale CEFR levels. This lends support to the claim that the LST B1–C1 test forms may be seen to be acceptably anchored on the LID scale.

Conclusion

This chapter has reported on the externally-referenced anchoring (ERA) of LanguageCert SELT tests (LST) at levels B1–C1. The study was pursuing two related research questions.

The first research question explored the extent to which good Rasch infit and outfit statistics would emerge from the externally-referenced anchoring of B1–C1 test forms. As has been described, the majority of B1, B2 and C1 test forms exhibited good Rasch infit and outfit statistics. This may be interpreted as a baseline of test quality.

The second research question explored the extent to which broadly bell-shaped item measure distributions would emerge from the analysis. The analyses generally exhibited a good match between CEFR levels B1–C1 and LID scale levels. Items on all test forms showed generally balanced distributions, with the majority of items in the majority test forms falling within the 25th to 75th percentiles – the percentiles point which broadly match the upper and lower end of the cut scores determined for respective B1–C1 CEFR levels.

The match in the current study between the externally-referenced LST B1–C1 anchored levels and LID scale CEFR B1–C1 levels supports the argument that LanguageCert LST B1–C1 tests have been well set, with the results of the study statistically verifying expert judgements. The fact that the majority of items on the B1–C1 test forms fell within the 25th to 75th percentiles confirms the claim that LST B1–C1 tests are well targeted at the appropriate CEFR levels.

The test forms and items have been shown to be located acceptably on the LID scale – and against CEFR levels. Against this backdrop, vertical anchoring can now be brought to bear to place composite tests for each CEFR level on to the LID and hence LanguageCert Global scales. The current research has explored ERA in the context of a limited number of tests at three CEFR levels. A subsequent study is currently underway which will properly road-test the ERA methodology by submitting a considerable number (i.e., at least 10) test forms at each CEFR level to analysis and scrutiny.

Notes

1. The LanguageCert System reports scores on the LanguageCert Global Scale of 0-100 that is derived directly from the 180-point LID scale. It provides candidates, employers, education institutions and government agencies an easy-to-understand results system. It applies across all the tests in the LanguageCert System. The Global Scale defines specific levels of attainment needed to fulfil certain requirements. For example, entrance into a university or for migration and employment purposes. The levels of attainment can relate to overall performance in an examination, performance by skill (e.g., speaking), or both these parameters.

References

- 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., & Palmer, A.S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- 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.
- Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021a). *Validating the LanguageCert Test of English scale: The paper-based tests*. London, UK: LanguageCert.
- Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021b). *Validating the LanguageCert Test of English scale: The adaptive test*. London, UK: LanguageCert.
- Coniam, D., Zhao, W., Lee, T., Milanovic, M., & Pike, N. (2022). The role of expert judgement in language test validation. *Language Education & Assessment*.
- Ebel, R. L. (1965). *Measuring educational achievement*. Prentice-Hall, NJ: Englewood Cliffs.
- 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.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Humphry, S. (2006). *The impact of differential discrimination on vertical equating*. ARC report.
- Lee, T., Milanovic, M., & Pike, N. (2022). Equating Rasch values and expert judgement through externally-referenced anchoring. *International Journal of TESOL Studies*, 4(1), 187-202.
- Linacre, J. M. (2018). *Winsteps: Rasch measurement computer program*. Winsteps.com: Beaverton, OR.
- Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, 10(3), 211–234.
- Papargyris, Y., & Yan, Z. (2022). Examiner quality and consistency across LanguageCert Writing Tests. *International Journal of TESOL Studies*, 4(1), 203-212.
- 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.
- Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, 16(4), 33-45. <https://doi.org/10.1111/j.1745-3992.1997.tb00606.x>.
- 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.

Appendix 1: LST B1: Fit Statistics and Person Reliabilities

Test no.	Rasch statistics summary		
T206	SELT B1 T206 PERSON 10810 INPUT 1314 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 44.0 51.7 154.41 13.89 .99 .1 .95 .1 P.SD 8.9 1.9 33.64 8.86 .14 .6 .57 .8 REAL RMSE 16.47 TRUE SD 29.33 SEPARATION 1.78 PERSON RELIABILITY .76 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 1112.2 1307.6 100.00 2.08 .99 .0 .94 -.4 P.SD 123.7 5.6 20.52 1.53 .14 2.8 .43 3.1 REAL RMSE 2.15 TRUE SD 20.41 SEPARATION 9.51 ITEM RELIABILITY .99		
	T207	SELT B1 T207 PERSON 10810 INPUT 1295 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 40.8 52.0 141.01 11.14 1.00 .1 .97 .1 P.SD 10.5 .0 33.05 7.00 .11 .7 .41 .8 REAL RMSE 13.16 TRUE SD 30.31 SEPARATION 2.30 PERSON RELIABILITY .84 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 1014.8 1295.0 100.00 1.79 .99 -.1 .97 -.3 P.SD 140.8 .0 19.76 .40 .14 3.4 .39 3.4 REAL RMSE 1.83 TRUE SD 19.68 SEPARATION 10.74 ITEM RELIABILITY .99	
		T208	SELT B1 T208 PERSON 10810 INPUT 1384 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 41.0 51.7 141.70 10.99 1.00 .1 .96 .1 P.SD 10.2 2.0 31.95 6.47 .11 .6 .48 .8 REAL RMSE 12.75 TRUE SD 29.29 SEPARATION 2.30 PERSON RELIABILITY .84 ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 1091.4 1375.1 100.00 1.76 .99 -.2 .96 -.4 P.SD 152.6 8.5 19.95 .36 .15 3.3 .37 3.4 REAL RMSE 1.79 TRUE SD 19.86 SEPARATION 11.00 ITEM RELIABILITY .99
			T209
T384			
	T409		
		T414	
			T446
T593			

Appendix 2: LST B2: Fit Statistics and Person Reliabilities

Test no.	Rasch statistics summary																																																																																										
T211	SELT B2 211 ----- <table border="1"> <thead> <tr> <th>PERSON</th> <th>2732</th> <th>INPUT</th> <th>528</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>31.0</td> <td>51.9</td> <td>132.15</td> <td>7.56</td> <td>1.00</td> <td>.0</td> <td>1.00</td> <td>.0</td> </tr> <tr> <td>P.SD</td> <td>10.2</td> <td>.7</td> <td>25.77</td> <td>2.64</td> <td>.15</td> <td>.9</td> <td>.41</td> <td>1.0</td> </tr> <tr> <td>REAL RMSE</td> <td>8.00</td> <td>TRUE SD</td> <td>24.50</td> <td>SEPARATION</td> <td>3.06</td> <td>PERSON RELIABILITY</td> <td>.90</td> <td></td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>ITEM</th> <th>52</th> <th>INPUT</th> <th>52</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>314.5</td> <td>527.1</td> <td>120.00</td> <td>2.26</td> <td>1.00</td> <td>-.1</td> <td>1.00</td> <td>.0</td> </tr> <tr> <td>P.SD</td> <td>101.5</td> <td>1.2</td> <td>22.91</td> <td>.25</td> <td>.14</td> <td>2.8</td> <td>.28</td> <td>2.6</td> </tr> <tr> <td>REAL RMSE</td> <td>2.27</td> <td>TRUE SD</td> <td>22.79</td> <td>SEPARATION</td> <td>10.05</td> <td>ITEM RELIABILITY</td> <td>.99</td> <td></td> </tr> </tbody> </table>	PERSON	2732	INPUT	528	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	31.0	51.9	132.15	7.56	1.00	.0	1.00	.0	P.SD	10.2	.7	25.77	2.64	.15	.9	.41	1.0	REAL RMSE	8.00	TRUE SD	24.50	SEPARATION	3.06	PERSON RELIABILITY	.90		ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	314.5	527.1	120.00	2.26	1.00	-.1	1.00	.0	P.SD	101.5	1.2	22.91	.25	.14	2.8	.28	2.6	REAL RMSE	2.27	TRUE SD	22.79	SEPARATION	10.05	ITEM RELIABILITY	.99	
	PERSON	2732	INPUT	528	MEASURED		INFIT	OUTFIT																																																																																			
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	31.0	51.9	132.15	7.56	1.00	.0	1.00	.0																																																																																			
P.SD	10.2	.7	25.77	2.64	.15	.9	.41	1.0																																																																																			
REAL RMSE	8.00	TRUE SD	24.50	SEPARATION	3.06	PERSON RELIABILITY	.90																																																																																				
ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT																																																																																				
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	314.5	527.1	120.00	2.26	1.00	-.1	1.00	.0																																																																																			
P.SD	101.5	1.2	22.91	.25	.14	2.8	.28	2.6																																																																																			
REAL RMSE	2.27	TRUE SD	22.79	SEPARATION	10.05	ITEM RELIABILITY	.99																																																																																				
T219	SELT B2 219 ----- <table border="1"> <thead> <tr> <th>PERSON</th> <th>2732</th> <th>INPUT</th> <th>569</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>31.7</td> <td>51.8</td> <td>135.44</td> <td>8.07</td> <td>1.00</td> <td>.0</td> <td>1.00</td> <td>.0</td> </tr> <tr> <td>P.SD</td> <td>11.4</td> <td>1.5</td> <td>29.79</td> <td>3.63</td> <td>.15</td> <td>.9</td> <td>.40</td> <td>.9</td> </tr> <tr> <td>REAL RMSE</td> <td>8.05</td> <td>TRUE SD</td> <td>28.45</td> <td>SEPARATION</td> <td>3.21</td> <td>PERSON RELIABILITY</td> <td>.91</td> <td></td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>ITEM</th> <th>52</th> <th>INPUT</th> <th>52</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>347.2</td> <td>567.0</td> <td>120.00</td> <td>2.25</td> <td>.99</td> <td>-.1</td> <td>1.00</td> <td>-.1</td> </tr> <tr> <td>P.SD</td> <td>103.4</td> <td>2.1</td> <td>23.37</td> <td>.30</td> <td>.15</td> <td>2.8</td> <td>.38</td> <td>2.8</td> </tr> <tr> <td>REAL RMSE</td> <td>2.27</td> <td>TRUE SD</td> <td>23.26</td> <td>SEPARATION</td> <td>10.24</td> <td>ITEM RELIABILITY</td> <td>.99</td> <td></td> </tr> </tbody> </table>	PERSON	2732	INPUT	569	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	31.7	51.8	135.44	8.07	1.00	.0	1.00	.0	P.SD	11.4	1.5	29.79	3.63	.15	.9	.40	.9	REAL RMSE	8.05	TRUE SD	28.45	SEPARATION	3.21	PERSON RELIABILITY	.91		ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	347.2	567.0	120.00	2.25	.99	-.1	1.00	-.1	P.SD	103.4	2.1	23.37	.30	.15	2.8	.38	2.8	REAL RMSE	2.27	TRUE SD	23.26	SEPARATION	10.24	ITEM RELIABILITY	.99	
	PERSON	2732	INPUT	569	MEASURED		INFIT	OUTFIT																																																																																			
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	31.7	51.8	135.44	8.07	1.00	.0	1.00	.0																																																																																			
P.SD	11.4	1.5	29.79	3.63	.15	.9	.40	.9																																																																																			
REAL RMSE	8.05	TRUE SD	28.45	SEPARATION	3.21	PERSON RELIABILITY	.91																																																																																				
ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT																																																																																				
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	347.2	567.0	120.00	2.25	.99	-.1	1.00	-.1																																																																																			
P.SD	103.4	2.1	23.37	.30	.15	2.8	.38	2.8																																																																																			
REAL RMSE	2.27	TRUE SD	23.26	SEPARATION	10.24	ITEM RELIABILITY	.99																																																																																				
T220	SELT B2 220 ----- <table border="1"> <thead> <tr> <th>PERSON</th> <th>2732</th> <th>INPUT</th> <th>547</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>33.5</td> <td>51.8</td> <td>139.19</td> <td>8.07</td> <td>1.00</td> <td>.1</td> <td>.98</td> <td>.0</td> </tr> <tr> <td>P.SD</td> <td>11.1</td> <td>2.2</td> <td>28.30</td> <td>3.41</td> <td>.14</td> <td>.8</td> <td>.32</td> <td>.9</td> </tr> <tr> <td>REAL RMSE</td> <td>8.76</td> <td>TRUE SD</td> <td>26.91</td> <td>SEPARATION</td> <td>3.07</td> <td>PERSON RELIABILITY</td> <td>.90</td> <td></td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>ITEM</th> <th>52</th> <th>INPUT</th> <th>52</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>352.2</td> <td>544.8</td> <td>120.00</td> <td>2.27</td> <td>1.00</td> <td>-.1</td> <td>.98</td> <td>-.1</td> </tr> <tr> <td>P.SD</td> <td>87.5</td> <td>1.8</td> <td>20.71</td> <td>.26</td> <td>.14</td> <td>2.7</td> <td>.28</td> <td>2.6</td> </tr> <tr> <td>REAL RMSE</td> <td>2.28</td> <td>TRUE SD</td> <td>20.58</td> <td>SEPARATION</td> <td>9.02</td> <td>ITEM RELIABILITY</td> <td>.99</td> <td></td> </tr> </tbody> </table>	PERSON	2732	INPUT	547	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	33.5	51.8	139.19	8.07	1.00	.1	.98	.0	P.SD	11.1	2.2	28.30	3.41	.14	.8	.32	.9	REAL RMSE	8.76	TRUE SD	26.91	SEPARATION	3.07	PERSON RELIABILITY	.90		ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	352.2	544.8	120.00	2.27	1.00	-.1	.98	-.1	P.SD	87.5	1.8	20.71	.26	.14	2.7	.28	2.6	REAL RMSE	2.28	TRUE SD	20.58	SEPARATION	9.02	ITEM RELIABILITY	.99	
	PERSON	2732	INPUT	547	MEASURED		INFIT	OUTFIT																																																																																			
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	33.5	51.8	139.19	8.07	1.00	.1	.98	.0																																																																																			
P.SD	11.1	2.2	28.30	3.41	.14	.8	.32	.9																																																																																			
REAL RMSE	8.76	TRUE SD	26.91	SEPARATION	3.07	PERSON RELIABILITY	.90																																																																																				
ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT																																																																																				
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	352.2	544.8	120.00	2.27	1.00	-.1	.98	-.1																																																																																			
P.SD	87.5	1.8	20.71	.26	.14	2.7	.28	2.6																																																																																			
REAL RMSE	2.28	TRUE SD	20.58	SEPARATION	9.02	ITEM RELIABILITY	.99																																																																																				
T363	SELT B2 363 ----- <table border="1"> <thead> <tr> <th>PERSON</th> <th>2732</th> <th>INPUT</th> <th>573</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>37.7</td> <td>51.8</td> <td>149.97</td> <td>9.42</td> <td>1.00</td> <td>.1</td> <td>.97</td> <td>.0</td> </tr> <tr> <td>P.SD</td> <td>10.6</td> <td>1.9</td> <td>30.75</td> <td>5.38</td> <td>.15</td> <td>.7</td> <td>.36</td> <td>.8</td> </tr> <tr> <td>REAL RMSE</td> <td>10.05</td> <td>TRUE SD</td> <td>28.78</td> <td>SEPARATION</td> <td>2.65</td> <td>PERSON RELIABILITY</td> <td>.88</td> <td></td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>ITEM</th> <th>52</th> <th>INPUT</th> <th>52</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>415.4</td> <td>571.1</td> <td>120.00</td> <td>2.40</td> <td>.99</td> <td>-.1</td> <td>.96</td> <td>-.2</td> </tr> <tr> <td>P.SD</td> <td>80.1</td> <td>1.7</td> <td>19.74</td> <td>.34</td> <td>.15</td> <td>2.7</td> <td>.29</td> <td>2.4</td> </tr> <tr> <td>REAL RMSE</td> <td>2.42</td> <td>TRUE SD</td> <td>19.60</td> <td>SEPARATION</td> <td>8.00</td> <td>ITEM RELIABILITY</td> <td>.98</td> <td></td> </tr> </tbody> </table>	PERSON	2732	INPUT	573	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	37.7	51.8	149.97	9.42	1.00	.1	.97	.0	P.SD	10.6	1.9	30.75	5.38	.15	.7	.36	.8	REAL RMSE	10.05	TRUE SD	28.78	SEPARATION	2.65	PERSON RELIABILITY	.88		ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	415.4	571.1	120.00	2.40	.99	-.1	.96	-.2	P.SD	80.1	1.7	19.74	.34	.15	2.7	.29	2.4	REAL RMSE	2.42	TRUE SD	19.60	SEPARATION	8.00	ITEM RELIABILITY	.98	
	PERSON	2732	INPUT	573	MEASURED		INFIT	OUTFIT																																																																																			
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	37.7	51.8	149.97	9.42	1.00	.1	.97	.0																																																																																			
P.SD	10.6	1.9	30.75	5.38	.15	.7	.36	.8																																																																																			
REAL RMSE	10.05	TRUE SD	28.78	SEPARATION	2.65	PERSON RELIABILITY	.88																																																																																				
ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT																																																																																				
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	415.4	571.1	120.00	2.40	.99	-.1	.96	-.2																																																																																			
P.SD	80.1	1.7	19.74	.34	.15	2.7	.29	2.4																																																																																			
REAL RMSE	2.42	TRUE SD	19.60	SEPARATION	8.00	ITEM RELIABILITY	.98																																																																																				
T385	SELT B2 385 ----- <table border="1"> <thead> <tr> <th>PERSON</th> <th>2732</th> <th>INPUT</th> <th>280</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>35.4</td> <td>51.9</td> <td>144.96</td> <td>8.86</td> <td>1.00</td> <td>.1</td> <td>1.00</td> <td>.1</td> </tr> <tr> <td>P.SD</td> <td>10.9</td> <td>.9</td> <td>31.00</td> <td>4.99</td> <td>.12</td> <td>.8</td> <td>.44</td> <td>1.0</td> </tr> <tr> <td>REAL RMSE</td> <td>10.17</td> <td>TRUE SD</td> <td>29.29</td> <td>SEPARATION</td> <td>2.88</td> <td>PERSON RELIABILITY</td> <td>.89</td> <td></td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>ITEM</th> <th>52</th> <th>INPUT</th> <th>52</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>190.8</td> <td>279.5</td> <td>120.00</td> <td>3.29</td> <td>1.00</td> <td>-.1</td> <td>1.00</td> <td>.0</td> </tr> <tr> <td>P.SD</td> <td>42.6</td> <td>.8</td> <td>20.02</td> <td>.42</td> <td>.15</td> <td>2.1</td> <td>.39</td> <td>2.1</td> </tr> <tr> <td>REAL RMSE</td> <td>3.32</td> <td>TRUE SD</td> <td>19.74</td> <td>SEPARATION</td> <td>5.95</td> <td>ITEM RELIABILITY</td> <td>.97</td> <td></td> </tr> </tbody> </table>	PERSON	2732	INPUT	280	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	35.4	51.9	144.96	8.86	1.00	.1	1.00	.1	P.SD	10.9	.9	31.00	4.99	.12	.8	.44	1.0	REAL RMSE	10.17	TRUE SD	29.29	SEPARATION	2.88	PERSON RELIABILITY	.89		ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	190.8	279.5	120.00	3.29	1.00	-.1	1.00	.0	P.SD	42.6	.8	20.02	.42	.15	2.1	.39	2.1	REAL RMSE	3.32	TRUE SD	19.74	SEPARATION	5.95	ITEM RELIABILITY	.97	
	PERSON	2732	INPUT	280	MEASURED		INFIT	OUTFIT																																																																																			
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	35.4	51.9	144.96	8.86	1.00	.1	1.00	.1																																																																																			
P.SD	10.9	.9	31.00	4.99	.12	.8	.44	1.0																																																																																			
REAL RMSE	10.17	TRUE SD	29.29	SEPARATION	2.88	PERSON RELIABILITY	.89																																																																																				
ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT																																																																																				
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	190.8	279.5	120.00	3.29	1.00	-.1	1.00	.0																																																																																			
P.SD	42.6	.8	20.02	.42	.15	2.1	.39	2.1																																																																																			
REAL RMSE	3.32	TRUE SD	19.74	SEPARATION	5.95	ITEM RELIABILITY	.97																																																																																				
T421	SELT B2 421 ----- <table border="1"> <thead> <tr> <th>PERSON</th> <th>2732</th> <th>INPUT</th> <th>235</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>28.9</td> <td>51.7</td> <td>126.60</td> <td>7.12</td> <td>1.00</td> <td>.0</td> <td>1.01</td> <td>.0</td> </tr> <tr> <td>P.SD</td> <td>10.2</td> <td>3.2</td> <td>23.14</td> <td>3.13</td> <td>.10</td> <td>.8</td> <td>.23</td> <td>.9</td> </tr> <tr> <td>REAL RMSE</td> <td>7.78</td> <td>TRUE SD</td> <td>21.79</td> <td>SEPARATION</td> <td>2.80</td> <td>PERSON RELIABILITY</td> <td>.89</td> <td></td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>ITEM</th> <th>52</th> <th>INPUT</th> <th>52</th> <th>MEASURED</th> <th></th> <th>INFIT</th> <th>OUTFIT</th> <th></th> </tr> <tr> <th>TOTAL</th> <th>COUNT</th> <th>MEASURE</th> <th>REALSE</th> <th>INMSQ</th> <th>ZSTD</th> <th>OHMSQ</th> <th>ZSTD</th> <th></th> </tr> </thead> <tbody> <tr> <td>MEAN</td> <td>130.5</td> <td>233.4</td> <td>120.00</td> <td>3.17</td> <td>1.00</td> <td>-.1</td> <td>1.01</td> <td>.0</td> </tr> <tr> <td>P.SD</td> <td>37.7</td> <td>1.1</td> <td>17.36</td> <td>.26</td> <td>.14</td> <td>2.3</td> <td>.22</td> <td>2.1</td> </tr> <tr> <td>REAL RMSE</td> <td>3.18</td> <td>TRUE SD</td> <td>17.06</td> <td>SEPARATION</td> <td>5.36</td> <td>ITEM RELIABILITY</td> <td>.97</td> <td></td> </tr> </tbody> </table>	PERSON	2732	INPUT	235	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	28.9	51.7	126.60	7.12	1.00	.0	1.01	.0	P.SD	10.2	3.2	23.14	3.13	.10	.8	.23	.9	REAL RMSE	7.78	TRUE SD	21.79	SEPARATION	2.80	PERSON RELIABILITY	.89		ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT		TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD		MEAN	130.5	233.4	120.00	3.17	1.00	-.1	1.01	.0	P.SD	37.7	1.1	17.36	.26	.14	2.3	.22	2.1	REAL RMSE	3.18	TRUE SD	17.06	SEPARATION	5.36	ITEM RELIABILITY	.97	
	PERSON	2732	INPUT	235	MEASURED		INFIT	OUTFIT																																																																																			
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	28.9	51.7	126.60	7.12	1.00	.0	1.01	.0																																																																																			
P.SD	10.2	3.2	23.14	3.13	.10	.8	.23	.9																																																																																			
REAL RMSE	7.78	TRUE SD	21.79	SEPARATION	2.80	PERSON RELIABILITY	.89																																																																																				
ITEM	52	INPUT	52	MEASURED		INFIT	OUTFIT																																																																																				
TOTAL	COUNT	MEASURE	REALSE	INMSQ	ZSTD	OHMSQ	ZSTD																																																																																				
MEAN	130.5	233.4	120.00	3.17	1.00	-.1	1.01	.0																																																																																			
P.SD	37.7	1.1	17.36	.26	.14	2.3	.22	2.1																																																																																			
REAL RMSE	3.18	TRUE SD	17.06	SEPARATION	5.36	ITEM RELIABILITY	.97																																																																																				

Appendix 3: LST C1: Fit Statistics and Person Reliabilities

Test no.	Rasch statistics summary																																																																																																				
T210	SELT C1 T210 ----- <table border="0"> <tr> <td>PERSON</td> <td>581</td> <td>INPUT</td> <td>135</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>30.6</td> <td>51.9</td> <td>150.01</td> <td>6.85</td> <td>1.00</td> <td>.0</td> <td>1.00</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>10.2</td> <td>.6</td> <td>21.06</td> <td>1.08</td> <td>.11</td> <td>.9</td> <td>.21</td> <td>.9</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>6.94</td> <td>TRUE SD</td> <td>19.88</td> <td>SEPARATION</td> <td>2.87</td> <td>PERSON RELIABILITY</td> <td>.89</td> <td></td> <td></td> </tr> </table> <table border="0"> <tr> <td>ITEM</td> <td>52</td> <td>INPUT</td> <td>52</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>79.6</td> <td>134.7</td> <td>140.00</td> <td>4.22</td> <td>1.00</td> <td>-.1</td> <td>1.00</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>19.8</td> <td>.5</td> <td>16.10</td> <td>.34</td> <td>.17</td> <td>2.0</td> <td>.25</td> <td>1.9</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>4.23</td> <td>TRUE SD</td> <td>15.53</td> <td>SEPARATION</td> <td>3.67</td> <td>ITEM RELIABILITY</td> <td>.93</td> <td></td> <td></td> </tr> </table>	PERSON	581	INPUT	135	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	30.6	51.9	150.01	6.85	1.00	.0	1.00	.0		P.SD	10.2	.6	21.06	1.08	.11	.9	.21	.9		REAL RMSE	6.94	TRUE SD	19.88	SEPARATION	2.87	PERSON RELIABILITY	.89			ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	79.6	134.7	140.00	4.22	1.00	-.1	1.00	.0		P.SD	19.8	.5	16.10	.34	.17	2.0	.25	1.9		REAL RMSE	4.23	TRUE SD	15.53	SEPARATION	3.67	ITEM RELIABILITY	.93		
	PERSON	581	INPUT	135	MEASURED		INFIT		OUTFIT																																																																																												
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	30.6	51.9	150.01	6.85	1.00	.0	1.00	.0																																																																																													
P.SD	10.2	.6	21.06	1.08	.11	.9	.21	.9																																																																																													
REAL RMSE	6.94	TRUE SD	19.88	SEPARATION	2.87	PERSON RELIABILITY	.89																																																																																														
ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT																																																																																													
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	79.6	134.7	140.00	4.22	1.00	-.1	1.00	.0																																																																																													
P.SD	19.8	.5	16.10	.34	.17	2.0	.25	1.9																																																																																													
REAL RMSE	4.23	TRUE SD	15.53	SEPARATION	3.67	ITEM RELIABILITY	.93																																																																																														
T222	SELT C1 T222 ----- <table border="0"> <tr> <td>PERSON</td> <td>581</td> <td>INPUT</td> <td>100</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>33.4</td> <td>52.0</td> <td>158.73</td> <td>8.18</td> <td>1.01</td> <td>.0</td> <td>1.01</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>11.2</td> <td>.0</td> <td>29.10</td> <td>3.82</td> <td>-.14</td> <td>.8</td> <td>.31</td> <td>.8</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>9.03</td> <td>TRUE SD</td> <td>27.66</td> <td>SEPARATION</td> <td>3.06</td> <td>PERSON RELIABILITY</td> <td>.90</td> <td></td> <td></td> </tr> </table> <table border="0"> <tr> <td>ITEM</td> <td>52</td> <td>INPUT</td> <td>52</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>64.2</td> <td>100.0</td> <td>140.00</td> <td>5.40</td> <td>.99</td> <td>.0</td> <td>1.01</td> <td>-.1</td> <td></td> </tr> <tr> <td>P.SD</td> <td>16.3</td> <td>.0</td> <td>21.76</td> <td>1.06</td> <td>-.15</td> <td>1.3</td> <td>.39</td> <td>1.3</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>5.50</td> <td>TRUE SD</td> <td>21.05</td> <td>SEPARATION</td> <td>3.83</td> <td>ITEM RELIABILITY</td> <td>.94</td> <td></td> <td></td> </tr> </table>	PERSON	581	INPUT	100	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	33.4	52.0	158.73	8.18	1.01	.0	1.01	.0		P.SD	11.2	.0	29.10	3.82	-.14	.8	.31	.8		REAL RMSE	9.03	TRUE SD	27.66	SEPARATION	3.06	PERSON RELIABILITY	.90			ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	64.2	100.0	140.00	5.40	.99	.0	1.01	-.1		P.SD	16.3	.0	21.76	1.06	-.15	1.3	.39	1.3		REAL RMSE	5.50	TRUE SD	21.05	SEPARATION	3.83	ITEM RELIABILITY	.94		
	PERSON	581	INPUT	100	MEASURED		INFIT		OUTFIT																																																																																												
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	33.4	52.0	158.73	8.18	1.01	.0	1.01	.0																																																																																													
P.SD	11.2	.0	29.10	3.82	-.14	.8	.31	.8																																																																																													
REAL RMSE	9.03	TRUE SD	27.66	SEPARATION	3.06	PERSON RELIABILITY	.90																																																																																														
ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT																																																																																													
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	64.2	100.0	140.00	5.40	.99	.0	1.01	-.1																																																																																													
P.SD	16.3	.0	21.76	1.06	-.15	1.3	.39	1.3																																																																																													
REAL RMSE	5.50	TRUE SD	21.05	SEPARATION	3.83	ITEM RELIABILITY	.94																																																																																														
T356	SELT C1 T356 ----- <table border="0"> <tr> <td>PERSON</td> <td>581</td> <td>INPUT</td> <td>115</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>36.1</td> <td>52.0</td> <td>163.88</td> <td>8.06</td> <td>1.00</td> <td>-.1</td> <td>.99</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>9.3</td> <td>.0</td> <td>24.82</td> <td>3.49</td> <td>.12</td> <td>.8</td> <td>.28</td> <td>.8</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>8.78</td> <td>TRUE SD</td> <td>23.22</td> <td>SEPARATION</td> <td>2.65</td> <td>PERSON RELIABILITY</td> <td>.87</td> <td></td> <td></td> </tr> </table> <table border="0"> <tr> <td>ITEM</td> <td>52</td> <td>INPUT</td> <td>52</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>79.9</td> <td>115.0</td> <td>140.00</td> <td>5.83</td> <td>1.00</td> <td>.0</td> <td>.99</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>18.1</td> <td>.0</td> <td>19.40</td> <td>.74</td> <td>.15</td> <td>1.4</td> <td>.33</td> <td>1.4</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>5.08</td> <td>TRUE SD</td> <td>18.73</td> <td>SEPARATION</td> <td>3.68</td> <td>ITEM RELIABILITY</td> <td>.93</td> <td></td> <td></td> </tr> </table>	PERSON	581	INPUT	115	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	36.1	52.0	163.88	8.06	1.00	-.1	.99	.0		P.SD	9.3	.0	24.82	3.49	.12	.8	.28	.8		REAL RMSE	8.78	TRUE SD	23.22	SEPARATION	2.65	PERSON RELIABILITY	.87			ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	79.9	115.0	140.00	5.83	1.00	.0	.99	.0		P.SD	18.1	.0	19.40	.74	.15	1.4	.33	1.4		REAL RMSE	5.08	TRUE SD	18.73	SEPARATION	3.68	ITEM RELIABILITY	.93		
	PERSON	581	INPUT	115	MEASURED		INFIT		OUTFIT																																																																																												
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	36.1	52.0	163.88	8.06	1.00	-.1	.99	.0																																																																																													
P.SD	9.3	.0	24.82	3.49	.12	.8	.28	.8																																																																																													
REAL RMSE	8.78	TRUE SD	23.22	SEPARATION	2.65	PERSON RELIABILITY	.87																																																																																														
ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT																																																																																													
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	79.9	115.0	140.00	5.83	1.00	.0	.99	.0																																																																																													
P.SD	18.1	.0	19.40	.74	.15	1.4	.33	1.4																																																																																													
REAL RMSE	5.08	TRUE SD	18.73	SEPARATION	3.68	ITEM RELIABILITY	.93																																																																																														
T364	SELT C1 T364 ----- <table border="0"> <tr> <td>PERSON</td> <td>581</td> <td>INPUT</td> <td>120</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>32.2</td> <td>52.0</td> <td>155.87</td> <td>7.54</td> <td>1.00</td> <td>.0</td> <td>1.00</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>10.9</td> <td>.0</td> <td>26.02</td> <td>2.59</td> <td>.11</td> <td>.8</td> <td>.55</td> <td>1.0</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>7.97</td> <td>TRUE SD</td> <td>24.76</td> <td>SEPARATION</td> <td>3.11</td> <td>PERSON RELIABILITY</td> <td>.91</td> <td></td> <td></td> </tr> </table> <table border="0"> <tr> <td>ITEM</td> <td>52</td> <td>INPUT</td> <td>52</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>74.3</td> <td>120.0</td> <td>140.00</td> <td>4.66</td> <td>.99</td> <td>-.1</td> <td>1.00</td> <td>-.1</td> <td></td> </tr> <tr> <td>P.SD</td> <td>18.4</td> <td>.0</td> <td>18.18</td> <td>.48</td> <td>.15</td> <td>1.5</td> <td>.54</td> <td>1.5</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>4.68</td> <td>TRUE SD</td> <td>17.56</td> <td>SEPARATION</td> <td>3.75</td> <td>ITEM RELIABILITY</td> <td>.93</td> <td></td> <td></td> </tr> </table>	PERSON	581	INPUT	120	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	32.2	52.0	155.87	7.54	1.00	.0	1.00	.0		P.SD	10.9	.0	26.02	2.59	.11	.8	.55	1.0		REAL RMSE	7.97	TRUE SD	24.76	SEPARATION	3.11	PERSON RELIABILITY	.91			ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	74.3	120.0	140.00	4.66	.99	-.1	1.00	-.1		P.SD	18.4	.0	18.18	.48	.15	1.5	.54	1.5		REAL RMSE	4.68	TRUE SD	17.56	SEPARATION	3.75	ITEM RELIABILITY	.93		
	PERSON	581	INPUT	120	MEASURED		INFIT		OUTFIT																																																																																												
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	32.2	52.0	155.87	7.54	1.00	.0	1.00	.0																																																																																													
P.SD	10.9	.0	26.02	2.59	.11	.8	.55	1.0																																																																																													
REAL RMSE	7.97	TRUE SD	24.76	SEPARATION	3.11	PERSON RELIABILITY	.91																																																																																														
ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT																																																																																													
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	74.3	120.0	140.00	4.66	.99	-.1	1.00	-.1																																																																																													
P.SD	18.4	.0	18.18	.48	.15	1.5	.54	1.5																																																																																													
REAL RMSE	4.68	TRUE SD	17.56	SEPARATION	3.75	ITEM RELIABILITY	.93																																																																																														
T386	SELT C1 T386 ----- <table border="0"> <tr> <td>PERSON</td> <td>581</td> <td>INPUT</td> <td>55</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>29.2</td> <td>51.9</td> <td>147.57</td> <td>6.90</td> <td>1.00</td> <td>.0</td> <td>1.01</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>9.7</td> <td>.8</td> <td>21.35</td> <td>1.39</td> <td>.12</td> <td>.8</td> <td>.26</td> <td>.9</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>7.04</td> <td>TRUE SD</td> <td>20.15</td> <td>SEPARATION</td> <td>2.86</td> <td>PERSON RELIABILITY</td> <td>.89</td> <td></td> <td></td> </tr> </table> <table border="0"> <tr> <td>ITEM</td> <td>52</td> <td>INPUT</td> <td>52</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>30.9</td> <td>54.9</td> <td>140.00</td> <td>6.67</td> <td>.99</td> <td>-.1</td> <td>1.01</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>9.4</td> <td>.3</td> <td>18.59</td> <td>.72</td> <td>.22</td> <td>1.5</td> <td>.40</td> <td>1.5</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>6.71</td> <td>TRUE SD</td> <td>17.34</td> <td>SEPARATION</td> <td>2.58</td> <td>ITEM RELIABILITY</td> <td>.87</td> <td></td> <td></td> </tr> </table>	PERSON	581	INPUT	55	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	29.2	51.9	147.57	6.90	1.00	.0	1.01	.0		P.SD	9.7	.8	21.35	1.39	.12	.8	.26	.9		REAL RMSE	7.04	TRUE SD	20.15	SEPARATION	2.86	PERSON RELIABILITY	.89			ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	30.9	54.9	140.00	6.67	.99	-.1	1.01	.0		P.SD	9.4	.3	18.59	.72	.22	1.5	.40	1.5		REAL RMSE	6.71	TRUE SD	17.34	SEPARATION	2.58	ITEM RELIABILITY	.87		
	PERSON	581	INPUT	55	MEASURED		INFIT		OUTFIT																																																																																												
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	29.2	51.9	147.57	6.90	1.00	.0	1.01	.0																																																																																													
P.SD	9.7	.8	21.35	1.39	.12	.8	.26	.9																																																																																													
REAL RMSE	7.04	TRUE SD	20.15	SEPARATION	2.86	PERSON RELIABILITY	.89																																																																																														
ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT																																																																																													
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	30.9	54.9	140.00	6.67	.99	-.1	1.01	.0																																																																																													
P.SD	9.4	.3	18.59	.72	.22	1.5	.40	1.5																																																																																													
REAL RMSE	6.71	TRUE SD	17.34	SEPARATION	2.58	ITEM RELIABILITY	.87																																																																																														
T588	SELT C1 T588 ----- <table border="0"> <tr> <td>PERSON</td> <td>581</td> <td>INPUT</td> <td>56</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>30.6</td> <td>52.0</td> <td>150.40</td> <td>7.02</td> <td>.99</td> <td>.0</td> <td>1.03</td> <td>-.1</td> <td></td> </tr> <tr> <td>P.SD</td> <td>9.7</td> <td>.2</td> <td>21.46</td> <td>1.12</td> <td>.12</td> <td>.8</td> <td>.31</td> <td>1.0</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>7.11</td> <td>TRUE SD</td> <td>20.25</td> <td>SEPARATION</td> <td>2.85</td> <td>PERSON RELIABILITY</td> <td>.89</td> <td></td> <td></td> </tr> </table> <table border="0"> <tr> <td>ITEM</td> <td>52</td> <td>INPUT</td> <td>52</td> <td>MEASURED</td> <td></td> <td>INFIT</td> <td></td> <td>OUTFIT</td> <td></td> </tr> <tr> <td>TOTAL</td> <td>COUNT</td> <td>MEASURE</td> <td>REALSE</td> <td>IMNSQ</td> <td>ZSTD</td> <td>OMNSQ</td> <td>ZSTD</td> <td></td> <td></td> </tr> <tr> <td>MEAN</td> <td>32.9</td> <td>56.0</td> <td>140.00</td> <td>6.77</td> <td>1.00</td> <td>-.1</td> <td>1.03</td> <td>.0</td> <td></td> </tr> <tr> <td>P.SD</td> <td>10.2</td> <td>.2</td> <td>21.08</td> <td>.87</td> <td>-.16</td> <td>1.2</td> <td>.34</td> <td>1.2</td> <td></td> </tr> <tr> <td>REAL RMSE</td> <td>6.82</td> <td>TRUE SD</td> <td>19.95</td> <td>SEPARATION</td> <td>2.92</td> <td>ITEM RELIABILITY</td> <td>.90</td> <td></td> <td></td> </tr> </table>	PERSON	581	INPUT	56	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	30.6	52.0	150.40	7.02	.99	.0	1.03	-.1		P.SD	9.7	.2	21.46	1.12	.12	.8	.31	1.0		REAL RMSE	7.11	TRUE SD	20.25	SEPARATION	2.85	PERSON RELIABILITY	.89			ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT		TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD			MEAN	32.9	56.0	140.00	6.77	1.00	-.1	1.03	.0		P.SD	10.2	.2	21.08	.87	-.16	1.2	.34	1.2		REAL RMSE	6.82	TRUE SD	19.95	SEPARATION	2.92	ITEM RELIABILITY	.90		
	PERSON	581	INPUT	56	MEASURED		INFIT		OUTFIT																																																																																												
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	30.6	52.0	150.40	7.02	.99	.0	1.03	-.1																																																																																													
P.SD	9.7	.2	21.46	1.12	.12	.8	.31	1.0																																																																																													
REAL RMSE	7.11	TRUE SD	20.25	SEPARATION	2.85	PERSON RELIABILITY	.89																																																																																														
ITEM	52	INPUT	52	MEASURED		INFIT		OUTFIT																																																																																													
TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD																																																																																														
MEAN	32.9	56.0	140.00	6.77	1.00	-.1	1.03	.0																																																																																													
P.SD	10.2	.2	21.08	.87	-.16	1.2	.34	1.2																																																																																													
REAL RMSE	6.82	TRUE SD	19.95	SEPARATION	2.92	ITEM RELIABILITY	.90																																																																																														

Chapter 3: Aligning LanguageCert SELT Tests to the LanguageCert Item Difficulty Scale

Tony Lee, Yiannis Papargyris, Michael Milanovic, Nigel Pike and David Coniam

Abstract

This chapter reports on the alignment of LanguageCert SELT tests to the LanguageCert Item Difficulty (LID) Scale. The chapter builds on the study reported in Chapter 2 which established, through the use of externally-referenced anchoring, that the LanguageCert SELT B1–C1 tests are robust.

The chapter explores the alignment of LanguageCert SELT tests in relation to the two objectively marked components of Listening and Reading. The use of externally-referenced anchoring enabled the robustness of the four CEFR test levels B1–C2 to be demonstrated.

As the chapter illustrates, the LanguageCert SELT tests in general assess at their designated CEFR level but also contain items which allow them to assess across levels. At the C1 level, there are items which assess above C1 and, at the other end, below C1. Likewise, at the B2 level, there are items which assess both above and below B2.

The value of being able to assess across levels provides valuable information for stakeholders such as tertiary level administrators, admissions tutors, immigration officials and so on. Stakeholders of the LanguageCert Academic and General test have identified this as a highly valued feature of these tests going forward.

Keywords: test alignment, SELT, IESOL, externally-referenced anchoring

Introduction

Since 2020, LanguageCert has been an approved provider, delivering Secure English Language Tests (SELT) tests to the UK Home Office for UK visas & immigration purposes, for movement to and for work in the UK. In the LanguageCert SELT Test (LST), four-skills tests are offered at a range of levels (B1 to C2), mapped to the Common European Framework of Reference (CEFR). Chapter 2 by Milanovic et al. illustrated how LanguageCert calibrates test material and aligns test forms to the respective CEFR levels. Building on the previous study, the current study demonstrates the alignment of all four LST levels (B1–C2) incorporating all B1 to C2 test forms produced since 2020.

The LST tests used in the current study consist of a number of the test forms for the respective CEFR levels delivered by LanguageCert in the 18-month period from mid-2020 to late 2021.

The LanguageCert SELT Tests

The LanguageCert SELT Test (LST) suite of tests form an integral part of the LanguageCert System. The suite comprises four tests from B1 to C2, each aligned to its respective CEFR level as well as three 2-skill tests ranging from A1-B1. Examination specifications reflect the requirements of the CEFR. Test materials writers conform to the highest international standards and have extensive expertise in, and knowledge and understanding of, the CEFR, the latter being crucial in ensuring validity and reliability (Hughes, 2003). Test items are linked to the CEFR by expert judgement, a methodology which has been shown to be robust (Coniam et al., 2022).

The B1-C1 tests comprise 52 items: 26 Listening and 26 Reading items. The C2 tests comprise 56 items: 30 Listening and 26 Reading items. In adhering to the key test qualities of validity and reliability (Bachman and Palmer, 2010), the LST tests assess the communicative skills that test takers will be expected to have mastered at particular levels of ability. Test content matches target test takers – in terms of grammar, functions, vocabulary, topics etc., and the tasks have correspondingly relevant ‘communicative’ contexts.

Each LST 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 referred to above 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 creating and developing test materials and aligning them to the CEFR. The LID scale may be found in Table 2 below.

Studies by Coniam et al. (2021a; 2021b) have validated and extended the LID scale beyond its original CTS origins to a Rasch-based calibration where all levels are statistically validated and linked. Rasch-based calibration provides greater reliability than classical statistics.

An overview of the methodology surrounding Rasch can be found in the Glossary of statistical terms and techniques at the end of the volume, along with an outline of the infit and outfit mean square statistics which are key to the interpretation of Rasch results in the context of data ‘fit’.

Test Data

Table 1 below provides detail on the number of test forms at each level and candidates.

Table 1: SELT IESOL test forms and candidatures

CEFR level	Test forms	Candidates
B1	9	10,808
B2	6	2,732
C1	6	581
C2	3	111

Via externally-referenced, or vertical, anchoring (see Chapter 2 and further detail below), test forms are anchored at the midpoint of the item distribution of a given scale. The C2 sample is small, as can be seen from Table 1 but as Lee et al. (2022) illustrate, externally-referenced anchoring is nonetheless a methodology that works even with small samples. On this basis, C2 is included in the current analysis.

The midpoints of the LID scale for the six CEFR levels are presented in Table 2. In line with the LanguageCert Global Scale, Table 2 includes correspondences between the LID scale and the Global Scale.

Table 2: LID scale

CEFR level	LID scale range	LID scale midpoint	Global scale range	Global scale midpoint
A1	51-70	60	10-19	15
A2	71-90	80	20-39	30
B1	91-110	100	40-59	50
B2	111-130	120	60-74	67
C1	131-150	140	75-89	82
C2	151-170	160	90-100	95

Externally-Referenced Anchoring

The methodology used in the current study is described more fully in Chapter 2 and is based on externally-referenced anchoring (ERA) (Lee et al., 2022). In ERA, test forms which have no common items but comprise items which have been set at predefined and well-accepted CEFR levels are anchored using the calibrated midpoints of a test form against the LID scale and against the CEFR. For each test level, the frame of reference (see Humphry, 2006) constitutes the respective CEFR scale locations calibrated through the test forms and items for that level. On the basis of vertical midpoint anchoring, ERA:

- enables an effective calibration of the items in each test form – given that no other restrictions are imposed on the items.
- reveals the items' goodness of fit between expertly-assigned values and calibrated item distributions.

The anchoring goodness of fit is then evaluated by two metrics:

1. The extent to which a test's midpoint corresponds to the LID scale level.
2. The fit in terms of the extent to which the item distribution around a test's midpoint includes most of the items in a given test. Such fit is determined by a broadly bell-shaped distribution of item measures with the majority of item measures being clustered around the mean and falling between the 25th to 75th percentiles (Lee et al., 2022).

Research Question

The research question being pursued in the current study is:

Can the four SELT tests (B1-C2) be accurately placed on the LID scale and hence against the CEFR?

Background Statistical Analysis

The reader is referred to the outline of the Rasch measurement model provided in the Glossary of statistical terms and techniques at the end of the volume.

Item Infit and Outfit

Accuracy mentioned in the research question above will be measured through good Rasch infit and outfit statistics emerging from the analysis at each of the four test levels. Analysis in the current study has been conducted via the Rasch analysis software Winsteps (Linacre, 2018). Appendix 1 provides detail on fit statistics. Most of the items in tests at all four LanguageCert SELT Test levels had infit and outfit fit statistics within the acceptable fit range of 0.7-1.3, indicating good fit to the Rasch model.

Reliability

Test reliability, for a 50-item test, is accepted at 0.7 or above (Ebel, 1965). The equivalent of classical test reliability in Rasch is person reliability (Anselmi et al., 2019). As Appendix 1 illustrates, 0.8 or better was achieved on all four levels of test, thus meeting reliability criteria.

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

Externally-referenced Anchoring Results

Test means and measures that emerged after the introduction of externally-referenced anchoring are now examined, in particular means recorded at the 25th, 50th and 75th percentiles. The 25th percentile will ideally be located half a logit (10 LID scale points) below and the 75th percentile half a logit above the test midpoint (Lee et al., 2022).

Summary analyses of the LST B1–C2 test forms are presented below. Acceptable values are in green font. Values which are greater than five LID scale points (a quarter of a logit) away from the established range are in red font.

Two sets of linked analyses for the composite LST tests are presented below. The first set provides a summary of percentile distribution values; the second provides a more visual impression in the form of item difficulty distribution graphs.

Table 3 provides the relevant detail for the composite LST tests. Each level has two sets of entries: the LID scale level range (in blue font) to the left-hand side and the distributions which emerged (in green font) to the right-hand side.

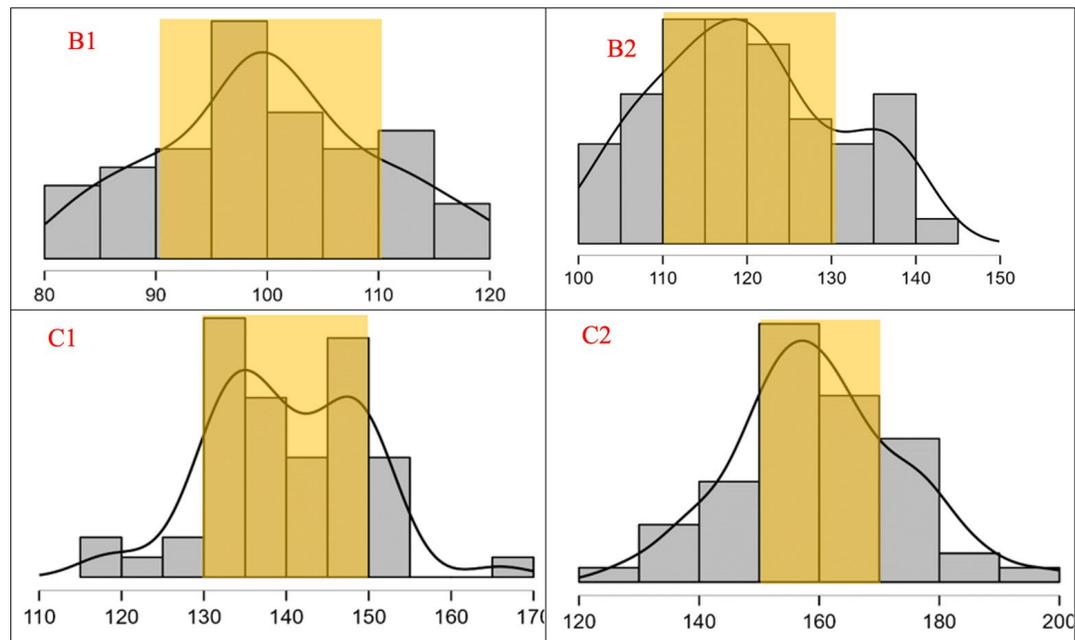
Table 3: Percentile distributions in composite LanguageCert SELT Test tests

No. of items	B1		B2		C1		C2	
Mean		52		52		52		56
SD		100		120.00		140.00		160
Maximum		9.59		10.83		9.28		14.09
75th p'tile		119.55		141.02		165.98		198.53
50th p'tile	110	105.64	130	126.43	150	147.69	170	167.96
25th p'tile		99.45		119.29		139.50		159.15
Minimum	91	94.04	111	112.78	131	133.45	151	150.72
		82.05		100.28		117.51		127.34

As can be seen, at the 25th percentile, all test levels are acceptably close to the lower LID scale range. Similarly, at the 75th percentile, all test levels are acceptably close to the upper LID scale range. Although there is a degree of divergence, the divergence is within the accepted half a logit (10 LID scale points) of difference (Zwick et al., 1999) which means that tests have been generally well targeted at their intended level.

To provide an accessible visual impression, test difficulty distributions are now presented in graph form in Figure 1. The green shading denotes the LID scale range for each test level. Frequency trend lines included across the scale for each test level provide a visual indication of the general shape of the distributions.

Figure 1: LanguageCert SELT Test tests: Test difficulty distributions



As can be seen, each level shows a broadly bell-shaped distribution, as confirmed by the best fit lines that wrap around the columns. The distributions are not perfect – C1 shows a somewhat irregular pattern in the centre of the graph. In general, however, the distributions are comparatively regular, indicating that the tests are performing as expected.

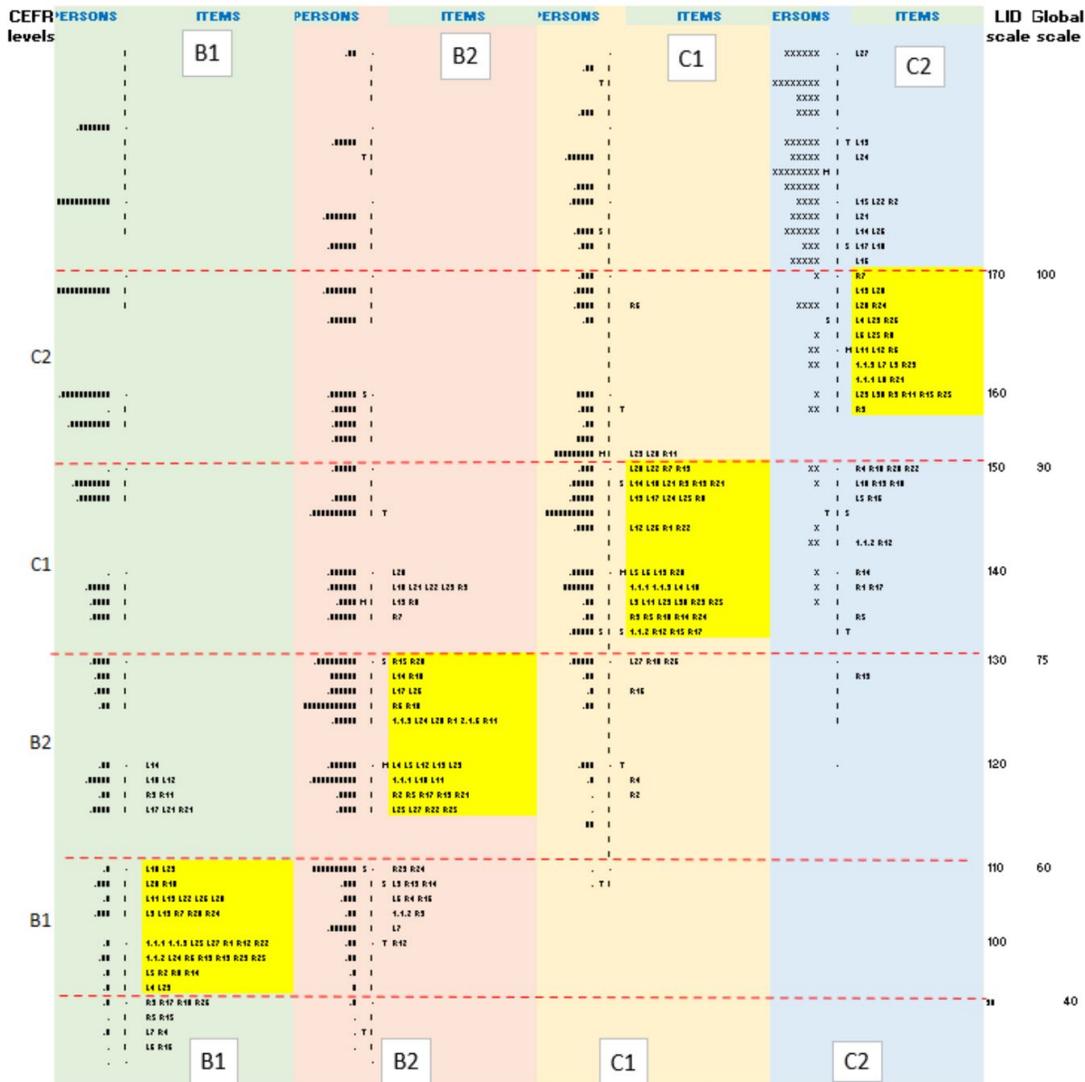
Placing LanguageCert SELT on the LID Scale

It has been established that the test forms have been well set and are robust in terms of fit statistics and reliability. The tests are located at appropriate points across the ranges of the LID scale, and hence at appropriate points against the CEFR.

Figure 2 below presents the Rasch person and item distributions on the LID and Global scales. The B1 test is green; the B2 salmon; the C1 beige; the C2 blue. LID scale values are to the right-hand side of the maps; CEFR levels to the left-hand side. The red tram lines indicate the LID scale cuts for each level. The highlighted yellow sections are the CEFR / test item match.

When reading the maps, it should be noted that candidates (persons) are located to the left-hand side of a particular map, items to the right-hand side. More able candidates are situated towards the upper left end of the map, and less able candidates towards the lower left end. More demanding items are situated towards the upper right end of the map while easier items are situated towards the lower right end.

Figure 3: LanguageCert SELT Test Common Scale



As can be seen from Figure 3, for each LST test, the majority of the items (the highlighted yellow sections) fall within the CEFR level for which they are intended. This is an indicator of validity, indicating that the LST tests are generally well set, and are being targeted at the appropriate level.

It is also clear from Figure 3 that while tests assess in general at a particular CEFR level, the tests also assess across levels. Taking the beige C1 test as an example and reading up from the bottom of the C1 row, it can be seen that the bulk of the items assess at C1 level, as intended. There are, however, a number of items which assess at B2, below C1, and another set which assess at C2, above C1.

Likewise, with the salmon-colour B2 test, the majority of items assess at B2 level, but substantial numbers also assess at B1 and at C1 levels. This is the value and utility of a common scale: the reach across levels. While tests in principle assess at a given level, with appropriate calibration, tests can also be used across levels.

Conclusion

This chapter has explored the alignment of LanguageCert SELT tests to the LID Scale. The use of externally-referenced anchoring has enabled the robustness of the four CEFR test levels B1–C2 to be demonstrated.

As the Rasch item/person maps illustrate, while the LST tests principally assess at their designated CEFR level, tests also contain items which assess across levels. At the C1 level, there are items which assess above and below C1. Likewise, at the B2 level, there are items which assess both above and below B2.

As stated earlier, being able to assess across levels provides valuable information for stakeholders such as tertiary level administrators, admissions tutors, immigration officials and so on. Stakeholders of the LanguageCert Academic and General test have identified this as a highly valued feature of these tests.

The research question pursued in the study was whether LanguageCert SELT tests could be accurately placed on the LID scale and hence the CEFR, accuracy being defined as good Rasch infit and outfit statistics being obtained in the analysis at each of the four test levels. Rasch levels were indeed within acceptable levels, supporting the claim that the tests are accurately placed.

This exercise forms part of the overall research drive that is being undertaken at LanguageCert to locate its various test products on the LID and hence LanguageCert Global Scale. The extensive research and calibration undertaken with the LanguageCert Test of English (Coniam et al., 2021a; b) is now being extended to other LanguageCert products. The research conducted with the SELT tests in the current study forms part of that endeavour.

Notes

1. The LanguageCert System reports scores on the LanguageCert Global Scale of 0-100 that is derived directly from the 180-point LID scale (see below). It provides candidates, employers, education institutions and government agencies an easy-to-understand results system. It applies across all the tests in the LanguageCert System. The Global Scale defines specific levels of attainment needed to fulfil certain requirements. For example, entrance into a university or for migration and employment purposes. The levels of attainment can relate to overall performance in an examination, performance by skill (e.g., speaking), or both these parameters.

The LanguageCert Global Scale



References

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., & Palmer, A.S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021a). *Validating the LanguageCert Test of English scale: The paper-based tests*. London, UK: LanguageCert.

Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021b). *Validating the LanguageCert Test of English scale: The adaptive test*. London, UK: LanguageCert.

Coniam, D., Zhao, W., Lee, T., Milanovic, M., & Pike, N. (2022). *The role of expert judgement in language test validation*. *Language Education & Assessment*.

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

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Humphry, S. (2006). *The impact of differential discrimination on vertical equating*. ARC report.

Lee, T., Milanovic, M., & Pike, N. (2022). Equating Rasch values and expert judgement through externally-referenced anchoring. *International Journal of TESOL Studies*, 4(1), 187-202.

Linacre, J. M. (2018). *Winsteps Rasch measurement computer program*. Winsteps.com: Beaverton, OR.

Milanovic, M., Lee, T., Coniam, D., & Papargyris, Y. (2022). *Externally-Referenced Anchoring of SELT tests*. London: LanguageCert.

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.

Appendix 1: LanguageCert SELT Test: Fit Statistics and Person Reliabilities

Test level	Rasch statistics summary
B1	<p>SELT B1 All</p> <pre> ----- PERSON 10810 INPUT 10810 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 42.0 51.8 140.85 11.27 1.00 .1 1.00 .1 P.SD 9.7 1.8 29.79 7.29 .05 .4 .25 .6 REAL RMSE 13.42 TRUE SD 26.59 SEPARATION 1.98 PERSON RELIABILITY .80 ----- ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 8731.7 10760.6 100.00 .58 1.00 -.2 1.00 -.3 P.SD 597.2 43.3 9.50 .06 .07 4.4 .18 4.9 REAL RMSE .59 TRUE SD 9.48 SEPARATION 16.19 ITEM RELIABILITY 1.00 ----- </pre>
B2	<p>SELT B2 All</p> <pre> ----- PERSON 2732 INPUT 2732 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 33.3 51.8 136.75 7.69 1.00 .0 1.00 .0 P.SD 11.2 1.8 26.17 3.99 .07 .7 .16 .8 REAL RMSE 8.66 TRUE SD 24.70 SEPARATION 2.85 PERSON RELIABILITY .89 ----- ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 1750.5 2722.8 120.00 .93 1.00 -.1 1.00 -.2 P.SD 258.8 6.8 10.72 .04 .08 4.1 .14 3.9 REAL RMSE .94 TRUE SD 10.68 SEPARATION 11.42 ITEM RELIABILITY .99 ----- </pre>
C1	<p>SELT C1</p> <pre> ----- PERSON 581 INPUT 581 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 32.4 52.0 153.57 7.03 1.00 .0 1.00 .0 P.SD 10.5 .4 22.54 2.64 .06 .7 .12 .7 REAL RMSE 7.51 TRUE SD 21.25 SEPARATION 2.83 PERSON RELIABILITY .89 ----- ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 361.8 580.6 140.00 1.96 1.00 -.1 1.00 -.1 P.SD 49.8 .7 9.19 .10 .09 2.4 .14 2.0 REAL RMSE 1.96 TRUE SD 8.98 SEPARATION 4.57 ITEM RELIABILITY .95 ----- </pre>
C2	<p>SELT C1</p> <pre> ----- PERSON 581 INPUT 581 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 32.4 52.0 153.57 7.03 1.00 .0 1.00 .0 P.SD 10.5 .4 22.54 2.64 .06 .7 .12 .7 REAL RMSE 7.51 TRUE SD 21.25 SEPARATION 2.83 PERSON RELIABILITY .89 ----- ITEM 52 INPUT 52 MEASURED INFIT OUTFIT TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD MEAN 361.8 580.6 140.00 1.96 1.00 -.1 1.00 -.1 P.SD 49.8 .7 9.19 .10 .09 2.4 .14 2.0 REAL RMSE 1.96 TRUE SD 8.98 SEPARATION 4.57 ITEM RELIABILITY .95 ----- </pre>

Chapter 4: LanguageCert SELT Writing Test Quality

David Coniam, Irene Stoukou, Tony Lee and Michael Milanovic

Abstract

This chapter focuses on a vital aspect of writing tests. It reports on a study into test quality on a sample of the LanguageCert SELT Writing Tests administered at CEFR levels B1 and B2 during the period 2021-2022. This was a large sample encompassing over 14,000 candidates, 60 examiners and 18 different tasks. Using Many-Facet Rasch Analysis (MFRA), the study explores the consistency of marking in terms of examiner, task, and rating scale fit and severity. The use of Many-Facet Rasch Analysis provides analysts with an understanding of test quality, superior to Classical Test Statistics. Further explanation is provided in the Glossary of statistical terms and techniques at the end of the volume.

Results from the study indicate that, for the different test facets, fit to the Rasch model was generally good. The task and rating scale severity ranges were generally within acceptable limits. Crucially, examiner fit was good, with only a small number of examiners exhibiting misfit. Against the backdrop of the analysis reported, the study concludes that the SELT Writing Tests pitched at CEFR levels B1 and B2 are robust and fit for purpose.

Keywords: test quality, Multi-faceted Rasch analysis, test facets

Introduction

One of the maxims of assessment is that tests should 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 and Palmer, 2010; Messick, 1989). Under such a precondition, the marking of candidates' writing 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). As Weigle (2002) observes, rating is a complicated process involving numerous factors – the candidate, the rater, the prompt, the rating scale etc – before a grade can be assigned to a script.

While scores awarded arise as a result of different facets in a Writing test – the examiners, the prompts, the rating scales – examiners are usually the facet which accounts for the largest source of variation, and hence inconsistency (Lumley and McNamara (1995). A considerable amount of research exists on examiner reliability (Saito, 2008; Webb et al., 1990); consistency (Elder et al., 2007; Lumley and McNamara, 1995); and severity (Engelhard and Myford, 2003). Other investigations of factors affecting examiners' rating have focused on: mother tongue, expertise, educational qualifications, professional background (Barkaoui, 2007; Cumming, 1990; Johnson and Lim, 2009; Shohamy et al., 1992).

From the issues just outlined, it follows that, for marking to be as consistent and accurate as possible, examiners need to be properly trained and standardised (Lumley and McNamara, 1995; Kang et al., 2019; Webb et al., 1990; Weigle, 1998). For details of the training of LanguageCert Writing Test examiners, see Papargyris and Yan, (2022).

Prompts, or tasks, need to be at the appropriate level, of comparative difficulty and free of bias as far as possible (Lim, 2009). Barkaoui and Knouzi (2012) explore writing tasks, describing how task variability needs to be controlled so that different tasks do not produce greatly different outputs, and do not affect scores awarded. In Weigle's (2002) terms, this means "construct irrelevant variance" should be minimised. LanguageCert task and item writers are of a high standard and have extensive expertise in, and understanding of, the different CEFR levels (Papargyris and Yan, 2022).

Rating scales need to interface with raters and tasks such that they also exhibit difficulty appropriate to the level being assessed, and possess good psychometric properties. Knoch et al. (2020) outline how rating scales may be evaluated for robustness.

SELT Writing Test Makeup

The data in this study were drawn from the administration of examinations at CEFR levels B1 and B2, which form part of LanguageCert's SELT suite of English language tests. In the LanguageCert SELT Writing Tests (LSWT), candidates complete two writing tasks which elicit a range of writing skills. Table 1 elaborates.

Table 1: Writing Test tasks

Level	Part 1: Candidates produce	Word length	Part 2: Candidates produce	Word length
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

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 become increasingly important as the CEFR level of the test increases.

Against the above backdrop, candidate responses are marked using an analytic mark scheme which matches the CEFR descriptors. Separate marks are awarded by marking examiners for four aspects of writing ability in the scripts produced by candidates. This set of criteria ensures that a wide range of writing skills are considered, thus enhancing the reliability and representativeness of test scores. Table 2 elaborates.

Table 2: Rating scale criteria

Accuracy and Range of Grammar
Accuracy and Range of Vocabulary
Organisation
Task Fulfilment

Data: Test Facets and the LID Scale

This section provides details of the dataset constructed for the analysis. This comprises the four facets used in the Many-Facet Rasch Analysis (MFRA) (detail provided below): the candidates, examiners, tasks, and rating scales. Table 3 provides the detail.

Table 3: Writing Test facet breakdown

CEFR level	Candidates	Examiners	Tasks	Rating scales
B1	11,054	58	18	4
B2	2,813	52	12	4

The focus in the current study is on CEFR level B due to candidature cohort size. The B1 candidature is over 11,000, while that of B2 is almost 3,000. The C level cohorts are considerably smaller and do not therefore form part of the current analysis. The sample sizes are a reflection of the number of applicants for the different visa types. The examiners constitute LanguageCert's trained cohort of examiners, who are trained and standardised to mark across levels (see Papargyris and Yan, 2022). There are a range of tasks: nine sets of Task 1s and Task 2s at B1, matching the larger candidature and six sets of tasks at B2.

The four rating scales were presented in Table 2. While the same four criteria are applied across levels, the demands posed by the criteria at a specific level reflect expectations of language ability at that level.

At LanguageCert, tests, items, and candidate test results are linked to the CEFR by means of the LanguageCert Item Difficulty (LID) scale. A description of the LID scale is provided in Chapter 2. LID scale ranges and midpoints for the two CEFR levels explored in the current study are presented in Table 4.

Table 4: LID scale ranges

CEFR level	LID scale range	Midpoint
A1	51-70	
A2	71-90	
B1	91-110	100
B2	111-130	120
C1	131-150	
C2	151-170	

An accepted first-line metric of examiner quality is that of correlations between examiners (see e.g., Tisi et al., 2013). Following accepted practice for analysing multiple facets in a performance test such as Writing, however, the best analytical instrument is MFRA (see e.g., Eckes, 2015).

In the current study, following an initial investigation of inter-examiner correlations, the main focus involves the use of Many-Facet Rasch Analysis (MFRA), which is conducted via the computer program FACETS (Linacre, 2020). The reader is referred to the outline of the Rasch measurement model and MFRA provided in the Glossary of statistical terms and techniques at the end of the volume.

Research Questions

The Research Questions pursued in the current study are as follows:

1. Do the different facets of examiner severity, candidate ability, task difficulty and rating scale difficulty exhibit good fit statistics?
2. Are task and rating scale difficulty appropriate to the test level?

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).

Data Analysis: Results and Discussion

Classical Test Analysis

Inter-examiner correlations are first provided for whole test scores, and individual task scores. Table 5 provides the detail.

Table 5: Inter-examiner correlations

CEFR level	Whole test	Task 1	Task 2
B1	0.86	0.84	0.85
B2	0.78	0.78	0.76

$p < .001$ for all correlations

As can be seen, against a preferred basis of 0.8, B1 and B2 whole test and task scores are good. While correlation analysis is seen as a first base in investigating issues such as examiner reliability, it is nonetheless viewed as being somewhat limited (Lunz et al., 1994). Analysis of a rather broader scope – such as that afforded by Many-Facet Rasch Analysis [MFRA] (see e.g., Eckes, 2015) – is recommended for performance tests such as Writing. And it is to MFRA that the discussion now moves.

Many-Facet Rasch Analysis

In the current study, as mentioned, four facets have been specified: candidates, examiners, tasks and rating scales. In the analysis, all things being equal (i.e., examiner severity, candidate ability, task difficulty and rating scale difficulty), measures will centre around zero logits (rescaled to the midpoint of the appropriate LID/CEFR level, with an SD of 20 [refer back to Table 4]). In terms of examiner judgements, a higher score indicates severity; a lower score indicates leniency. For candidates, a higher score indicates higher language ability, with a lower score indicating lower language ability. For tasks, a higher score indicates the task is more difficult, while a lower score indicates that the task is easier. For rating scales, a higher score indicates a more demanding scale.

In the analysis below, two perspectives are provided. A picture of global data-model fit is first provided for the two test levels. This is followed by the variable map which exemplifies the ‘ruler’ concept and how all facets may be viewed together.

Overall Data-Model Fit

A key focus in Rasch is that of overall data-model 'fit'. This is the difference between expected and observed scores, and can be observed through the number of unexpected responses. Satisfactory model fit is indicated when 'unexpected responses' account for no more than 5% of (absolute) standardised residuals (Linacre, 2002).

Table 6: Unexpected responses

Level	Valid responses	Unexpected responses
B1	94,772	957 (1.48%)
B2	25,696	175 (0.68%)

As can be seen from Table 6, for both test levels, the number of unexpected responses reported against valid responses used for estimating model parameters in the analysis was less than 5%. This is an indicator of acceptable data-model fit.

Facet Maps

A useful visual guide for understanding Rasch analysis is the facet map which provides a view of how the different facets are located on the scale. Figure 1 below presents a composite picture of the variable maps produced by FACETS for the B1 and B2 Writing Tests. The composite picture of both facet maps permits an appreciation to be gained not only of how the individual facets sit on the ruler for their specific test, but also provides a comparative picture of both tests.

Logit measures for both tests have been rescaled (from the standard logit midpoint of zero and an SD of 1) in line with LID scale ranges (Table 4). The midpoints, which are indicated by green bands, are set at 100 for B1 and 120 for B2. SDs for both levels are 20.

Candidates range across the whole ability spectrum, covering approximately 10 logits at each level, thus satisfying the requirements of the SELT tests for visa purposes. As a consequence of wide candidate variation, examiners will also show wide variation, as may be seen in the Appendices.

For current purposes, the map in Figure 1 has been limited to details of tasks and rating scales since it is preferable that these elements be within the specified difficulty domains for the respective CEFR level.

Figure 1: B1 and B2 facet maps

	B1		B2		LID
	Tasks	Scales	Tasks	Scales	
					143
					138
					137
					136
					135
					134
					133
					132
					131
					130
					129
					128
					127
					126
					125
					124
					123
					122
					121
					120
					119
					118
					117
					116
					115
					114
					113
					112
					111
					110
					109
					108
					107
					106
					105
					104
					103
					102
					101
					100
					99
					98
					97
					96
					95
					94
					93
					92
					91
					90
					89
					88
					87
					86
					85
					84
					83
					82
					81
					80
					79
					78
					77
Tasks	Scales	Tasks	Scales	LID	

Rating scales	Tasks	Scales	Tasks	Scales	LID
ARG Accuracy and range of grammar					
IO Organisation					
ARV Accuracy and range of vocabulary					
TF Task Fulfilment					

As can be seen from the maps, for the B1 test, the central zone (91-110 LID scale points) – contains all 12 tasks and three of the four rating scales (TF [Task Fulfilment] is marked leniently – see below).

Similarly, for the B2 test, the central zone (111-130 LID scale points) – contains all 18 tasks and three of the four rating scales (TF is again marked leniently).

The facet maps are useful as a visual guide to how the facets are located together on the one map, or 'ruler'. A more detailed analysis of the different test facets is now provided.

Analysis of Test Facets

In the data output and analysis presented below, infit and LID measures are reported for the examiner, task, and rating scale facets. In the tables, the infit data shows the 'big picture' in that it scrutinises the internal structure of a facet. Acceptable ranges of fit are generally taken as 0.5-1.5 (Lunz and Stahl, 1990).

Examiners

Appendix 1 presents the examiner fit statistics (sorted by infit) for the two test levels.

Table 7 presents the picture of examiner fit. Three examiners exhibited misfit at B1, with three misfitting examiners at B2. This figure of approximately 5% is acceptable, given the number of examiners.

Table 7: Examiner fit summary

CEFR level	Examiners	LID scale range (logits)	Examiners exhibiting misfit
B1	58	100 (5)	3
B2	52	65 (3.5)	3

The degree of examiner severity ranges from five logits between the 58 examiners on B1 to three and a half logits with the 52 B2 examiners. Such ranges are not unusual. Eckes (2005), in an analysis of the German Test-DaF Writing test, reports an examiner severity spread of 4.26 logits. Park (2004) reports an examiner severity range of 5.24 logits.

it should be noted that the facet of examiner 'severity/leniency' is not a value judgement. Severity reflects an examiner's tendency to award a rating lower than deserved while leniency reflects an examiner's tendency to award a rating higher than deserved. Severity/leniency should be understood in terms relative to the examiner facet alone without reference to other facets in the calibration or the calibrated Rasch measures in absolute terms.

In general, the picture with the B1 and B2 tests reported above is indicative of a good baseline of examiner consistency.

Tasks

Appendix 2 presents the task fit statistics (sorted by LID measure) for the two test levels. Table 8 presents task fit and difficulty.

Table 8: Task fit summary

CEFR level	Tasks	LID scale range: Measures (logits)	Misfit
B1	18	8 (0.4)	-
B2	12	10 (1.0)	-

All task fit values are good, indicating that the tasks generally perform well. The degree of task severity is limited, within half a logit for B1 and one logit for B2. While not absolute, the more demanding Task 2s have higher LID values, appearing at the more difficult end of the spectrum. This is possibly because the Task 2s are required to be longer, and hence impose greater cognitive demands on candidates, leading to the assessment of a wider range of ability (see e.g., Crossley, 2020; Rubin and Raftery, 1986).

Rating Scales

Appendix 3 presents the rating scale fit statistics (sorted by LID measure) for the two test levels. Table 9 presents scale fit and difficulty. All task fit values are good, within acceptable levels, an important baseline.

Table 9: Rating scale fit summary

CEFR level	Scales	LID scale range (logits)	Misfit
B1	4	18 (0.9)	-
B2	4	29 (1.5)	-

The four rating scales show good model fit, with the range among the different scales extending to approximately one logit. The rating scales nonetheless illustrate a pattern observed in previous research: that the most demanding scales tend to be those involving the formal 'expressive' categories – grammar and syntax, for example (Pollitt and Hutchison, 1987). The Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, and Organisation scales were within a half logit range of one another. Task Fulfilment, the least 'formal' scale, was the most leniently marked, as this type of scale has generally tended to be (Coniam, 2005). While English language teacher-examiners have a clear idea of how to interpret the formal categories, they are less clear about the demands of scales such as Task Fulfilment.

Conclusion

This study has examined the issue of facet quality across the LanguageCert SELT B1 and B2 Writing Tests. The study employed inter-examiner correlations initially, but, for the most part, has drawn on Many-Facet Rasch Analysis in its exploration of test quality.

The research questions in the study centred around the extent to which the different test facets exhibited good fit statistics, and how far task and rating scale difficulty were appropriate to test level.

Inter-examiner correlations were good for B1 and B2 levels.

In terms of the analysis of the test facets, examiner fit to the Rasch model was generally good – a key background consideration. There was a range in terms of examiner severity, but this was consistent with severity ranges from previous studies and to an extent reflected the wide ability range of the candidature.

Regarding tasks, all task fit values were good, and task difficulty values indicated that the tasks generally performed well. The task difficulty range was under a logit, and tasks can be seen to be appropriate for their intended level.

The analysis of the rating scales illustrated a somewhat familiar pattern. While the scales showed good model fit, severity range among the scales extended to approximately a logit and a half on the B2 test. This was largely due to the fact that, on the two tests, the Task Fulfilment scale was most leniently marked – as this type of scale generally tends to be. A tightening up of expected performances in the Task Fulfilment scale would help to better target rating expectations.

In light of the analysis of the data reported here, the SELT B1 and B2 English Language Writing Tests may be seen to be both robust and fit for purpose.

It should be noted that, as also stated in Chapter 3, the use of (MFRA) provides more detailed, nuanced and reliable information for analysts and stakeholders than Classical Test Analysis. This is particularly important in high-stakes tests of English language proficiency such as the LanguageCert Academic examination (LCA). Chapter 1 has demonstrated how MFRA was used to ensure that both the LCA and LCG are valid, reliable, quality examinations,

References

- Bachman, L. F., & Palmer, A.S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing writing*, 12(2), 86-107.
- 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*. Leiden, NL: Brill.
- Coniam, D. (2005). The impact of wearing a face mask in a high-stakes oral examination: An exploratory post-SARS study in Hong Kong. *Language Assessment Quarterly*, 2(4), 235-261.
- Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021). *Validating the LanguageCert Test of English scale: The adaptive test*. London, UK: LanguageCert.
- Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3), 415-443.
- Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-51.

- Eckes, T. (2015). *Introduction to many-facet Rasch measurement*. Frankfurt am Main: Peter Lang.
- Elder, C., Barkhuizen, G., Knoch, U., & Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37-64.
- Engelhard Jr, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and Composition program with a many-faceted Rasch model. *ETS Research Report Series*, 2003(1), i-60.
- Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, 26(4), 485-505.
- Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, 36(4), 481-504.
- Knoch, U., Zhang, B. Y., Elder, C., Flynn, E., Huisman, A., Woodward-Kron, R., Manias, E., & McNamara, T. (2020). 'I will go to my grave fighting for grammar': Exploring the ability of language-trained raters to implement a professionally-relevant rating scale for writing. *Assessing Writing*, 46, 100488.
- Lim, G. S. (2009). Prompt and rater effects in second language writing performance assessment.
- Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85-106.
- Linacre, J. M. (2020) *FACETS computer program for many-facet Rasch measurement*. Beaverton, Oregon: Winsteps.com.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater 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.
- Lunz, M. E., Stahl, J. A., & Wright, B.D. (1994) Interjudge reliability and decision reproducibility. *Educational & Psychological Measurement*, 54, 913-925.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition). New York: American Council on Education.
- Papargyris, Y., & Yan, Z. (2022). Examiner quality and consistency across LanguageCert Writing Tests. *International Journal of TESOL Studies*, 4(1), 203-212. doi.org/10.46451/ijts.2022.01.13.
- Park, T. (2004). An investigation of an ESL placement test of writing using manyfacet Rasch measurement. *Studies in Applied Linguistics and TESOL*, 4(1), 1-21.
- Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing*, 4, 72-92.
- Rubin, D. L., & Rafoth, B. A. (1986). Social cognitive ability as a predictor of the quality of expository and persuasive writing among college freshmen. *Research in the Teaching of English*, 9-21.
- Shohamy, E., Gordon, C. M., & Kraemer, R. (1991). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, 76, 27-33.
- Tisi, J., Whitehouse, G., Maughan S., & Burdett, N. (2013). *A review of literature on marking reliability research (Report for Ofqual)*. Slough: NFER.
- 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.
 Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Appendix 1: Examiner Fit Statistics (sorted by Infit) B1 Examiner Fit Statistics

(Logits rescaled to mean of 100; SD of 20) Yellow=largest and smallest severity values; green=misfit

<i>Examiner</i>	<i>LID</i>	<i>Infit</i>	<i>S.E.</i>
86041	96.31	1.79	7.10
546621	115.28	1.58	0.80
1655216	75.58	1.43	5.34
1664875	84.9	1.40	1.54
46342	92.29	1.36	0.58
1676912	75.78	1.31	1.17
808145	80.1	1.30	6.27
1652253	91.93	1.28	3.34
1643606	92.57	1.26	1.60
1672790	84.39	1.26	1.28
1652250	90.64	1.24	1.03
708446	125.49	1.20	4.73
181343	145.49	1.19	8.63
2112799	75.31	1.19	1.62
1664751	152.09	1.14	2.55
5941	122.55	1.13	10.99
2028104	97.46	1.13	4.64
1668578	62.13	1.10	1.91
1655206	99.08	1.09	2.48
2112802	115.4	1.07	3.23
1685135	126.57	1.07	5.12
5813	129.51	1.06	1.03
1655196	104.77	1.05	2.36
1673573	143.65	1.04	4.27
1652261	94.11	1.04	0.79
1685125	121.8	1.03	0.73
124236	95.59	1.02	24.1
953535	126.71	1.02	1.53
1681139	77.6	1.00	1.07
1664747	53.26	1.00	1.28
28729	124.95	1.00	1.09
1664753	84.79	0.99	1.02
2116474	84.71	0.98	1.05
1676916	78.98	0.98	1.08
1681140	56.35	0.96	1.99
17955	108.45	0.95	10.6
1366256	111.29	0.95	3.92
1652245	94.44	0.94	1.64
1643603	112.1	0.94	1.07
continued from previous column			
<i>Examiner</i>	<i>LID</i>	<i>Infit</i>	<i>S.E.</i>
8925	110.32	0.92	3.37
1667700	96.37	0.92	2.64
1655211	107.94	0.91	1.32
1672777	70.72	0.91	1.21
1648183	99.01	0.9	3.83
14592	102.42	0.89	0.77
2069067	98.17	0.88	1.06
1664778	66.02	0.86	1.35
1655247	111.32	0.79	3.05
2187924	75.80	0.79	1.4
2433349	80.42	0.76	14.93
1858871	114.98	0.76	2.93
2248452	102.98	0.75	0.88
2228716	144.41	0.69	9.73
18078	74.83	0.68	17.05
1668577	104.00	0.68	1.23
2085519	109.77	0.63	4.41
1211463	124.27	0.5	11.89
19459	98.22	0.48	0.46

B2 Examiner Fit Statistics

(Logits rescaled to mean of 120; SD of 20) Yellow=largest and smallest severity scores; green=misfit

<i>Examiner</i>	<i>LID</i>	<i>Infit</i>	<i>S.E.</i>
2028104	113.31	1.68	8.69
546621	133.21	1.53	1.16
1643606	106.29	1.36	2.34
1676912	121.22	1.34	1.94
46342	104.13	1.34	0.89
1664875	137.54	1.29	2.58
1652250	119.84	1.27	1.67
1366256	97.61	1.24	10.51
2248452	128.82	1.22	1.85
1672777	142.53	1.19	2.33
86041	122.54	1.17	7.76
1680800	84.70	1.17	2.80
1676916	109.72	1.16	1.68
1672790	115.88	1.15	2.22
1655216	99.88	1.12	15.55
1668578	111.72	1.10	2.53
1664753	99.79	1.09	2.04
1668577	103.88	1.07	2.58
1652253	103.13	1.06	5.22
1664747	101.87	1.06	2.31
2112799	121.50	1.05	3.12
1655196	144.61	1.03	3.33
1655206	137.45	1.02	3.19
5813	130.13	1.02	13.01
1664751	150.78	1.00	4.29
1652261	131.95	1.00	1.69
1655247	103.03	0.96	3.92
1685125	124.76	0.93	1.00
1643603	148.57	0.92	1.83
continued from previous column			
<i>Examiner</i>	<i>LID</i>	<i>Infit</i>	<i>S.E.</i>
2069067	119.56	0.90	2.09
1648183	135.66	0.88	7.2
2116474	118.14	0.87	1.93
1681140	110.46	0.87	2.72
1685135	140.62	0.86	6.06
1681139	109.17	0.85	1.94
14592	126.27	0.83	1.28
1673573	116.92	0.83	7.58
17955	131.92	0.81	6.42
1858871	131.77	0.75	5.71
1664778	124.66	0.74	1.88
28729	126.88	0.73	1.48
953535	134.20	0.71	3.3
1652245	117.85	0.71	3.52
1655211	118.56	0.69	2.22
1667700	117.76	0.68	5.58
2187924	116.50	0.67	2.45
15559	119.42	0.65	11.13
808145	107.45	0.64	9.5
708446	130.25	0.60	12.05
1211463	92.11	0.60	11.39
19459	121.05	0.54	0.71
2085519	104.6	0.34	10.86

Appendix 2: Task Fit Statistics (sorted by LID Measure)

B1 (Mean: 100; SD: 20)				B2 (Mean: 120; SD: 20)			
Task ID	LID	Infit	S.E.	Task ID	LID	Infit	S.E.
3268	104.07	1.05	0.91	1058	126.21	0.90	1.32
0084	103.32	0.99	0.69	2092	124.4	1.05	0.93
0106	103.00	1.04	0.98	2090	122.12	0.96	1.32
0082	102.51	0.99	0.72	1064	121.73	0.93	1.27
0093	102.25	1.00	0.69	2085	119.3	0.98	0.91
0101	101.82	1.02	0.73	2100	119.08	0.97	1.27
0096	100.37	0.91	0.67	1061	118.54	0.93	0.89
0065	99.84	1.06	0.73	1059	118.43	0.95	0.94
0063	99.57	1.01	0.65	2083	118.08	1.05	0.90
0052	99.12	0.94	0.73	2094	118.01	1.02	0.90
0081	98.88	0.90	0.74	1054	117.76	1.01	0.90
3267	98.79	1.01	0.92	1056	116.34	0.92	0.92
0069	98.66	1.05	0.98				
0062	98.46	0.99	0.67				
0055	97.72	0.93	0.70				
0053	97.64	0.97	0.73				
0060	97.51	0.94	0.69				
0099	96.47	0.99	0.67				

Appendix 3: Rating Scale Statistics (sorted by LID Measure)

Rating scale	Abbreviation
Task Fulfilment	TF
Accuracy and range of grammar	ARG
Accuracy and range of vocabulary	ARV
Organisation	IO

B1 (Mean: 100; SD: 20)				B2 (Mean: 120; SD: 20)			
Scale	LID	Infit	S.E.	Scale	LID	Infit	S.E.
IO	106.93	1.02	0.35	ARG	129.38	0.73	0.55
ARG	106.81	0.81	0.33	IO	128.02	1.21	0.59
ARV	98.37	0.79	0.34	ARV	123.67	0.81	0.56
TF	87.89	1.38	0.37	TF	98.94	1.24	0.61

Chapter 5: Exploring Item Bank Stability Through Live and Simulated Datasets

Tony Lee, David Coniam and Michael Milanovic

Abstract

LanguageCert manages the construction of its tests, exams and assessments using a sophisticated item banking system which contains large amounts of test material that is described, inter alia, in terms of content characteristics such as: macroskills, grammatical and lexical features; and measurement characteristics such as Rasch difficulty estimates and fit statistics. In order to produce content and difficulty equivalent test forms, it is vital that the items in any LanguageCert bank manifest stable measurement characteristics.

The current Chapter is the first of two linked studies exploring the stability of one of the item banks developed by LanguageCert [Note 1]. This particular bank has been used as an adaptive test bank and comprises 827 calibrated items. It has been administered to over 13,000 test takers, each of whom have taken approximately 60 items. The purpose of the two exploratory studies is to examine the stability of this adaptive test item bank from both statistical and operational perspectives.

The study compares test taker performance in the live dataset (with over 13,000 test takers who each takes approximately 60 items) with a simulated 'full' dataset generated using model-based imputation. Simulation regression lines showed a good match and Rasch fit statistics were also good – thus indicating that items comprising the adaptive item bank are of high quality both in terms of content and statistical stability. Potential future stability was confirmed by results obtained from a Bayesian ANOVA. As mentioned above, item bank stability is important when item banks are used for multiple purposes – two such purposes in the LanguageCert context being both adaptive testing and the construction of linear tests. The current study therefore lays the groundwork for a subsequent follow-up study where the utility of this adaptive test item bank is verified by the construction, administration and analysis of a number of linear tests.

Keywords: item banks, stability, simulated dataset, Rasch, Bayesian ANOVA

Introduction

This chapter reports on a study investigating the stability and robustness of one of the item banks developed by LanguageCert. Given that both linear paper-based and adaptive high-stakes tests are produced from such item banks, key issues that need to be confirmed are (i) item bank stability and (ii) item measurement quality, in terms of tests generated from such item banks (see Mills and Steffen, 2000). These issues are important because test quality is a vital consideration for any organisation administering high-stakes examinations, especially examinations such as the LanguageCert Academic (LCA) and General (LCG) examinations.

Item Bank Size and Stability

Operationally, a key question is how to establish the stability of an item bank from a measurement perspective. In this context, an item bank containing 827 items was used both as an adaptive test bank and for the generation of linear tests. 'Stability' may be defined here from two perspectives. The first is that model-fit statistics remain within acceptable ranges, even at the extreme ends of the percentile spectrum. The second, from an operational perspective, is that tests derived from the item bank produce comparable results when run with test takers.

One of the early researchers into item banking with particular reference to adaptive testing some five decades ago was Choppin (1968). Choppin's starting point was that an item bank of around 500 items was required, calibrated on 2,000 test takers. Ree (1981) conducted simulations of different adaptive test scenarios with differing test taker and item bank sample sizes. His recommendations, to an extent, echoed Choppin's findings: that an item bank comprising at least 200 items and calibrated on 2,000 test takers might be an acceptable starting point. Wainer (2000) describes an item pool consisting of some 800 items. Voss and Blumenthal (2020) describe a pool of 1,071 items calibrated on some 4,200 test takers.

Other researchers have nonetheless recommended rather larger item bank sizes. Derner et al. (2008) in discussing the construction of an item bank to measure technical skill attainment mentions 9,000 items as an optimal size, resources permitting. Similarly, Rudner (2009), in describing the development of the GMAT, states that for reasons such as security and broad construct coverage, the GMAT comprised over 9,000 items.

Among the limited number of researchers who have investigated stability (see e.g., Gao and Chen, 2005; Weiss and von Minden, 2012; Sahin and Weiss, 2015) studies have tended to focus on the theoretical construct in terms of how many (or rather how few) items might be necessary for information to be provided at appropriate θ levels (theta, for personal ability estimates) and with item parameters accurately estimated. While these studies provide an informative backdrop, the study reported in the current chapter differs somewhat in its approach to stability. The simulated 'full dataset' study reported in the current Chapter 4 is followed up in the next chapter (i.e., Chapter 5) with an study of the direct construction, administration and analysis of real world tests from the live item bank.

The LanguageCert adaptive item bank described in the current chapter contained (as of late 2021) 827 items, and subsets of approximately 60 items have been administered (as adaptive tests) to approximately 13,000 test takers. This gave a live dataset of 0.78 million data points against a theoretical maximum of 10.66 million

data points. The items, thus, were more than those Choppin (1965) estimated as necessary, and close to the number specified by Wainer (2000). They are fewer than the 9,000 items suggested by Derner et al (2008) when developing the GMAT. The 13,000 sample of test takers in the study far exceeds those suggested by Choppin (Ibid), Werner (Ibid), and Vlss and Blumenthal (Ibid).

In assessment situations where items need to be calibrated to a common scale, analysis need to take account of extensive amounts of missing data (Roth, 1998). This is particularly the case with the item bank and adaptive test as they stood at the end of 2021. As mentioned above, the adaptive bank contains many hundreds of items, with responses available for each test taker for only a small number of items. For the reliability of such an item bank to be demonstrated, item statistics therefore need to be computed such that missing values in the dataset are taken account of. This may be achieved by using imputed values (Peugh and Enders, 2004).

A number of methods for evaluating the effect of missing data have been explored: model-based imputation (Huisman and Molenaar, 2001); pairwise deletion (Zhang and Walker, 2008); maximum likelihood (Schminkey et al, 2016); and multiple imputation (Li et al., 2015). The consensus would appear to converge on model-multiple imputation, the method that was adopted in the current study.

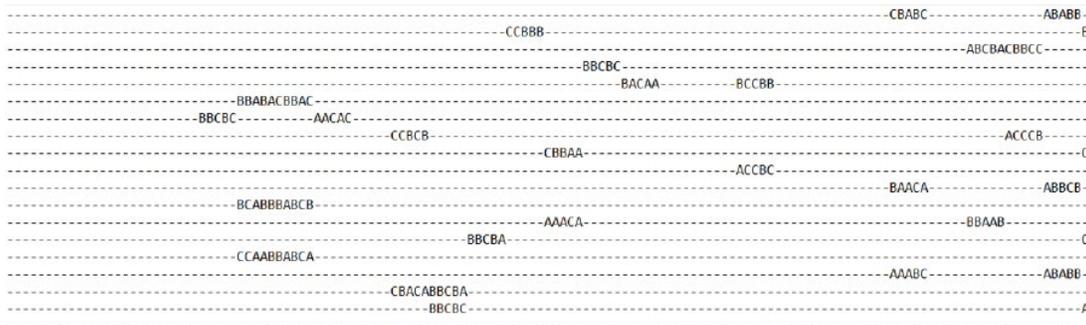
Multiple Imputation in the Current Study

The current study describes the analysis of the adaptive test item bank where missing data is simulated using the Rasch measurement program Winsteps (Linacre, 2018). Imputed missing data values have been generated via model-based multiple imputation, with the starting point for the simulation being the 13,000 test takers, with their individual responses to 60 items.

Methodology

Figure 1 below presents a snapshot of data of actual test taker responses in the adaptive test dataset.

Figure 1: Data points in the live LTE adaptive test dataset



As may be seen from Figure 1, the data occurs as chunks, spread out across a vast data space.

Bayesian statistical methods describe the conditional future probability of an event. An overview of the methodology surrounding Bayesian statistics and interpretations can be found in the Glossary. A brief recap of some key points is nonetheless presented below.

In Bayesian statistics, the critical statistic is the Bayes Factor (BF) – the ratio of likelihood between the null and the alternative hypothesis. Proposed cutoff levels for interpreting the strength of a Bayes Factor (Jeffreys, 1961) range from 1 (no evidence for the alternative hypothesis) to 10-30 (strong evidence), to 30-100 (very strong evidence), to above 100 (extremely strong evidence for the alternative hypothesis).

The credible interval in Bayesian statistics represents the range in which a specified percentage, e.g., 95%, of cases would fall. It has a direct interpretation as “the probability that p is in the specified interval” (Hoekstra et al., 2014).

Results

To explore the research questions, the simulation was run to explore how a potentially-complete dataset compared with the live dataset which contained missing data values.

Table 1 presents the regression line calculated for the simulation.

Table 1: Simulation regression line

Sample size	R ²
13,000	0.99

As can be seen, the simulation shows an extremely good match with the regression line of 0.99 – markedly above the target of 0.75 (Ringle and Sinkovics, 2009). This result gives an indication of the stability of the calibration of the adaptive item bank under investigation, which is underpinned, it has to be assumed, by the quality of the items which constitute the bank.

Item-model Fit Statistics

Table 2 below presents a comparison of item-model fit statistics between the live ('sparse') and simulated ('full') dataset. Unacceptably high values are flagged in red font.

Table 2: Live and Simulated datasets (N=827): Item fit statistics

Percentile statistics	Infit		Outfit	
	Live dataset	Simulated dataset	Live dataset	Simulated dataset
Mean	1.03	1.00	1.07	1.00
Std. Deviation	0.34	0.01	0.45	0.03
Minimum (1st p'tile)	0.54	0.98	0.98	0.93
25th p'tile	0.93	1.00	0.89	0.99
50th p'tile	0.97	1.00	0.95	1.00
75th p'tile	1.05	1.00	1.10	1.01
Maximum (99th p'tile)	4.76	1.02	5.18	1.37

As can be seen, at the 25th and 75th percentiles, fit statistics for values in the existing live dataset are well within the acceptable range of 0.5 – 1.5. It is only at maximum values that both infit and outfit mean squares emerge as being unacceptably high.

The simulated 'full' dataset presents a picture of stability – even at minimum and maximum percentile values (See Table 2). The larger standard deviations which emerge with the live dataset may be accounted for by the fact that each test taker in the live dataset has only 60 data points, as opposed to over 820 in the case for every item in the simulations.

The results using the simulated data suggest that the quality of the test items in the adaptive test item bank is high and that the adaptive test as it is currently calibrated would appear to be robust.

Person-model Fit Statistics

The explorations above have been at item level. To further explore the stability of the item bank, person-model fit statistics are now reported. Person values and misfit, it should be noted, are sometimes more difficult to derive clear interpretations out of than item values – certainly when these are all calibrated – due to the fact that test takers may guess, leave blanks, cheat etc. (see Meijer, 1996).

The current study builds on research by Coniam et al. (2021), which documented different phases of measurement scale development for the LanguageCert Test of English (LTE), validating the LanguageCert Item Difficulty (LID) scale. Test taker results are reported against CEFR (the Common European Framework of Reference for Languages) levels, which have been defined on the basis of LanguageCert Item Difficulty (LID) scale scores; these are laid out in Table 3 below.

Table 3: LID scale

CEFR level	Mid point
A1	60
A2	80
B1	100
B2	120
C1	140
C2	160

The LID scale in Table 3 above is key for the interpretation of Table 4 below, which presents person-model fit statistics for the live and simulated datasets, with unacceptably high values again in red font. LID values are also included in the table in order to provide a more in-depth picture of comparability.

Table 4: Live and simulated datasets (N=13,000): Person fit statistics

Percentile statistics	Live dataset			Simulated dataset		
	LID values	Infit	Outfit	LID values	Infit	Outfit
Mean	120.80	1.00	1.03	121.12	1.00	1.00
SD	18.86	0.17	0.39	18.93	0.04	0.15
Minimum (1st p'tile)	37.96	0.43	0.19	42.03	0.84	0.46
25th p'tile	108.46	0.89	0.80	108.63	0.97	0.92
50th p'tile	120.45	0.98	0.94	120.82	1.00	0.98
75th p'tile	134.43	1.10	1.15	134.76	1.03	1.05
Maximum (99th p'tile)	180.64	2.00	8.47	182.84	1.19	3.73

As can be seen, at the 25th, 50th and 75th percentiles, LID measures are constant with both datasets – indicative that the simulated dataset is a good extrapolation of the live dataset.

Fit statistics are within acceptable values. It is, again, only at maximum values that outfit mean squares in particular emerge as unacceptably high. This may well be due to outliers, i.e., test takers who have scored higher than they might have been expected to as a result of correct guesses. There is less misfit in infit and outfit mean square values in the simulated dataset than in the live dataset. This again suggests that – even though indications are that values computed from the current live dataset are stable and reliable – as the dataset increases in size, its stability will improve even further.

Bayesian Statistic Results

Bayesian statistics permit, as mentioned, the exploration of the probability-based future robustness of the adaptive test. To this end, a Bayesian ANOVA was run on the simulated dataset. The Bayesian H0 for ANOVA (as with the null hypothesis in standard [frequentist] statistics), is that there will be no significant difference among test means.

The descriptives for the simulation are presented in Table 5.

Table 5: Simulation Descriptives (N=827)

		95% Credible Intervals	
Mean	SD	Lower	Upper
100.29	36.15	97.82	102.76

The 95% credible intervals indicate that the fluctuations of the item bank mean in future events would be only two and a half LID scale points (about an eighth of a logit) above and below the mean with an extreme difference of about five LID scale points – approximately one quarter of a CEFR level. Since divergences within half a logit (10 LID scale points) are viewed as being bounds (Zwick et al., 1999), fluctuations of 2.5 LID scale points can be seen as non significant.

Against the above backdrop, the overall estimation from the Bayesian ANOVA is provided in Table 6.

Table 6: Bayesian ANOVA estimations

Models	P(M)	P(M data)	BF _M	BF ₀₁	error %
Null model	0.5	1	14,830	1	
Simulations	0.5	0.00006	0.00006	14,830	0.0007

In Bayesian ANOVA, the critical statistic is the BF01 Bayes Factor. This represents the ratio of BF0 (the null hypothesis of nil mean differences) to BF1 (the alternative hypothesis of existence of mean difference). The target Bayes Factor was 30-100; the figure of 14,830 obtained is far beyond this figure, into the range of above 100: “extreme evidence” (after Jeffreys, 1961) in favour of the no difference in mean in the ANOVA results. This figure is a strong indicator of robustness.

Conclusion

This study has explored how an item bank used for adaptive testing purposes may be assessed in terms of robustness. In the study, item bank stability was investigated using a simulated ‘full’ dataset generated through model-based imputation. Three research questions were pursued in this study.

RQ 1 investigated whether the regression line (R2) value of the simulation would be a minimum of 0.75. The R2 values for the simulation was 0.99, a strong indicator of stability of the calibration.

RQ 2 investigated whether Rasch infit and outfit statistics would be within acceptable ranges at the 25th and 75th percentiles. For both live dataset values and simulated dataset values (the latter using the ‘full’ dataset)

at both percentiles, fit statistics were well within acceptable ranges. There was evidence of misfit with outfit mean squares although this was only at the maximum value end of the scale.

RQ 3 investigated whether the Bayes Factor would be in the range of at least 30-100. The Bayes Factor which emerged was 14,830 – well above the target of 100, and indicative of “extreme evidence”.

The conclusion which may be drawn from the comparison of the ‘full’ and (comparatively sparser) live dataset was that as the live dataset expands in terms of data points (i.e., items and test takers), stability is likely to improve further. Such predicted stability lends credence to the claim that the items that comprise the adaptive item bank are of good quality and have been well set – and, furthermore, lends support to the robustness of the bank as an assessment instrument.

It has been noted both above and in Chapter 1 that item bank stability is a crucial element of a good, high-quality test, especially high-stakes tests where reliable and accurate outcomes are crucial for candidates and all other stakeholders. This is particularly the case with the LanguageCert Academic (LCA) examination where immigration and education opportunity decisions are based on the LCA results.

The current study has been laying the background for a follow-up study. The ground work – item bank stability – has now been established. Therefore, the follow-up study involves a real-world use of the item bank. This study will involve the construction, administration to a representative sample of test takers, and analysis of a number of linear tests derived from the adaptive item bank. This study is reported in Chapter 6 (see Coniam et al., 2022).

While the explorations reported in the current chapter relate to the analysis of a specific item bank, the methodology may be useful to any researcher developing an item bank. Creating a simulated ‘full’ dataset allows for a view of the stability of the item bank to be evaluated, with the two statistics used in this current study offering a picture of stability. A regression line above 0.75 gives a first line indication of the stability of the calibration of the item bank. The crucial figures, however, are calibration values and the Rasch infit and outfit statistics at the 25th and 75th percentiles. Since, in Rasch measurement, the starting point of measurement is the mid-point (the 50th percentile), the CEFR level of a test should start in the middle of the item distribution. The central 50% of the item distribution (25% to 75% of the items) is therefore the range most precisely measuring the CEFR level. Such a distribution shows that 50% of the items are well targeted at the CEFR level intended by the test, and means that half of the items in the test are around the CEFR level presumed by the test.

The fact that infit and outfit figures were within acceptable values provides further evidence of stability in the item bank. Finally, a Bayesian ANOVA permits a prediction to be made as to the likely future stability of the item bank. If the Bayes Factor obtained from the ANOVA is 30-100 or higher, this is further “very strong evidence” as to the likely long term stability of the item bank.

Notes

1. Reference is made in this chapter to “one” item bank. It should be noted that LanguageCert tests access multiple parallel item banks.

References

- Choppin, B. (1968). Item Bank using sample-free calibration. *Nature*, 219, 870-872.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). Validating the LanguageCert Test of English scale: The adaptive test. LanguageCert: London, UK.
- Coniam, D., Lee, T., Milanovic, M. (2022). Exploring Item Bank Stability in the Creation of Multiple Test Forms. LanguageCert: London, UK.
- Derner, S., Klein, S., & Hilber, D. (2008). Assessing the Feasibility of a Test Item Bank and Assessment Clearinghouse: Strategies to Measure Technical Skill Attainment of Career and Technical Education Participants. MPR Associates, Inc.
- Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three parameter logistic model. *Applied Measurement in Education*, 18(4), 351-380.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21, 1157-1164.
- Huisman, M., & Molenaar W. I. (2001). Imputation of missing scale data with item response models. In Boomsma, A., van Duijn, M., & Snijders, T. (Eds.). *Essays on item response theory* (pp. 221-244). New York: Springer-Verlag.
- Jeffreys, H. 1961. *Theory of probability* (3rd ed.). New York: Oxford University Press.
- Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: a flexible tool for handling missing data. *Jama*, 314(18), 1966-1967.
- Linacre, J. M. (2012). *A user’s guide to WINSTEPS*. Chicago, IL: Winsteps.com.
- Linacre, J. M. (2018). *Winsteps Rasch measurement computer program*. Beaverton, OR.
- Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession*, 13, 425-444.
- Meijer, R. R. (1996). Person-fit research. *Applied Measurement in Education*, 9(1), 3-8.
- Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In *Computerized adaptive testing: Theory and practice* (pp. 75-99). Dordrecht: Springer.
- Mislevy, R., & Wu, P. (1988). *Inferring examinee ability when some item responses are missing* (RR-88-48-ONR). Princeton NJ: Educational Testing Service. <https://doi.org/10.21236/ADA201421>.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525-556. <https://doi.org/10.3102/00346543074004525>.
- Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing*, 20, 277-319.

- Roth, P. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537-560.
- Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing* (pp. 151-165). New York, NY: Springer.
- Sahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences*, 15(6), 1585-1595.
- Schminkey, D. L., von Oertzen, T., & Bullock, L. (2016). Handling missing data with multilevel structural equation modeling and full information maximum likelihood techniques. *Research in Nursing & Health*, 39(4), 286-297.
- Voss, S., & Blumenthal, Y. (2020). Assessing the Word Recognition Skills of German Elementary Students in Silent Reading-Psychometric Properties of an Item Pool to Generate Curriculum-Based Measurements. *Education Sciences*, 10(2), 35.
- Vriens, M., & Melton, E. (2002). Managing missing data. *Marketing Research*, 14(3), 12.
- Weiss, D. J., & von Minden, S. V. (2012). A comparison of item parameter estimates from Xcalibre 4.1 and Bi-log-MG. St. Paul, MN: Assessment Systems Corporation.
- Zhang B., & Walker C. M. (2008). Impact of missing data on person-model fit and person trait estimation. *Applied Psychological Measurement* 32(6) 466-479.



Chapter 6: Exploring Item Bank Stability in the Creation of Multiple Test Forms

David Coniam, Tony Lee and Michael Milanovic

Abstract

The engine facilitating the construction of LanguageCert tests is a complex item banking system. These item banks contain large amounts of test material covering a wide range of content and construct characteristics. They are calibrated on the basis of Rasch difficulty estimates and fit statistics, and classical test statistics analysis.

When effectively constructed and managed, item banks allow for the creation of test forms which are consistent and comparable both in terms of content and difficulty. This is relevant not only when creating tests intended to measure at a particular level such as CEFR level B1 but also when developing tests which measure across multiple levels from A1 to C2.

This chapter reports the second in a set of linked studies investigating one of the item banks developed by LanguageCert UK. The first study (see Chapter 5) explored item bank stability in terms of model fit and regression line statistics in both the live dataset (13,000 test takers, each doing 60 items) and in a simulated 'full' dataset generated via model-based imputation (i.e., 13,000 test takers, each having done all 827 items). The purpose of the current study involved submitting the item bank to a real-world test in order to examine the quality of actual tests derived from the item bank. To achieve this, three tests were compiled from the calibrated item bank, and subsequently administered to a sample of test takers. In the analysis of the three tests, good fit statistics emerged, with high correlations between each test – an indicator of robust joint calibration and further evidence as to the stability of the item bank.

The chapter concludes with the claim that the items that comprise the item bank have been well set, with strong support for the robustness of the item bank as a clearinghouse from which many different tests may be constructed.

Keywords: item bank stability, Rasch, test development and administration

Introduction

This chapter reports on a study investigating the stability and robustness of one of the item banks developed by LanguageCert UK [Note 1]. Given that both paper-based and adaptive high-stakes tests are produced from these item banks, key issues are item bank stability and item quality in terms of tests generated from the item bank (Mills and Steffen, 2000). Indeed, test quality is of the utmost importance for any organisation administering high-stakes examination.

In the previous study, reported in Chapter 5 of this volume (Coniam et al., 2022), an overview was first presented of issues relevant to item bank size (Choppin, 1968; Ree, 1981; Derner et al., 2008; Rudner, 2009) and item bank stability (Gao and Chen, 2005; Weiss and von Minden, 2012; Sahin and Weiss, 2015). Following this, the makeup of the LanguageCert adaptive test item bank and test taker dataset was outlined. The adaptive item bank contains 827 items, with subsets of approximately 60 items administered (as adaptive tests) to approximately 13,000 test takers.

By using imputed values (Peugh and Enders, 2004), via the software *Winsteps* (Linacre, 2018), a simulated 'full' dataset was constructed on the basis of responses for each test taker being imputed for all 827 items based on test takers' actual responses. From the existing 0.78 million data points (13,000 x 60), a full dataset containing 10.66 million data points (13,000 x 827) was therefore generated.

In Chapter 5, two RQs were pursued in the study involving simulations.

The first investigated whether the regression line (R^2) value of the simulation would be a minimum of 0.75 [the rule of thumb for 'substantial' R^2 values – Ringle and Sinkovics (2009)]. The R^2 values for the simulation were, in fact, 0.99, a strong indicator of stability of the calibration.

The second investigated whether Rasch infit and outfit statistics would be within acceptable ranges at the key 25th and 75th percentiles. For both live dataset values and simulated dataset values at the both percentiles, fit statistics were well within acceptable ranges.

The conclusion drawn from the comparison of the 'full' and live (comparatively sparser) dataset was that as the live dataset expands in terms of data points (i.e., items and test takers), stability is likely to improve further. It was recommended that to provide corroborating evidence to support this claim, the item bank needed to be submitted to a real-world test whereby actual tests were generated, run and analysed. It is this procedure which the current study is now pursuing.

Assessment Context

The exploration reported in the current chapter relates to the LanguageCert Test of English (LTE). 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 LTE comprises three products, three level-agnostic tests, as in Table 1 below.

Table 1: LanguageCert Test of English (LTE) 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 test takers at all CEFR levels
(3) an adaptive test measuring CEFR A1-C2	Test for test takers at all CEFR levels

Reference is made in this chapter to “one” item bank. There are, however, a number of LTE item banks, with different banks used at different times to produce both paper-based and adaptive tests. Test taker results are reported against CEFR (the Common European Framework of Reference for Languages) levels, which have been defined on the basis of LID scale scores; these are laid out in Table 2 below.

Table 2: LID scale

CEFR level	Mid point
A1	60
A2	80
B1	100
B2	120
C1	140
C2	160

Rasch Measurement

The current study, as mentioned, is predicated on the use of the Rasch model. An overview of the methodology surrounding Rasch can be found in the Glossary of statistical terms and techniques, together with an outline of the infit and outfit mean square statistics which are key to the interpretation of Rasch results in the context of data ‘fit’.

The Study: Real-World Validation

While the results from the previous study suggest an a priori robust item bank, test taker results obtained from the item bank need be seen to be stable irrespective of what tests are constructed or derived – paper-based or adaptive – from the item bank. Consequently, in the current study, three live tests, produced in mid 2021 and administered to actual test taker groups, were analysed against calibrated values.

Three level-agnostic A1-C2 tests, each consisting of 110 items, were compiled from the recalibrated item bank. The three tests were constructed with a number of overlapping items from the item bank, thereby permitting

anchoring and direct comparisons to be made. The three tests were administered to a sample of European university students (average age 23 years), whose English language proficiency was estimated by their professors to range from B1 to C1.

Research Questions (RQs)

The current study is pursuing the following RQs.

RQ 1. Will Rasch infit and outfit statistics be within acceptable ranges: between 0.5-1.5 at the 25th and 75th percentiles?

RQ 2. Will LID values at the 50th percentile for all three tests be close to the usual calibration mid-point of 100 (i.e., within half a logit, or 10 LID points).

RQ 3. Will Pearson correlations between LID values across the three tests be over 0.8.?

Results

Table 3 below provides details of the number of different items in each test (i.e., overlapping items have been removed) and the number of test takers for each test.

Table 3: Test items and test takers

	Test 1	Test 2	Test 3	Totals
Discrete items	110	77	25	212
Test takers	108	241	228	577

The dataset comprised 212 discrete items administered to 577 test takers.

As mentioned, item difficulty in LTE tests is predicated on the overarching LanguageCert Item Difficulty (LID) scale (Table 2 above). For calibration purposes, the mid-point of the scale is set at 100 (B1 in LID value terms), with a standard deviation (SD) of 20 (see Coniam et al., 2021). This was the starting point for the analysis, with the Test 1 items first anchored with these values against the entire item bank. Table 4 presents a summary of the analysis of the whole item bank of 827 items with the 110 items in Test 1.

Table 4: Combined analysis of Test 1 with whole item bank

PERSON 13650 INPUT 13648 MEASURED					INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	36.3	59.3	123.29	6.69	.98	-.1	1.03	.1
P.SD	10.2	12.0	27.88	.89	.16	1.1	.50	1.2
REAL RMSE	6.75	TRUE SD	27.05	SEPARATION	4.01	PERSON RELIABILITY		.94
ITEM 934 INPUT 934 MEASURED					INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	529.9	866.5	100.16	2.76	1.00	-.2	1.07	.0
P.SD	462.6	582.8	41.21	3.50	.12	3.1	.37	3.7
REAL RMSE	4.46	TRUE SD	40.97	SEPARATION	9.19	ITEM	RELIABILITY	.99

As may be seen in the section highlighted in green in Table 4 above, infit and outfit figures are good – very close to 1.0. Reliability figures are high for both persons and items, being in the high 0.9 range.

Following the calibration of Test 1 to the whole item bank, Tests 2 and 3 were then calibrated against Test 1.

Table 5 presents the comparative results for the three tests.

Statistic	Test 1	Test 2	Test 3
Valid	110	77	25
Mean	100.00	97.26	101.55
Std. Deviation	28.22	38.73	26.34
Maximum (99 th percentile)	150.34	176.18	135.28
75 th percentile	121.26	126.93	125.75
50 th percentile	104.44	99.76	105.20
25 th percentile	83.49	69.63	82.67
Minimum (1 st percentile)	16.11	-8.41	53.19

As can be seen, at the 50th percentile, LID values for all three tests are close to the mean of 100 (B1). Means are likewise quite comparable at the 75th percentile of 120, where values are one logit (20 points) higher, at B2. There is some divergence at the lower end of the scale at the 25th percentile: Tests 1 and 3 are one logit lower around 80 – CEFR level A2. Test 2 exhibits a wider range of difficulty, and is another half logit lower at 69.63. Despite the divergences (some of which may be attributed to guessing), it can be seen that the three tests are broadly aligned.

Table 6 presents the Pearson correlations between LID values across the three tests.

Table 6: Correlations between LID values

		Test 1	Test 2
Test 2	Correlation	0.923	—
	p	< .001	—
Test 3	Correlation	0.917	0.964
	p	< .001	< .001

The scores from all three tests correlate very highly – above 0.9, with p values $< .001$. This is an indicator that the joint calibration of the three tests is robust.

Conclusion

The current study has explored how a dataset such as the LanguageCert Test of English (LTE) adaptive test may be assessed in terms of robustness. In the study, the item bank was submitted to a real-world test where-by three tests were compiled from the calibrated item bank, and administered to a representative sample of test takers. Three research questions were pursued in this study.

RQ 1 investigated whether Rasch infit and outfit statistics would be within acceptable ranges: between 0.5-1.5 at the 25th and 75th percentiles. Infit and outfit statistics were within acceptable ranges.

RQ 2 investigated whether LID values at the 50th percentile for all three tests would be within half a logit of the usual calibration midpoint of 100. At the 50th percentile, LID values for the three tests were within five LID scale points (a quarter of a logit) of the mean of 100.

RQ 3 investigated whether Pearson correlations between LID values across the three tests would be above 0.8. Test scores from all three tests correlated at above 0.9.

The previous – background simulation – study (Lee et al., 2022) was indicative of current, and future, stability. The current study's real-world testing of tests compiled from the item bank and administered to a representative sample of test takers provides corroborating evidence for the above claim. In the current study, good fit statistics, comparable LID score levels at major percentile levels, and high inter-test correlations emerged on the three tests: all of which underscore the stability of the item bank.

The conclusion that may therefore be drawn from the current study is that the items that comprise the item bank used in the construction of LTE tests are of good quality and have been well set. This in turn supports the claim regarding the robustness of the LTE as an assessment instrument. As has been pointed out in Chapter 1, the LanguageCert Academic and General examinations developers have drawn on the considerable assessment expertise of their research colleagues and the evidence of their research to ensure that the high-stakes examinations they are specifying, developing, piloting, and administering will be of the highest possible quality, adhering to strict standards and guidelines.

Notes

1. Reference is made in this chapter to “one” item bank. It should be noted that LanguageCert tests access multiple parallel item banks.

References

- Choppin, B. (1968). Item Bank using sample-free calibration. *Nature*, 219, 870-872.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). Validating the LanguageCert Test of English scale: The adaptive test. *LanguageCert*: London, UK.
- Derner, S., Klein, S., & Hilber, D. (2008). Assessing the Feasibility of a Test Item Bank and Assessment Clearinghouse: Strategies to Measure Technical Skill Attainment of Career and Technical Education Participants. MPR Associates, Inc.
- Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, 18(4), 351-380.
- Lee, T., Coniam, D., & Milanovic, M. (2022). Exploring Item Bank Stability Through Live and Simulated Datasets. *Journal of Language Testing & Assessment* (2022) Vol. 5: 13-21.
- Linacre, J. M. (2012). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.
- Linacre, J. M. (2018). *Winsteps Rasch measurement computer program*. Beaverton, OR.
- Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession*, 13, 425-444.
- Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In *Computerized adaptive testing: Theory and practice* (pp. 75-99). Springer, Dordrecht.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525-556.
- Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing*, 20, 277-319.
- Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing* (pp. 151-165). Springer, New York, NY.
- Sahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory & Practice*, 15(6), 1585-1595.
- Weiss, D. J., & von Minden, S. V. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bi-log-MG*. St. Paul, MN: Assessment Systems Corporation.
- Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, 16(4), 33-45..





SECTION 3: ORIGINAL RESEARCH



Chapter 7: The Role of Expert Judgement in Language Test Validation

David Coniam, Tony Lee, Michael Milanovic, Nigel Pike and Wen Zhao

Abstract

The calibration of test materials generally involves the interaction between empirical analysis and expert judgement. This chapter explores the extent to which familiarity with a scale might affect expert judgement as a component of test validation in the calibration process. It forms part of a larger study that investigates the alignment of the LanguageCert suite of tests, Common European Framework of Reference (CEFR), the China Standards of English (CSE) and China's College English Test (CET).

In the larger study, Year 1 students at a prestigious university in China were administered two tests – one with items based on China's College English Test (CET), designed by the university's staff, and the other a CEFR-aligned test developed by LanguageCert with items taken from the LanguageCert Test of English (LTE) item bank. Comparable sections of the CET and the LTE involved sets of discrete items targeting lexico-grammatical competence.

In order to ascertain whether expert judges were equally comfortable placing test items on either scale (CET or CEFR), a group of professors from the university in China, who set the CET-based test, were asked to expert judge the CET items against the nine CSE levels with which they were very familiar. They were then asked to judge the LTE items against the six CEFR levels, with which they were less familiar. Both sets of expert ratings and the test taker responses on both tests were then calibrated within a single frame of reference and located on a single scale – the scale developed and used by LanguageCert.

In the analysis of the expert ratings, the CSE-familiar raters exhibited higher levels of agreement with the empirically-derived score levels for the CET items than they did with the equivalent LTE items. This supports the proposition that expert judgement may be used in the calibration process where the experts in question have a strong knowledge of both the test material and the standards against which the test material is to be judged. However, when the judges were asked to place LTE items on the CEFR scale the results suggest that expert judgement may be less reliable in the evaluation of unfamiliar items against less familiar standards. The study

supports the proposition that expert judgement can be a useful dimension of item calibration but highlights the importance of familiarity in improving the accuracy of such judgement.

This is the case with the LanguageCert Academic and General examinations where the item writers as well as assessors/raters must be of the highest possible quality and fully aligned to the CEFR levels in the first instance. As these exams become more widely used in different parts of the world, similar alignments will need to take place with other relevant frameworks.

Keywords: expert judgement, test validation, reading and usage, CEFR, CSE

Introduction

This chapter focuses on a comparison of the ratings given by expert raters when they are familiar with the scales of an examination that they are rating, contrasted with the same expert raters rating a similar, albeit different, examination using the rating scales for that examination with which they are less familiar.

The chapter focuses on the issue of examiner scale familiarity as a vital component for validity when high stakes examinations that use expert raters are being calibrated. The study reported here provides data for the vital, underlying component of validity as part of a larger study. The larger study illustrates how calibration of rating data from both the CEFR (Common European Framework of Reference) and the CSE (China Standards of English) is possible.

The CSE and the CEFR

For the past two decades, the CEFR has come to be accepted as illustrating standards of language ability by many stakeholders: policy makers, exam bodies and test developers (Deygers et al., 2018). Not only in Europe, but in many countries around the world (Little, 2007), the CEFR has become the common currency for specifying levels of language ability (Figueras, 2012). The CSE reflects an overarching notion of language ability, in which language knowledge and strategies co-function in order to perform a language activity. The development of the CSE endeavours to pull together all of China's different English language curriculums and assessment instruments into one overarching framework.

Jin et al. (2017) describe the development of a Common Chinese Framework of Reference for English (CCFR-E): Teaching, Learning, Assessment, which began to be developed in 2014. The China Standards of English (CSE), which emerged from the research was released in 2018, and has three major levels, each subdivided into three sublevels. Figure 1 illustrates the two frameworks and how they are aligned.

Figure 1: CEFR and CSE levels

Common European Framework of Reference (CEFR)		China Standards of English (CSE)	
Label	Level	Level	Label
Proficient User	C2	Level 9	Advanced Stage
	C1	Level 8	
Independent	B2	Level 7	
	B1	Level 6	Intermediate Stage
Basic User	A2	Level 5	
	A1	Level 4	
		Level 3	Elementary Stage
	Level 2		
	Level 1		

The Use of Expert Judgement in Language Assessment

'Expert judgement' in language assessment has been a long-accepted practice in test development both in item writing and in the estimation of item difficulty – which in turn impacts level setting and cut scores. In the case of test setting, the use of the 'expert' is critical. In a study of minimally-trained item writers, Coniam (2009) reported such personnel as achieving a quality setting rate of only approximately 20%; i.e., items that may be defined as having good item statistics (see Falvey et al., 1994). A number of ground rules for the setting of good items was proposed by Haladyna and Downing (1989); many of these also appear in Alderson et al.'s (1995) discussion of the qualities needed of an "expert item writer". To be able to produce good tests efficiently – with good items and an accurate reflection of a given proficiency level – it is therefore clear that test item writers need to have both familiarity with and experience of the test they are engaging with, as well as being well-trained.

There has been considerable discussion of the use of expert judgement in standard setting, with research on one side supporting the use of experts (e.g., Shiotsu, 2010), with some dissenting voices from other quarters (see e.g., Mehrens, 1995; Alderson and Kremmel; 2013).

Studies comparing expert judge ratings against expected difficulty or empirical scores have reported mixed results. Studies which required independent predictions of judges have reported correlations in the 0.3 range (see e.g., Melican et al., 1989; Hambleton et al., 2003). In contrast, studies which provided raters with both a clear framework and adequate training have reported higher correlations – in the 0.7 range (see e.g., Attali et al., 2014). Lu and Read (2021), whose study compared two groups of experts' judgements on reading task item content, reported a general convergence of about 53% of the items.

Generally, the use of expert judgement has been widely employed in the field of language assessment for test validation and standard setting (e.g., Bachman et al, 1995). In recent expert judgement validation studies, judges have reportedly reached comparatively high levels of agreement (e.g., Gao and Rogers, 2011; van Steensel et al., 2013).

Against the above backdrop, the current study presents a comparative picture. In the first instance, a single group of experts rate items from a test with which they are familiar against a scale that has standards with which they are also familiar. In the second instance, the same group of experts rate items from a test with which they are not familiar against a scale that has standards with which they are much less familiar.

Within this context, four sets of data, consisting of discrete item subtests, constitute the dataset [Note 1]. The dataset is drawn from the College English Test (CET) that is calibrated to the CSE, and the LanguageCert Test of English (LTE), that is calibrated with the CEFR scales,

- Test taker results on a CET test
- Test taker results on an LTE test
- Expert judgement of CET discrete test items against the CSE
- Expert judgement of LTE discrete test items against the CEFR

Both sets of test taker responses to, and expert ratings of, test items were then calibrated together within a single scale of reference; following which the expert ratings were analysed on a single scale – the scale developed and used by LanguageCert.

Current Study: Assessment Instruments, Test Taker Sample, Research Questions

This section briefly outlines the background and make-up of the tests and the self-assessment ratings which test takers completed.

Test Material

In late 2020, approximately 2,500 Year 1 CET students took a 65-item multiple-choice reading and language use test prepared by experts from the China university. Three months later, this same set of students took a 53-item multiple-choice reading and language use CET test adapted from previously-validated LanguageCert Test of English (LTE) material (Coniam et al., 2021). The items in the LTE test used in the study were selected on the basis of having been calibrated to represent the spectrum of difficulty across the six CEFR levels.

Item difficulty in LTE tests is predicated on the overarching LanguageCert Item Difficulty (LID) scale; see Table 1. This scale lays out item difficulty levels generally adopted in LanguageCert assessments (Coniam et al., 2021).

Table 1: LID scale

CEFR level	LID scale range
A1	51-70
A2	71-90
B1	91-110
B2	111-130
C1	131-150
C2	151-170

For analysis and calibration purposes, 100 is taken as the mid-point of the scale. To this end, Rasch logit values are rescaled to a mean of 100 and a standard deviation (SD) of 20 (see Coniam et al., 2021).

As Zhao and Coniam (2022) describe, in a comparative analysis of the make-up of the two reading and usage tests, despite some differences, the content of the two tests, and even the order in which the different sections of the test appeared to test takers, exhibited a great deal of similarity.

The data in the current study involved discrete items which assessed grammar, vocabulary and usage. There were 30 such items in the CET test and 23 items in the LTE test. Appendix 1 provides detail on the two tests; Section 2 are the items in focus in the current study. There were eight expert judges, professors from the Foreign Studies Dept, all of whom had been involved in setting CET items for their students at the university. Given the relevance and status of the CSE in China, the eight expert judges had a clear picture of standards in the CSE. Also, given the fact that they were all English language professionals, most had knowledge of, albeit not in-depth familiarity with, the CEFR.

Before rating took place, training and standardisation sessions were conducted for the expert raters participating in the study. First, they rated sample CET items using the nine-level CSE. Second, they rated sample CEFR items using the six-level CEFR. Following this, raters were given the 30 College English Test items (CET) to rate against the nine CSE levels and 23 LTE items to rate against the six CEFR Levels.

The outcomes of the current study feed into the larger study mentioned previously, the overarching purpose of which involves exploring the alignment between the CEFR and the CSE in the context of the reading and usage element of LanguageCert tests. Consensus by the China expert raters on the CSE-related items will lend support to validating and triangulating the alignment that emerged between the datasets in the student self-assessments and performance on the CSE- and CEFR-related tests (see Zhao and Coniam, 2022).

Research Questions

The hypothesis being pursued in the current study is that raters will demonstrate a high level of agreement rating the CET items against the CSE, with which they are very familiar. A lower level of agreement, it is hypothesised, will be obtained from the expert ratings of the LTE items against the CEFR, with which they are

less familiar. On the basis of previous research findings, the research questions in the current study are therefore as follows:

RQ 1: On the CET items, will a high level of agreement between assessed and expert-rated values be obtained, and will such agreement be exhibited via a kappa value of 0.8 ('strong agreement')?

RQ 2: On the LTE items, will a moderate level of agreement between assessed and expert-rated values be obtained, and will such agreement be exhibited via a kappa value of 0.6 ('substantial agreement')?

Statistical Analysis

This section briefly outlines the statistics used in the current study.

Rasch Measurement

The manner for gauging test fitness-for-purpose in the current study, and for linking the data – the two different tests and self-assessments – involves the use of Rasch measurement. An overview of the methodology surrounding Rasch can be found in the Glossary, along with an outline of the infit and outfit mean square statistics which are key to the interpretation of Rasch results in the context of data 'fit'.

Kappa

Cohen's Kappa is a statistical measure for examining the agreement between two rated categories. It aids in determining the implementation of a given coding system.

In the current study, Kappa helps to assess levels of agreement between the two variables -- rater mapped and assessed CEFR levels. Following recoding as 1-6 (A1=1 through to C2=6), weighted kappa values are calculated via SPSS. According to Landis and Koch (1977), a level of 0.21 – 0.40 for kappa indicates 'fair agreement', 0.41 – 0.6 'moderate agreement', 0.61 – 0.8 'substantial agreement', and 0.8 or better 'strong' agreement. These are the values referred to in the research questions laid out above.

Data and Frame of Reference

To recap, there are four sets of assessment data in the current study (see Appendix 1):

- 30 CET items expert rated against the nine CSE levels
- 23 LTE items expert rated against the six CEFR levels
- test taker scores on the 53-item LTE
- test taker scores on the 65-item CET

Since all four datasets were collected from the same test takers, the data configuration may be taken as a unified collection, in that all data are referenced to the same candidates and to their English language ability. The person links (Boone, 2016) in the four datasets embrace a coherent frame of reference (FOR), defined by Humphry (2006) as “compris[ing] a class of persons responding to a class of items in a well-defined assessment context.” While the expert judges only rated the discrete items in each test (23 in the LTE, and 30 in the CET), all items in both tests (53 LTE items, and 65 CET items) were included in the calibration. The reason for this is that the assessed locations – the assessed values – of the expert-judged items need to be expressed in the context of the whole FOR, that is within the FOR of the total number of items (118) in both CET and LTE tests.

In the one frame of reference calibration, scores for all four elements were converted to the same measurement scale – LID scale values, as laid out in Table 1 above.

Results

Since the current study involves rater judgement of item difficulty, a baseline in the analysis involves rater consistency in the judgements made. A summary of the analysis of the eight judges’ consistency is presented below. As mentioned above, acceptable tolerance for fit is generally taken as ranging from 0.5 to 1.5 (Lunz and Stahl, 1990).

Table 2 reports the judges’ rating of the CET items, and Table 3 their rating of the LTE items.

Table 2: CET items rated by expert judges

Fair		Model	Infit	Outfit	
Average	Measure	S.E.	MnSq	MnSq	Judges
4.25	-2.44	0.29	0.64	0.67	1
5.37	0.05	0.27	1.14	1.13	3
5.84	1.03	0.26	0.87	0.89	4
4.28	-2.36	0.29	1.50	1.42	5
6.17	1.71	0.26	0.89	0.88	6
5.16	-0.38	0.27	0.67	0.65	7
6.53	2.40	0.27	1.20	1.17	8
5.37	0.00	0.27	0.99	0.97	Mean
0.89	1.89	0.01	0.31	0.28	SD Sample

RMSE .27 Adj (True) S.D. 1.87 Separation 6.89 Strata 9.52 Reliability .98

Fixed (all same) chi-square: 281.8 d.f.: 6 significance (probability): .00

Table 3: LTE items rated by expert judges

Fair		Model	Infit	Outfit	
Average	Measure	S.E.	MnSq	MnSq	Judges
4.25	-2.44	0.29	0.64	0.67	1
5.37	0.05	0.27	1.14	1.13	3
5.84	1.03	0.26	0.87	0.89	4
4.28	-2.36	0.29	1.50	1.42	5
6.17	1.71	0.26	0.89	0.88	6
5.16	-0.38	0.27	0.67	0.65	7
6.53	2.40	0.27	1.20	1.17	8
5.37	0.00	0.27	0.99	0.97	Mean
0.89	1.89	0.01	0.31	0.28	SD Sample

RMSE .25 Adj (True) S.D. .94 Separation 3.72 Strata 5.30 Reliability .93
Fixed (all same) chi-square: 44.4 d.f.: 3 significance (probability): .00

Reliability for both tests was high at 0.9, an important baseline criterion. Fit statistics were generally good. When the experts judged the CET items, infit and outfit figures for all eight judges were good, indicating that they all fit the model. When the experts judged the LTE items, Judge 1's infit and outfit statistics were below the 0.5 threshold, indicating misfit. The general picture, however, is that rater consistency is good.

Test Taker and Expert Judge Results: Equating Data-sets

Table 4 presents the results for the scores which emerged from test takers' scores on the tests and from judges' ratings of item difficulty. Both tests, it should be noted, were anchored at 100 -- the mid-point of the LanguageCert scale, at which all LanguageCert tests are anchored (see Chapter 7, this volume).

Table 4: Test takers' mean scores and judges' mean ratings of common items

Item type	Items	Mean LID value	SD	Reliability
CET assessed	30	104.28	22.43	1.00
CET rated	30	96.25	20.22	0.84
LTE assessed	23	102.96	32.10	0.97
LTE rated	23	104.89	32.77	1.00

As a baseline, reliabilities for all four elements of the dataset were high, above 0.8.

On the LTE test (anchored at 100), test takers' mean score (the "assessed" mean) was 102.96. The expert judges' mean rating of the 23 discrete LTE items (the "rated" mean) was 104.89, a difference of 1.93.

On the CET test (also anchored at 100), test takers' mean score (the "assessed" mean) was 104.28. The expert judges' mean rating of the 30 discrete CET items (the "rated" mean) was 96.25, a difference of 8.03.

While the standard deviations (SD) are broadly comparable within each pair of tests, the SDs differ considerably between tests. In light of this, the means and SDs need to be aligned into a single frame of reference (see Linacre, 2009). Such alignment is detailed in the following section.

Equating the Two Sets of Test Results

The two sets of test results are, as mentioned, from two different samples (test taker responses and judges' ratings), and hence from different frames of reference (FOR). As may be observed from Table 4, the two datasets have different means and SDs. Consequently, the two different datasets need to be aligned to a single FOR, such that, in Rasch terms, the mean orientations of the scale (the zero logit point), and the logit widths are expressed in terms of similar values, as well as being aligned to the LID scale (see Table 1).

With the current datasets, two factors need to be considered (Linacre, 2009: 298) for expert-rated item values to be mapped onto "assessed" item values and for dataset differences to be aligned:

- (1) the differences between the expert-rated item means and the assessed means
- (2) the proportional differences between expert-rated and assessed standard deviations

Regarding (2), Linacre (2009: 299) states that if the proportional differences between the SDs for both the expert-rated and assessed items for both datasets are close to 1, the width of the rated and assessed scales may be taken as identical.

Table 5 is an expansion of Table 4 and includes the results for parameters (1) and (2) above for assessed and rated LTE and CET items.

Table 5: Assessed and rated LTE and CET item differences of common items

1	2	3	4	5	6	7
	Items	N	LID values	Assessed minus Rated	SD	Assessed over Rated
CET assessed	30	2,328	104.28	+8.03	22.43	1.11
CET rated	30	8	96.25		20.22	
LTE assessed	23	4,218	102.96	-1.93	32.10	0.98
LTE rated	23	8	104.89		32.77	

As may be seen from Table 5, both SDs are close to 1.0 (1.11 for the CET and 0.98 for the LTE). Against this backdrop, it is therefore only parameter (1) above which needs to be brought to bear; parameter (2) does not need to be invoked.

Parameter (1) translates into the following item mapping rule (Linacre, 2009: 298):

- mapped CET item mean = assessed CET item mean – rated CET item mean
- mapped LTE item mean = assessed LTE item mean – rated LTE item mean

Mapped fit

With the parameters above set, it is now possible to map levels against the rated and assessed items, and to obtain mapped fits against CEFR levels (or LID values). For ease of reference, Table 1, LID values, in reproduced below.

Table 1: LID scale

CEFR level	LID scale range
A1	51-70
A2	71-90
B1	91-110
B2	111-130
C1	131-150
C2	151-170

CET Items

Table 6 presents the results – as LID scale values – for the CET items. For each item, Column 2 provides test taker scores as LID values, with Column 3 expressing these values as CEFR levels. Column 4 provides the original expert-rated score. Column 5 provides the adjustments to the original scores, adding 8.03 to each. Column 6 expresses the Column 5 scores as CEFR levels. Column 7 presents the difference between Columns 6 and 3 – the rater mapped level against the assessed level.

Table 6: CET items: Candidate assessed values and expert rating values (sorted by CEFR level)

[1]	[2]	[3] (from [2])	[4]	[5] ([4]+8.03)	[6] (from [[5])	7 ([6]-[3])
CET item number	Test taker assessed measure	Test taker assessed score, as CEFR level	Expert rating: original assessment	Expert rating – plus mapped value (+8.03)	Expert rating, as CEFR level	Rater mapped level vs assessed level
C329	62.5	A1	51.3	59.3	A1	=
C341	75.4	A2	64.2	72.2	A2	=
C343	77.8	A2	66.6	74.6	A2	=
C317	78.92	A2	67.72	75.72	A2	=
C334	79.66	A2	68.46	76.46	A2	=
C345	83.61	A2	72.41	80.41	A2	=
C339	96.54	B1	85.34	93.34	B1	=
C326	98.92	B1	87.72	95.72	B1	=
C331	103.44	B1	92.24	100.24	B1	=
C318	104.3	B1	93.1	101.1	B1	=
C344	104.58	B1	93.38	101.38	B1	=
C327	107.71	B1	96.51	104.51	B1	=
C335	108.19	B1	96.99	104.99	B1	=
C322	109.47	B1	98.27	106.27	B1	=
C328	110.8	B2	99.6	107.6	B1	-1
C321	111.33	B2	100.13	108.13	B1	-1
C342	115.88	B2	104.68	112.68	B2	=
C336	116.02	B2	104.82	112.82	B2	=
C320	118.71	B2	107.51	115.51	B2	=
C325	118.99	B2	107.79	115.79	B2	=
C338	120.28	B2	109.08	117.08	B2	=
C337	127.62	B2	116.42	124.42	B2	=
C333	128.64	B2	117.44	125.44	B2	=
C319	132.06	C1	120.86	128.86	B2	-1
C332	136.65	C1	125.45	133.45	C1	=
C330	141.77	C1	130.57	138.57	C1	=
C316	142.23	C1	131.03	139.03	C1	=
C323	144.93	C1	133.73	141.73	C1	=
C340	149.96	C1	138.76	146.76	C1	=
C324	158.32	C2	147.12	155.12	C2	=

Table 7 presents a summary of the CET item fit between rater mapped levels and candidate assessed levels.

Table 7: Fit of rater mapped levels to candidate assessed levels: CET items

Match	(N=30)
Over-rated by one level	0
Exact fit	27 (90.0%)
Under-rated by one level	3 (10.0%)

As can be seen, 27/30 (90%) of the expert ratings of the CET items matched the values which emerged through the calibration of the CET. Three items were under-rated by one level, one level being a degree of discrepancy which is generally taken as acceptable in the marking of examinations (see e.g., Attali and Burstein, 2005; Coniam, 2009).

After recoding rater mapped CEFR levels against assessed CEFR levels as 1-6 (A1=1 through to C2=6), weighted kappa was calculated. With the CET items, a kappa of 0.92 ($p < .001$) emerged – a ‘strong’ agreement between the two variables.

LTE Items

Table 8 now presents the results for the LTE items. For each item, as before, Column 2 provides test taker scores as LID values, expressed as CEFR levels in Column 3. Column 4 provides the original expert-rated score. Column 5 provides the adjustments to the original scores, subtracting -1.93 from each. Column 6 expresses Column 5 scores as CEFR levels. Column 7 presents the difference between Columns 6 and 3 – the rater mapped level against the assessed level.

Table 8: LTE items: Candidate assessed values and expert rating values (sorted by CEFR level)

[1]	[2]	[3] (from [2])	[4]	[5] ([4] - 1.93)	[6] (from [[5])	7 ([6]-[3])
LTE item number	Test taker assessed measure	Candidate assessed score, as CEFR level	Expert rating: original assessment	Expert rating – plus mapped value (-1.93)	Expert rating, as CEFR level	Rater mapped level vs assessed level
L82	69.56	A1	55.75	53.82	A1	=
L75	80.19	A2	66.38	64.45	A1	-1
L76	84.96	A2	71.15	69.22	A1	-1
L74	94.71	B1	80.9	78.97	A2	-1
L77	93.96	B1	80.15	78.22	A2	-1
L80	105.23	B1	91.42	89.49	A2	-1
L81	109.89	B1	96.08	94.15	B1	=
L83	90.06	B1	76.25	74.32	A2	-1
L85	103.21	B1	89.4	87.47	A2	-1
L87	103.45	B1	89.64	87.71	A2	-1
L94	108.82	B1	95.01	93.08	B1	=
L73	115.85	B2	102.04	100.11	B1	-1
L78	110.55	B2	96.74	94.81	B1	-1
L84	110.55	B2	96.74	94.81	B1	-1
L89	111.51	B2	97.7	95.77	B1	-1
L90	115.97	B2	102.16	100.23	B1	-1
L91	124.62	B2	110.81	108.88	B1	-1
L92	114.29	B2	100.48	98.55	B1	-1
L93	116.52	B2	102.71	100.78	B1	-1
L95	114.82	B2	101.01	99.08	B1	-1
L79	146.13	C1	132.32	130.39	C1	=
L86	149.55	C1	135.74	133.81	C1	=
L88	132.33	C1	118.52	116.59	B2	-1

Table 9 presents a summary of the LTE item fit between rater mapped levels and candidate assessed levels.

Table 9: Fit of rater mapped levels to candidate assessed levels: LTE items

Match	(N=23)
Over-rated by one level	-
Exact fit	5 (21.7%)
Under-rated by one level	18 (78.3%)

The expert ratings of the LTE items matched test takers' LTE scores much less closely than they did with the CET items. Only 5/23 (21.7%) of the expert ratings on the LTE items matched test takers' scores on the LTE items. 18 items were under-rated by one level.

With rater mapped and assessed CEFR levels again recoded as before, a weighted kappa of 0.40 ($p < .001$) emerged between the two variables – only a ‘fair’ agreement.

Discussion and Conclusion

In the expert ratings by the eight judges, a much better correspondence was observed between raters judging CET item difficulty than there was with the raters judging LTE item difficulty.

With the CET items, a 90% exact fit was recorded against the 30 CET items. In contrast, with the LTE items, the exact fit was considerably lower, at only 21.7%.

Discrepancies between items in both tests differed, it must be stated, by only one CEFR level: there were no instances of two-level discrepancies. This suggests that raters were generally within acceptable ranges, even if their ratings did not all exhibit an expert, i.e., exact, match. This was particularly the case with the LTE items. The judges were less familiar with the CEFR than they were with the CSE. Their ratings were, nonetheless, within generally tolerable ranges with a discrepancy of one level in six being within the range of acceptability for marking consistency purposes.

The purpose of the current study has been to make a meaningful contribution to the discussion surrounding the viability of expert judgement. The current study has explored expert judgement from two perspectives, namely with a single set of expert judges who rated two sets of items. These experts were very well acquainted with one set of items (the CET items) and the scale (the CSE) against which to assess them. As language teaching and assessment professionals, they were familiar (although less so than with the CSE) with the other set of items (the LTE items) and the scale (the CEFR) against which the LTE items were to be assessed.

The study pursued two research questions.

RQ 1 stated that, with the CET items, a ‘strong’ level of agreement of 0.8 (or 64% ‘shared variance’) would be achieved between assessed and expert-rated values. Of the 30 CET items, a kappa value of 0.92 emerged and 27, or 90.0%, of the items recorded an exact fit between the expert-rater scores and the empirical (‘assessed’) test taker scores.

RQ 2 set out lower expectations: that, with the LTE items, a ‘substantial agreement’ of 0.6 (or 36% ‘shared variance’) would be achieved between assessed and expert-rated values. A Kappa value of 0.40 emerged, with an exact fit only recorded between five, or 21.7%, of the 23 LTE items.

The conclusion that emerges from the current study is that judges who are very familiar with their own assessment situation in terms of test material, test constructs, assessment levels etc, are able to make more accurate assessments than are judges who are less familiar with the material they are assessing, and the levels at which test items should be assessed. While the current results might appear to be somewhat self-evident, the results lend support to the argument that, with adequate training and standardisation, and a strong background in the material to be judged, expert judgement is a methodology that may be reliably utilised in test validation.

As has been mentioned, the current expert judgement study forms part of a larger study, whose overarching purpose involves exploring potential alignment between the CEFR-based LanguageCert tests and the CSE in the context of the reading and usage components. What the current study reveals in the context of experts rating within their own assessment domains is that China experts may be reliably used to rate China CSE-linked test items, and experts who rate (and set) items within the context of the CEFR may be used to rate LanguageCert's CEFR-linked test items (Zhao and Coniam, 2022).

The following chapter, Chapter 8, shows how the results described in this chapter are seen to be significant in light of aligning LanguageCert examination results with the widespread, much used, and respected assessment framework, the Common European Framework of Reference (CEFR) which, is also the assessment framework against which the LanguageCert Academic examination will be aligned, as described in Chapter 1 of this volume.

Limitations

One limitation of the current study is that conclusions have been drawn on the basis of statistical data. Greater validity would be achieved if interviews were conducted with expert raters probing how they perceived standards in the CSE and the CEFR and how they applied these in their ratings.

A second limitation lies within the context of the exploration of the comparability of CSE and CEFR levels. The current study has only involved Chinese raters rating CSE- and CEFR-linked test items. A parallel study currently being planned involves redoing the current study from the opposite perspective: that is, having UK-based expert setters of CEFR-linked items rate the two sets of items used in the current study.

Notes

1. The College English Test (CET) is China's ESL test which examines the English proficiency of undergraduate and postgraduate students in China. It is intended to ensure that Chinese tertiary students reach English language levels specified in the National College English Teaching Syllabuses (see Mini, 2018).
2. 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 level-agnostic qualification is offered in two paper-based versions measuring CEFR levels A1-B1 or A1-C2, and as an adaptive test measuring CEFR levels A1-C2 (see Coniam et al., 2021).

References

- Alderson, J. C., and Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im) possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, 30(4), 535-556.

- Attali, Y., & Burstein, J. (2005). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3).
- Bachman, L., Davidson, F., Ryan, K., & Choi, I-C (1995). *An investigation of the comparability of the two tests of English as a foreign language: the Cambridge-TOEFL comparability study*. Cambridge: Cambridge University Press.
- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7, 303–310.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). *Validating the LanguageCert Test of English scale: The paper-based tests*. London, UK: LanguageCert.
- Council of Europe. (2009). *Relating language examinations to the Common European Framework of References for Languages: Relating Language Examinations to the 'Common European Framework of Reference for Languages: Learning, Teaching, Assessment' (CEFR). A Manual*. Strasbourg: Council of Europe.
- Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 level and the dream of a common standard. *Language Assessment Quarterly*, 15(1), 44-58.
- Falvey, P., Holbrook, J., & Coniam, D. (1994). *Assessing students*. Hong Kong: Longman.
- Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, 66(4), 477–485.
- Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). *Anchor-based methods for judgmentally estimating item difficulty parameters (LSAC Research Report 98-05)*. Newtown, PA: Law School Admission Council.
- Holmes, S. D., Meadows, M., Stockford, I., & He, Q. (2018). Investigating the Comparability of Examination Difficulty Using Comparative Judgement and Rasch Modelling. *International Journal of Testing*, 18(4), 366-391.
- Humphry, S. (2006). *The impact of differential discrimination on vertical equating*. ARC report. Western Australia: Department of Education & Training.
- Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1), 69-81.
- Jin, Y., Wu, Z., Alderson, C., & Song, W. (2017). Developing the China Standards of English: challenges at macro-political and micropolitical levels. *Language Testing in Asia*, 7(1), 1-19.
- Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer.
- Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Linacre, J. M. (2009). *A User's guide to Winsteps-ministep: Rasch-model computer programs*. Program Manual 3.68. O. Chicago, IL.
- Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4), 645–655.
- Liu, X., & Read, J. (2021). Investigating the skills involved in reading test tasks through expert judgement and verbal protocol analysis: Convergence and divergence between the two methods. *Language Assessment Quarterly*, 18(4), 1-25.
- Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

- Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments (Vol. 2, pp. 221-263). Washington DC: National Assessment Governing Board and National Center for Education Statistics.
- Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement*, 49, 467–478.
- Mini, G. (2018). An Introduction to China's College English Test (CET). WENR. World Education News+ Reviews [Elektronnyj resurs]. <https://wenr.wes.org/2018/08/an-introduction-to-chinas-college-english-test-cet>.
- Ministry of Education of the People's Republic of China. (2018). China's Standards of English Language Ability. Beijing: Ministry of Education.
- Ryan, J., & Brockmann, F. (2009). A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory. Washington, DC: Council of Chief State School Officers.
- Shiotsu, T. (2010). Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners. Cambridge: Cambridge University Press.
- Zhao, W., & Coniam, D. (2022). Using self-assessments to investigate comparability of the CEFR and CSE: An exploratory study using the LanguageCert Test of English. *International Journal of TESOL Studies*, 4(1), 169-186.

Appendix 1: CET and LTE Tests: Subtest Breakdown

Section	CET	LTE
1	Cloze: 15 items One cloze passage Assessing grammar, syntax, discourse, vocabulary	Cloze: 15 items Three cloze passages Assessing grammar, syntax, discourse, vocabulary
2	Discrete items: 30 items Assessing grammar, syntax, vocabulary, usage	Discrete items: 23 items Assessing grammar, syntax, vocabulary, usage
3	Reading comprehension: 20 items Four reading comprehension passages, each with 5 items Assessing a range of reading comprehension skills	Reading comprehension: 15 items Three reading comprehension passages, each with 5 items Assessing a range of reading comprehension skills
	65 items	53 items



Chapter 8: Using Self-Assessments to Investigate Comparability of the CEFR and CSE: An Exploratory Study Using the LanguageCert Test of English

Wen Zhao and David Coniam

Abstract

This chapter reports on an exploratory comparability study between the Common European Framework of Reference for Languages (CEFR) and the China Standards of English (CSE). Established equivalences are exhibited via the LanguageCert Test of English of reading and language use for the CEFR and a comparable test of reading and language use produced by a top-tier China university. A large sample of test takers participated in the study, first sitting the two comparable tests of reading and language use, and subsequently completing a number of self-assessment Can-Do statements related to the CEFR and the CSE.

Validity of the dataset was established by linking both tests and sets of self-assessments to a single frame of reference using a third test whose robustness and values had been previously established. While there were some divergences between how the two frameworks aligned – more notably towards the lower ends of the scales – correspondences which emerged between the CEFR and CSE frameworks were broadly in accordance with those reported in other studies referenced in this chapter. The current study therefore sets the groundwork for determining the correspondence between LanguageCert Tests, aligned to both the CEFR, and the CSE.

Keywords: self-assessments, Rasch, reading and language use, comparability

Introduction

The current study is the first step in aligning LanguageCert's different tests – currently aligned to the CEFR – to other key frameworks or assessments, in this case the China Standards of English (CSE) framework. To frame the study, the following section presents detail of studies of Can-Do self-assessment Instruments which have been used to validate approaches to learning and to establish comparability between assessments. Background to the CSE and CEFR is then presented, along with a description of studies which have investigated the correspondence between these frameworks.

Self-Assessment of Language Abilities

Over the past two decades, self-assessment has been shown to be of value in assisting learners to evaluate their own language ability (Bailey, 1998).

The benefits of self-assessment (SA) have been explored in a number of studies and shown to make worthwhile contributions in both learning and assessment. In the context of learning, for example, Butler (2018) illustrated the value of SA in the self-regulated learning process; Babaii et al., (2016) showed how SA aided self-awareness in learning, Dann (2002) showed its value in promoting learner autonomy; and De Saint-Leger (2009) demonstrated how SA was associated with learner confidence and hence performance.

In the area of language assessment, SA has been shown to offer a range of potential benefits. Bachman and Palmer (1996) demonstrated how SA permitted learners to self-assess themselves in an interactive, yet low-anxiety, manner. Oscarson (1989) showed how SA could help expand the range of assessment, emphasizing the fact that assessment should be the responsibility of both learners and teachers. Of relevance to the current study, Liu and Brantmeier (2019) reported a study of young learners in China who were able to quite accurately self-assess their abilities in reading and writing. As outlined below, Peng et al. (2021) explored the alignment of the CSE and the CEFR frameworks, in large part through the use of self-assessment descriptors.

Jones (2014) presented a description and analysis of the large-scale use of 'Can-Do' self-assessment descriptors [Note 2], established in the 1990s, to provide common levels of proficiency across European languages via the ALTE (Association of Language Testers in Europe) Framework. Jones concludes that, despite there being some variation across different educational systems in Europe, students of different languages were, on the whole, reasonably accurate in estimating their relative ability.

The use of instruments such as Can-do statements in self-assessment has been validated in a number of other studies: see e.g., Brown et al., 2014; Summers et al., 2019.

The CSE and the CEFR

For the past two decades, the CEFR has been accepted as illustrating standards of language ability by many stakeholders: policy makers, publishers, exam bodies and test developers (Deygers et al., 2018). Not only in Europe, but in many countries around the world (Little, 2007), the CEFR has become the common currency

for specifying levels of language ability (Figueras, 2012). The CSE reflects an overarching notion of language ability, with which language knowledge and strategies co-function in performing a language activity. Its development attempts to pull together China's various English language curriculums and assessment instruments into one overarching framework.

Jin et al. (2017) describe the development of the "Common Chinese Framework of Reference for English (CCFR-E): Teaching, Learning, Assessment" which began to be developed in 2014. The CCFR-E was finalised in 2018, being released and renamed as the "China Standards of English" (CSE). The CSE has three major levels, each subdivided into three sublevels. Figure 1 illustrates.

Figure 1: CEFR and CSE levels

Common European Framework of Reference		China Standards of English	
Level	CEFR	Level	CSE
Proficient User	C2	Level 9	Advanced stage
	C1	Level 8	
Independent	B2	Level 7	
	B1	Level 6	Intermediate stage
Basic User	A2	Level 5	
	A1	Level 4	
		Level 3	Elementary stage
		Level 2	
		Level 1	

Previous CSE / CEFR Equivalence Studies

Alderson (2017) discusses a range of studies exploring the CSE and its correspondence to the CEFR. This is supported by the discussion by Jin et al. (2017) and by research by Zhao et al. (2017), investigating the linking of College English vocabulary levels with the CEFR. Figure 2 presents a summary of the results of the different studies.

Dunlea et al. (2019) describe a comprehensive study involving all four language skills that explored the relationship between the British Council's Aptis test and IELTS with China's Standards of English Language Ability. The methodology involved expert judgement of items against CSE and CEFR levels and the assignment of CSE descriptors against tasks. Following this, the proposed levels were field tested in an "external evaluation" exercise, where Chinese teachers rated their own students against the proposed matched levels. As Figure 2 below illustrates, CSE L2 appeared to correspond to CEFR A1, CSE L3 to A2, CSE L4 / L5 to CEFR B1, CSE L6 / L7 to CEFR B2, CSE L8 to CEFR C1 and CSE L9 to CEFR C2.

Peng and associates have undertaken a number of studies investigating correspondences between CEFR and CSE levels. Level A0, it should be noted, denotes a level below CEFR A1. These studies are discussed below.

Peng et al. (2021) report on a study attempting to establish level correspondences between CEFR and CSE levels using difficulty estimates of all published descriptors (467 for the CEFR and 1,051 for the CSE) of ratings by English language teachers and students. While there was close correspondence at the top and bottom ends of the scale, there was overlap in the middle levels. Peng et al. (2021) report CSE L1 as corresponding to CEFR A0, CSE L2 to CEFR A1, CSE L2 / L3 to CEFR A2, CSE L4 / L5 to CEFR B1, CSE L6 / L7 to CEFR B2, CSE L7 / L8 to CEFR C1, and CSE L9 to CEFR C2.

In another study, Peng (2021) investigated level alignments between the CSE and CEFR writing descriptors. Results indicated a general correspondence between CSE and CEFR levels. While there was some overlap, CSE L1 / L2 corresponded to CEFR A1, CSE L3 to CEFR A2, CSE L4 / L5 to CEFR B1, CSE L6 to CEFR B2, CSE L7 to CEFR C1, CSE L8 to CEFR C1 / C2, and CSE L9 to CEFR C2.

In a further study, Peng and Liu (2021) attempted to align CSE listening skill levels with those of the CEFR. Results indicated that CSE listening descriptors tended to spread across several adjacent CEFR levels. CSE L1 corresponded to CEFR A1, CSE L2 to CEFR A2, CSE L3 to CEFR A2 / B1, CSE L4 to CFR B1, CSE L5 to CEFR B1 / B2, CSE L6 to CEFR B2 / C2, CSE L7 / L8 to CEFR C1, and CSE L9 to CEFR C2.

Figure 2: CFR/CSE Comparative Mappings from previous studies

Dunlea et al. (2019) All skills		Peng et al. (2021) All skills		Peng (2021) Writing		Peng and Liu (2021) Listening	
CSE	CEFR	CSE	CEFR	CSE	CEFR	CSE	CEFR
L9	C2	L9	C2	L9	C2	L9	C2
L8	C1	L7-L8	C1	L8	C1-C2	L7-L8	C1
L6-L7	B2	L6-L7	B2	L7	C1	L6	B2-C1
L4-L5	B1	L4-L5	B1	L6	B2	L5	B1-B2
L3	A2	L2-L3	A2	L4-L5	B1	L4	B1
L2	A1	L2	A1	L3	A2	L3	A2-B1
L1		L1	A0	L1-L2	A1	L2	A2
						L1	A1
							A0

The different studies outlined in Figure 2 contribute to the level alignment between the CSE and the CEFR. As may be seen, while there is a degree of agreement in the correspondence between the two studies, there are also divergences. These may be due to a number of factors: the samples; the tests; the judges used in the ratings.

Current study: Assessment Instruments, Test Taker Sample, Research Questions

This section briefly outlines the background and make-up of the tests and the self-assessment ratings which test takers completed. The methodology employed in the current study differs from that used in the Dunlea et al. (2019) and Peng et al. (2021) studies. The principal methodology in the latter two involved the use of

expert ratings. In the current study, a large sample of test takers took a live LanguageCert test, which was then calibrated in a single frame of references with their self-assessment ratings.

The use of this ancillary methodology is testament to LanguageCert's researchers' approach to maintaining the high quality of its examinations by using effective methods to enquire into their examinations in order to maintain their high quality.

Test Material

In late 2020, approximately 2,500 Year 1 non-English major college students took a 65-item multiple-choice reading and language use test prepared by experts from the university involved in the current study. Three months later, this same set of students took a 53-item multiple-choice reading and language use test adapted from existing and previously validated LanguageCert Test of English (LTE) material (Coniam et al., 2021). The items in the LTE test used in the study were selected on the basis of representing the spectrum of difficulty across the six CEFR levels.

Item difficulty in LTE tests is predicated on the overarching LanguageCert Item Difficulty (LID) scale; see Table 1. This scale lays out item difficulty levels generally adopted in LanguageCert assessments (Coniam et al., 2021).

Table 1: LID scale

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

For analysis and calibration purposes, 100 has been taken as the mid-point of the scale. To this end, Rasch logit values are rescaled to a mean of 100 and a standard deviation (SD) of 20 (see Coniam et al., 2021).

Appendix 1 provides a comparative analysis of the make-up of the two reading and usage tests. As may be seen, the CET test is slightly longer than the LTE test; also, all CET items are 4-option multiple-choice whereas the LTE items are 3-option multiple-choice. Despite these differences, the content of the two tests, and even the order in which the different sections of the test appeared to test takers, exhibit a great deal of similarity.

Can-Do Self-Assessment Descriptors

Both the CEFR and the CSE contain large arrays, for all skill areas, of Can-Do descriptors (see e.g., <https://www.cultofpedagogy.com/can-do-ell/> for examples of how such descriptors help classroom teachers understand what learners at different levels of proficiency should be able to do).

To reflect the focus of the current study, two sets of Can-Do self-assessment descriptors were assembled for reading and language use for each framework. A set of 22 Can-Do statements, related to the CSE, was compiled by the China university staff who designed the CET test used in the current study. Another set of 16 Can-Do statements related to the CEFR was compiled by members of the LanguageCert research and assessment team. All Can-Do statements were framed as Yes/No questions so that test takers rated themselves dichotomously (i.e., as can / cannot) on each statement. The relevant Can-do statements may be found in Appendices 2 and 3.

The composite set of 38 items were then intermingled. This was intended to forestall respondents trying to guess where their own estimated ability level might terminate.

Test and Self-Assessment Profile Administration

The first test (the CET) was administered in late 2020. In early 2021, the second test (the LTE) was administered. Immediately after the administration of the second test, test takers completed both sets of Can-Do self-assessments. These were all presented bilingually in both English and Chinese.

Self-Assessment Can-Do Statements and Research Questions

Against the backdrop outlined above, the current study pursued two main Research Questions.

RQ1: To what extent can self-assessment Can-Do statements be validly used to establish correspondences between the CEFR and CSE frameworks?

RQ2: To what extent are correspondences between the CEFR and CSE frameworks in line with those reported in previous studies?

Statistical Analysis: Rasch Measurement

The manner for gauging test fitness-for-purpose in the current study, and for linking the data from the two different tests and self-assessments, involves the use of Rasch measurement. An overview of the methodol-

ogy surrounding Rasch can be found in the Glossary, along with an outline of the infit and outfit mean square statistics which are key to the interpretation of Rasch results in the context of data ‘fit’.

Data and Frame of Reference

To recap, there are four sets of assessment data in the current study: the 65-item CET test, the 53-item LTE test, 22 CSE-referenced Can-Do ratings and 16 CEFR-referenced Can-Do ratings. Since all four datasets were collected from the same test takers, the data configuration may be taken as a unified collection, in that all data are referenced to the same candidates and to their English language ability. The person links (Boone, 2016) in the four datasets embrace a coherent frame of reference (FOR), defined by Humphry (2006) as “compris[ing] a class of persons responding to a class of items in a well-defined assessment context.”

In order to calibrate the four datasets in the current study onto the LanguageCert Item Difficulty (LID) scale (see Table 1), a previously calibrated test (henceforth referred to as “Test 3”) from the Coniam et al. (2021) study was incorporated into the data. As a subset of Test 3, the LTE test in the current study provides a set of item links (Boone, 2016). With sets of both person links and item links established, the LTE test could then be linked to Test 3. Following this, the other datasets in the study – the CET test and the two sets of self-assessments – could then be calibrated against Test 3 onto the LID scale. This resulted in all five assessment datasets being included into one single FOR.

Analysing within a Single Frame of Reference

As mentioned, Test 3 was the anchoring frame, having been previously anchored to the LID scale. Against this backdrop, the composite analysis is presented in Figure 3 below.

Figure 3: Composite analysis of three tests and two sets of self-assessments

1	2	3	4	5	6
Test takers	Composite analysis	LTE test	CET test	CSE Can-Dos	CEFR Can-Dos
MEASURE HEADROW = MEAS - TEST1					
170	170	170	170	170	170
140	140	140	140	140	140
130	130	130	130	130	130
120	120	120	120	120	120
110	110	110	110	110	110
100	100	100	100	100	100
90	90	90	90	90	90
80	80	80	80	80	80
70	70	70	70	70	70
60	60	60	60	60	60
50	50	50	50	50	50
40	40	40	40	40	40
30	30	30	30	30	30
20	20	20	20	20	20
10	10	10	10	10	10
0	0	0	0	0	0
Test takers	Composite dataset	LTE test	CET test	CSE Can-Dos	CEFR Can-Dos

In Figure 3, Column 2 contains the analysis of the amalgamated five datasets, i.e., 158 items in the four datasets under investigation plus the PB4 items. Column 3 contains the 53-item LTE test, Column 4 the 65-item CET test, Column 5 the 22 CSE-referenced Can-Do ratings, and Column 6 the 16 CEFR-referenced Can-Do ratings.

To recap, item links in the overall dataset are established between the 53 items in the LTE test and Test 3. Person links are established via the two tests and the two sets of self-assessments. All five datasets may therefore be seen to be within an overall FOR – the composite analysis to the far left of the person-item map in Figure 3. Against the overall picture of calibration, which is centred at 100, the mid-point of B1, it may be seen that the means for the two tests are slightly higher than the overall mean. Tables 2 and 3 present fit and reliability details on the two tests.

Table 2: Summary analysis: 53-item LTE test

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	2281.1	4218.7	105.33	.78	1.06	3.77	1.11	4.23
MAX.	4034.0	4265.0	168.63	1.38	1.23	9.90	1.51	9.90
MIN.	272.0	4137.0	48.18	.65	.93	-4.28	.78	-5.42
MODEL RMSE	.80	TRUE SD	26.19	SEPARATION	32.77	ITEM	RELIABILITY	1.00
S.E. OF ITEM MEAN = 3.63								

Table 3: Summary analysis: 65-item CET test

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	1411.6	2328.2	104.28	1.01	.99	.20	1.00	.46
MAX.	2150.0	2335.0	158.32	1.58	1.18	9.90	1.35	9.90
MIN.	271.0	2293.0	62.50	.86	.87	-9.90	.68	-9.90
MODEL RMSE	1.02	TRUE SD	22.23	SEPARATION	21.82	ITEM	RELIABILITY	1.00
S.E. OF ITEM MEAN = 2.78								

Tables 2 and 3 show that the two tests fit the Rasch model well, with mean infit and outfit figures well within the 0.5 to 1.5 range, and high reliability figures. The means of both tests are very comparable, a quarter of a logit above the overall mean of 100. The LTE test mean was 105.33, and the CET test 104.28.

Tables 4 and 5 now present fit and reliability details for the two sets of self-assessments.

Table 4: Summary analysis: 22 CSE Can-Do statements

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	2484.2	3888.5	95.29	.82	.89	-6.04	.83	-6.74
MAX.	3739.0	3980.0	136.74	1.36	.95	-1.35	.92	-2.55
MIN.	908.0	3823.0	50.61	.69	.82	-9.90	.72	-9.90
MODEL RMSE	.83	TRUE SD	23.90	SEPARATION	28.67	ITEM	RELIABILITY	1.00
S.E. OF ITEM MEAN = 5.22								

Table 5: Summary analysis: 16 CEFR Can-Do statements

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	2389.4	3877.2	95.66	.89	.90	-4.90	.80	-6.26
MAX.	3724.0	3954.0	138.02	1.39	.96	-1.43	.93	-2.21
MIN.	874.0	3836.0	49.67	.69	.84	-9.90	.61	-9.90
MODEL RMSE	.91	TRUE SD	30.67	SEPARATION	33.59	ITEM	RELIABILITY	1.00
S.E. OF ITEM MEAN =	7.92							

From Tables 4 and 5, it can also be seen that the two sets of self-assessments fit the Rasch model; mean infit and outfit figures are within the 0.5 to 1.5 range, and reliability figures are again high. The means of both two sets of self-assessments are again comparable, although this time a quarter of a logit below the overall mean of 100 – both being around 95. This slightly lower score is indicative that, on the self-assessments, test takers have tended to slightly over-rate themselves – a not uncommon phenomenon (Kruger and Dunning, 1999; Dunning et al., 2004).

The difference between the item means of the Can-Do ratings, and the LTE and CET assessment results are within half a logit (10 LID scale points): a difference which is generally accepted within Rasch measurement as being non-significant (Zwick et al., 1999). The conclusion that may be drawn is that test takers can be considered sufficiently objective in their self-assessments to permit tentative correspondences to be drawn between CSE and CEFR levels. The next section explores the correspondences.

Establishing Correspondences between CSE and CEFR Levels

Given that the two sets of self-assessments have been established as valid and broadly comparable, this section presents sets of tables – one at each CEFR level – which incorporate Can-Do statements within corresponding CEFR and CSE levels. Tables are presented one at a time for each CEFR level, in line with LID score ranges for the corresponding CEFR level. The tables are laid out such that the left-hand half of the table includes the detail for the CEFR level: the relevant Can-do statement, the LID value assigned in the current single FOR calibration, and the CEFR level for the Can-do, as laid down in formal documentation. The right-hand half of the table then includes corresponding detail for the CSE level: Can-do statements and their CSE level which fall into the LID value range for the CEFR level.

Table 6 presents the joint analysis for CEFR level C1, for which the LID range is 131-150 scale points.

Table 6: CEFR and CSE Can-Do Statement Level Comparisons: C1 (131-150)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.	138.02	C1			
I can understand specialised articles and longer technical instructions, even when they do not relate to my field.	137.77	C1			
			L7	136.74	I can comprehend academic papers or scientific and technical literature in relevant fields of study and evaluate the research methods.
I can understand long and complex factual and literary texts, appreciating distinctions of style.	133.82	C1			
I can extract necessary information and the points of the argument from articles and reference materials in my specialised field without consulting a dictionary.	130.74	C1			

Within the C1 CEFR LID range of 131-150, four CEFR C1 self-assessments were found, along with one CSE Level 7 self-assessment. The fit would appear to be CEFR C1 --> CSE L7.

Table 7 presents the joint analysis for CEFR level B2, for which the LID range is 111-130 scale points.

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L7	129.72	I can understand linguistically complex novels and materials related to culture and appraise their linguistic features.
			L6	128.73	I can understand the terminology of operational texts in related professional areas.
			L7	127.85	I can understand book reviews in relevant fields of inquiry.
			L6	127.27	I can understand novels and argumentative texts comprised of relatively complex language.
I can scan through rather complex texts, e.g. articles and reports, and can identify key passages.	118.74	B2			
			L5	117.63	I can understand the common figures of speech in stories pertaining to social life written in relatively complex language.
I can understand in detail specifications, instruction manuals, or reports written for my own field of work	116.58	B2			
			L5	116.41	I can infer the content of an entire book or text by scanning the table of contents.
I can read texts dealing with topics of general interest, such as current affairs, without a dictionary, and can understand multiple points of view.	115.69	B2			

Within the B2 CEFR LID range of 111-130, three CEFR C1 self-assessments were found, along with six CSE self-assessments, of which two were at L5, two at L6 and two at L7. The B2 CEFR / CSE fit would appear to be broader, i.e., CEFR B2 --> CSE L5-L7.

Table 8 presents the joint analysis for CEFR level B1, for which the LID range is 91-110.

Table 8: CEFR and CSE Can-Do Statement Level Comparison Chart: B1 (91-110)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L4	95.4	I can extract the key information in practical forms of writing (e.g. memos or notes).
I can understand the plot of longer narratives written in plain English.	95.15	B1			
			L6	94.63	I can infer the author's attitudes with the help of diction or rhetorical devices.
			L6	93.86	I can understand and summarise the main features of the objects in expository writing.
			L5	90.93	I can extract detailed information (e.g. characters, scenic spots) from prose essays.

Within the B1 CEFR LID range, one CEFR B1 self-assessment was found, along with four CSE self-assessments, of which one was at L4, one at L5, and two at L6. The B1 CEFR / CSE fit would therefore also appear to be quite broad, i.e., CEFR B1--> CSE L4-L6.

Table 9 presents the joint analysis for CEFR level A2, for which the LID range is 71-90.

Table 9: CEFR and CSE Can-Do Statement Level Comparison Chart: A2 (71-90)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L4	89.84	I can analyse the authors' viewpoints on familiar social phenomena in short, simple pieces of argumentative writing.
			L5	89.07	I can read arguments on common topics and commentary on familiar topics.
			L5	88.32	I can generalise duly from what has been read while reading.
I can search the internet or reference books, and obtain school- or work-related information, with the help of a dictionary.	87.43	A2			
			L4	86.37	I can discover the key information or details by skimming, scanning, and/or browsing.
I can understand clearly written instructions (e.g. for playing games, for filling in a form, for assembling things).	83.72	A2			
I can understand the main points of English newspaper and magazine articles adapted for educational purposes.	79.90	A2			
			L4	77.65	I can understand details (e.g. time, character, place) in travel notes.
			L4	75.88	I can read short, simple stories, prose essays, and expository writing.
			L3	75.61	I can understand the authors' viewpoints in short, simple letters.

Within the A2 CEFR LID range, three CEFR A2 self-assessments were found, along with seven CSE self-assessments, of which one was at L3, four at L4, and two at L5. The A2 CEFR / CSE fit would therefore appear to be mainly CEFR A2 --> CSE L4-L5.

Table 10 presents the joint analysis for CEFR level A1, for which the LID range is 51-70.

Table 10: CEFR and CSE Can-Do Statement Level Comparison Chart: A1 (51-70)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L3	68.48	I can improve my understanding with reference to key words or topic sentences.
			L3	68.23	I can understand linguistically simple stories.
			L2	67.23	I can pick out the key information in notes or notices.
I can understand the main points of texts dealing with everyday topics (e.g. life, hobbies, sports) and obtain the information I need.	62.61	A1			
I can understand short narratives and biographies written in simple words.	60.80	A1			
I can understand texts of personal interest (e.g. articles about sports, music, travel, etc.) written with simple words.	60.28	A1			
I can understand very short reports of recent events such as text messages from friends' or relatives', describing travel memories, etc.	59.61	A1			

Within the A1 CEFR LID range, four CEFR A1 self-assessments were found, along with three CSE self-assessments, of which one was at L2, and two at L3. The broad A1 CEFR / CSE fit would appear to be CEFR A1 --> CSE L3.

Finally, below CEFR A1, there was one fit between the CEFR and CSE. Table 11 presents this fit.

Table 11: CEFR and CSE Can-Do Statement Level Comparison Chart: A0 (below 51)

CEFR			CSE		
CEFR Can-Do Statements	LID value	CEFR level	CSE level	LID value	CSE Can-Do Statements
			L1	50.61	I can understand short, linguistically simple articles on daily life.
I can understand very short, simple, everyday texts, such as simple posters and invitation cards.	49.67	A0			

In this mapping, low A1 (“A0”) fitted with CSE L1.

From the above set of tables showing the comparative fit of the CEFR and CSE levels, it is now possible to produce an overall tentative mapping of how the CEFR scale, as represented by the LTE, may be mapped against the CSE. Table 12 presents the match. It should be noted that there was insufficient data to calibrate CEFR level C2.

Table 12: CEFR / CSE fit in LTE study

CEFR	China CSE
A0	L1
A1	L2-L3
A2	L3-L5
B1	L4-L6
B2	L5-L7
C1	L7
C2	N/A

As can be seen from Table 12, as might perhaps be expected, while there is not a one-to-one match between the levels in the two frameworks, as one moves up the scale, there is a graduated fit between the CEFR and the CSE.

Figure 4 below, presents a reworking of Figure 2, which includes the alignments proposed in the Dunlea et al. (2019) [henceforth the ‘Dunlea’ study] and Peng and associates’ (2021) studies [henceforth the ‘Peng’ studies], together with the alignments as they have emerged empirically in the current study.

Figure 4: Formal CEFR / CSE Mapping

LC Mapping		Dunlea et al. (2019)		Peng et al. (2021)		Peng (2021)		Peng and Liu (2021)	
Reading & Language Use		All skills		All skills		Writing		Listening	
CSE	CEFR	CSE	CEFR	CSE	CEFR	CSE	CEFR	CSE	CEFR
	C2	L9	C2	L9	C2	L9	C2	L9	C2
L7	C1	L8	C1	L7-L8	C1	L8	C1-C2	L7-	C1
L5-L7	B2	L6-L7	B2	L6-L7	B2	L7	C1	L8	
L4-L6	B1	L4-L5	B1	L4-L5	B1	L6	B2	L6	B2-C1
L3-L5	A2	L3	A2	L2-L3	A2	L4-	B1	L5	B1-B2
L2-L3	A1	L2	A1	L2	A1	L5		L4	B1
L1	A0	L1		L1	A0	L3	A2	L3	A2-B1
						L1-	A1	L2	A2
						L2		L1	A1
									A0

The different mappings show both similarities and differences.

The current study mapped A0 onto L1, as did the Peng (all skills) study.

The current study mapped A1 against L2 / L3. Dunlea et al. mapped A1 to L2, and the Peng studies mapped A1 to L1 / L2.

The current study mapped A2 more broadly against L3-L5. The Dunlea study mapped A2 to L3 while the Peng studies mapped A2 against L2 / L3.

The current study mapped B1 against L4 / L6. The Dunlea and Peng studies mapped B1 against L4 / L5.

The current study mapped B2 against the bottom end of L5 to L7. The Dunlea study mapped B2 against L6 / L7 and the Peng studies mapped B2 against L5 / L7.

The current study mapped C1 at L7, whereas in the Dunlea study C1 mapped at L8 and at L7 / L8 in Peng’s studies.

There was no data for C2 in the current study.

The results of the current study can be seen to echo the mappings of the previous studies, although the mappings which have emerged suggest a slightly more lenient fit than that reported in other studies (see below) – as for example with CEFR C1 being located against CSE L7 in the current study as against CSE L7 / L8 in the Peng studies and CSE L8 by Dunlea. This is mirrored at the lower end of the scale, where the current study does not suggest direct one to one matches. There are a number of possible reasons for these divergences. A key difference is that the current study empirically matched levels against performance, as opposed to an expert-rater-focused methodology. Another reason may be attributed to the fact that only one skill – essentially

reading – has been explored in this chapter, whereas the other studies examined all four skills. A third is that the sample was limited at the top end of the ability spectrum to C1-level test takers.

Conclusion

The current study pursued two Research Questions.

The first research question was whether self-assessment Can-Do statements may be validly used to establish correspondences between the CEFR and CSE frameworks. As was illustrated, from a comprehensive analysis of both test and Can-Do self-assessment responses, respondents tended to slightly over-estimate their abilities on both the CEFR and the CSE. These over-estimations were minimal, however, in that mean values were only a quarter of a logit higher than might have been expected. Secondly, the over-estimations were consistent with the scales for both frameworks.

The second research question was that correspondences between the CEFR and CSE frameworks would be broadly in accordance with those proposed by previous studies. While there have been some divergences, more notably towards the lower end of the scales, the correspondences proposed in the current study broadly echo those reported in previous studies.

A range of correspondences may well be expected from different studies, exploring different assessment instruments. Difficulties in accurate alignment have been commented on by other researchers: Papageorgiou et al., 2015; North & Piccardo, 2018. Peng insightfully comments that “the CSE is a local standard with granular levels reflecting Chinese learners’ requirements and progress [....] while the CEFR is a framework for reference with broad bands of proficiency and is intended to be adapted or further developed for specific contexts and uses”. In the current study, the assessment context has focused on reading and language use, whereas the Dunlea et al. (2019) and the Peng et al. (2021) studies examined all four language skills, as well as writing and listening, which Peng (2021) and Peng and Liu (2021) respectively explored.

From a wider, and methodological, perspective, the use in the current study of a single frame of reference to calibrate self-assessment ratings directly against performance adds to the armoury of tools available to assessment professionals in linking exercises such as those between two different tests, or by providing a larger perspective between two different assessment frameworks. The use of innovative methods is characteristic of LanguageCert’s aim to keep their research fresh and relevant in their determination to maintain the quality of their examinations.

The approach adopted in the current study may be useful for other assessment situations, where Can-Do ratings may be incorporated at the end of an assessment session. This may even be done in a user-friendly manner where individual candidates rate subsets of Can-Do ratings, which are then linked via common items to cover a range of Can-Do aspects. Can-Do ratings also provide further opportunities for the validation of examinations because they allow for triangulation of data and, therefore, more robust findings.

Limitations

A limitation of the current study was that the investigation of test types was limited to reading and language use. Future studies will broaden this by extending the investigations conducted in the current study to other language skills.

Notes

1. The 211 and 985 projects were initiatives undertaken by China in the 1990s to develop world-class universities in China. The current top-tier university is a “211/985” university.
2. The CEFR framework comprises descriptors laying out what a student can do as a particular skill when they have completed a given level. A descriptor for Reading at A2, for example, is: “I can understand short narratives and biographies written in simple words.”

References

- Alderson, J. C. (2017). Foreword to the special issue “The common European framework of reference for languages (CEFR) for English language assessment in China” of language testing in Asia. *Language Testing in Asia*, 7, 20.
- Babaii, E., Taghaddomi, S., Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners’ and teachers’ criteria. *Language Testing*, 33(3), 411–437.
- Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions and directions*. New York: Heinle & Heinle.
- Boone, W. J. (2016). Rasch analysis for instrument development: why, when, and how?. *CBE—Life Sciences Education*, 15(4).
- Brown, G. T., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, 22(4), 444–457.
- Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals*, 47(2), 261–285.
- Butler, Y. G. (2018). The role of context in young learners’ processes for responding to self-assessment items. *The Modern Language Journal*, 102(1), 242–261.
- Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. Brigham Young University Testing Services and the Department of Instructional Science. <http://testing.byu.edu/info/handbooks/betteritems.pdf>.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). *Validating the LanguageCert Test of English scale: The paper-based tests*. London, UK: LanguageCert.
- Council of Europe. (2009). *Relating language examinations to the Common European Framework of References for Languages: Learning teaching, assessment*. Strasbourg: Language Policy Division. Educational Testing Service. (2012). *The official guide to the TOEFL test (4th Ed.)*. New York: McGraw-Hill.

- Dann, R. (2002). *Promoting assessment as learning: Improving the learning process*. Routledge.
- De Saint-Léger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals*, 42(1), 158-178.
- Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 level and the dream of a common standard. *Language Assessment Quarterly*, 15(1), 44-58.
- Dunlea, J., Spiby, R., Wu, S., Zhang, J., & Cheng, M. (2019). China's standards of English language ability: Linking UK exams to the CSE. https://www.britishcouncil.org/sites/default/files/linking_cse_to_uk_exams_5_0.pdf.
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5, 69-106.
- Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.
- Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, 66(4), 477-485.
- Gu, M. (2018). An Introduction to China's College English Test (CET). WENR. *World Education News+ Reviews*, August 2018. <https://wenr.wes.org/2018/08/an-introduction-to-chinas-college-english-test-cet>.
- Humphry, S. (2006). The impact of differential discrimination on vertical equating. ARC report.
- Jin, Y., Wu, Z., Alderson, C., & Song, W. (2017). Developing the China Standards of English: challenges at macro-political and micropolitical levels. *Language Testing in Asia*, 7(1), 1-19.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6), 1121.
- Lee, T., Milanovic, M., Coniam, D., & Pike, N. (2021). Externally-referenced anchoring: equating expert judgement and Rasch measurement values in LanguageCert IESOL English language tests. London, UK: LanguageCert.
- Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4), 645-655.
- Liu, H., & Brantmeier, C. (2019). "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System*, 80, 60-72.
- Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession*, 13, 425-444.
- Ministry of Education of the People's Republic of China. (2018). *China's Standards of English Language Ability*. Beijing: Ministry of Education.
- North, B., & Piccardo, E. (2018). *Aligning the Canadian Language Benchmarks (CLB) to the Common European Framework of References (CERF)*.
- Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels. Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.
- Peng, C. (2021). Aligning the CSE with the CEFR: Level alignment in writing ability. *Foreign Language World*, 5, 84-93.
- Peng, C., & Liu, J. (2021). The listening skill level alignment of the CSE with the CEFR. *Foreign Language Educator*, 5, 43-50.
- Peng, C., Liu, J., & Cai, H. (2021). Aligning China's Standards of English Language Ability with the Common European Framework of Reference for Languages. *The Asia-Pacific Education Researcher*, 1-11.

- Summers, M. M., Cox, T. L., McMurry, B. L., & Dewey, D. P. (2019). Investigating the use of the ACTFL can-do statements in a self-assessment for student placement in an Intensive English Program. *System*, 80, 269-287.
- Wright, B. D. (1997). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 97-116.
- Zhao, W., Wang, B., Coniam, D., & Xie, B. (2017). Calibrating the CEFR against the China Standards of English for College English vocabulary education in China. *Language Testing in Asia*, 7(1), 1-18.
- 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 9: Online Invigilation of English Language Examinations: A Survey of Past China Candidates' Attitudes and Perceptions

David Coniam

Abstract

Drawing on a previous large-scale study examining the reactions of past candidates to the use of online invigilation – or online 'proctoring' (OLP) – in the delivery of high-stakes English language examinations (Coniam et al., 2021), this chapter reports the responses of the subset of China candidates in the sample. China is a rapidly-expanding market for LanguageCert's English language examinations. It is therefore instructive to gauge the market and its candidates' needs, as we move forward, by analysing the responses of past China candidates to OLP as a mode of sitting an exam.

The chapter first sets the scene in terms of the initially gradual, and then accelerated 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.

Detail is then presented from the survey exploring the reactions to and perceptions of OLP by the China respondents (N=64), comparing this sample with the larger world-wide sample, all of whom had taken an English language examination via OLP. A strong endorsement by the China cohort 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. The results are indicative of a broad acceptance of OLP, pointing to strong future uptake of the OLP mode of test delivery.

This is an encouraging finding in the aftermath of the initial growth of online proctoring that increased exponentially during the Covid pandemic. Initial resistance to OLP has decreased in changes to life, lifestyle, approaches to working, and now assessment, brought about by the need to innovate and make changes to typical practices pre-pandemic. In countries without access to highly-trained assessors, the discipline and security engendered by well-administered, well-thought-out, and highly developed OLP can only benefit candidates and other stakeholders in the search for secure modes of assessment.

Keywords: English language examinations, survey, Chinese candidates, attitudes and perceptions

Background

Online Delivery of Learning and Teaching

The generally accepted mode for 'delivery' of both teaching and assessment has long been that of a teacher providing input to a class of students from the front (see Wiesenbergs and Stacey, 2008).

Views of how education is delivered are changing, however, along with the uptake and acceptance of technology across all facets of society (Lim and Wang, 2016), augmented and accentuated by the COVID-19 pandemic (TPD@Scale for the Global South, 2020). The traditional mode of delivery is consequently being rethought along with the use of more innovative and interactional methods. Todd (2020), for example, describes how COVID-19 was a strong mover in the adoption of online teaching.

Over the past decade, developments in technology have permitted greater uptake of 'blended' learning and teaching (Lim and Wang, 2016). There has been a global move towards experimenting with, if not embracing, various types of synchronous and asynchronous modes of teaching (Lim and Graham, 2021).

While the expanded use of technology has brought about a change in mindset in terms of the delivery of teaching content using online instruction and facilitation, it is nonetheless the general expectation that examinations will still broadly occur in a face-to-face situation (Ahlawat et al., 2014). While there has been some take-up of technology in the area of assessment, this has been less than it has been with teaching. It is still generally accepted that assessment – high-stakes assessment in particular – should be conducted in a pen-and-paper medium, in front of an examiner/invigilator in a centre such as a school hall. Upon completion, candidates' test papers are collected, then passed on to markers, with a considerable time lapse before results are available. Table 1 outlines different test administration possibilities.

Table 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

Non-traditional forms of test administration might involve, for example, candidates taking oral examinations remotely by an examiner, as well as candidates completing pen-and-paper (or computer-based) tests at their own home under the invigilation of an examiner in another location (via video link). A final permutation is candidates sitting examinations in an unsupervised mode, as in take-home exams (Bengtsson, 2019), or unsupervised computer adaptive tests (Thompson, 2017).

As mentioned, while online learning technologies have more recently come to be accepted for the competent delivery of learning and teaching, the delivery of assessment via online mode has been beset with problems and challenges (Hussein et al., 2020). A key issue revolves around academic integrity: as, for example, when examinations are taken remotely, and even more so when the location is a candidate's home.

Online Delivery of Assessment

A number of issues – both positive and negative – need to be examined in the context of the online delivery of assessment.

As mentioned, the COVID-19 pandemic has forced a major rethink of how education is delivered, with many educational institutions moving quite rapidly to partial or even total online delivery of classes (Gardner, 2020). In a discussion of technology-enhanced assessment (TEA) during the COVID-19 pandemic in Pakistan, Khan and Jawaid (2020) note how the three areas of teaching, learning, and assessment need to be equally embraced in terms of access and delivery, with an emphasis on changes in attitudes to the online delivery of assessment being key.

Clark et al. (2020) comment on the 'fit' of instructional practices within a course in terms of online vs face to face teaching and assessment. They comment on the issue of 'continuity'; that where a course is intended at the outset to be a distance learning one, then online assessment should naturally fall into place. While many teachers managed to adapt their teaching practices reasonably well in online modes during the COVID-19 pandemic (see Chaturvedi et al., 2021), assessment was problematic, with traditional practices prevailing. Mismatches between intended course outcomes and assessment conducted online tended to be greater than was the case with traditional paper-based assessments which were more aligned with intended course outcomes (see Gil-Jaurena and Softic, 2016). A suggestion by Clark et al. (2020) is for classes to begin in a blended/distance format. Such a format, they argue, permits intended course outcome/assessment gaps to be narrowed, with students becoming acclimatised to an online environment, more able to see a fit between online assessment and course content, and hence more prepared for taking online exams.

Comparability of Results from Exams Taken in Online / 'Standard' Model

In a study of groups of students given online non-proctored exams, Ardid et al. (2015) reported such participants scoring higher than those sitting online-proctored exams. This disparity, they suggest, raises concerns about security and honesty in terms of how non-proctored assessments can be conducted satisfactorily.

Differences between online proctored exams and proctored paper and pencil exams have been investigated by, for example, Alexander et al. (2001). In the context of a computer technology course, they found no significant differences in student performance on proctored online exams and proctored paper and pencil exams.

Benefits and Drawbacks

On the positive side, candidates may take an online-proctored exam in the comfort (and safety) of their own home – an important factor in times of a pandemic where movements are restricted or for those with disabilities. Convenience and speed are another factor to be considered; an exam may be delivered via computer, and results may therefore be obtained more rapidly.

It should also be borne in mind that many typical exams – especially high-stakes school and university exams – involve candidates sitting in halls and writing by hand for two to three hours. Since most assignments written over the course of an academic year will have involved multiple drafts on a word processor, it may well be argued that the traditional mode of administering exams compromises validity because traditional examination conditions do not reflect real life (Mogey et al. 2012). Writing an exam using a word processor on a locked-down computer may be viewed as a more valid mode in which to complete an examination.

As mentioned, one major concern revolves around 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, and getting students to set their own goals, and plan, monitor and control their cognition (Boekaerts and Corno, 2005). In contrast, expectations of assessment (and in particular high-stakes assessment) are that this will be the work of one student, or one candidate, working on their own, with no external support. In line with traditional views of comparability (and hence reliability), this therefore means that the same assessment should be delivered to all candidates at the same time. Such a requirement involves issues of security, honesty and fairness, all of which leads to concerns that some candidates gain an 'advantage' over others or of different aspects of malpractice take place.

Security

A major issue of discussion in the context of the online delivery of examinations has centred on security for different types of online examinations. Foster and Layman (2013) state that online (i.e., human) invigilation

should emphasise the “critical use of the Internet and automated processes to produce a secure solution in monitoring test takers” and provide a thorough analysis of security in online invigilation.

Foster presents an extensive list of key security features – a useful overview by which high-stakes assessment may be viewed. Features that he lists (see Table 2) range from the management and training of the invigilator (the “proctor”), to interaction with the candidate, to the stability of the Internet, to data transfer encryption.

Table 2: Key security features in OLP examinations (after Foster and Layman, 2013)

Features	
1.	Online proctor during exam
2.	Continuous Internet
3.	Encryption for data transfer
4.	Schedule availability
5.	Proctor management
6.	Interaction with test takers
7.	Prevent proctor view of screen
8.	Later video review proctoring
9.	Later video review capable
10.	Control during test session
11.	Automated proctoring
12.	Lockdown
13.	Authentication
14.	Webcam
15.	Logs/records
16.	Program customisation
17.	Effectiveness research

Foster and Layman (2013) describe how systems can provide levels of security which make online proctoring of examinations viable. They comment 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 the monitoring of online examinations than in traditional face-to-face settings (Rose, 2009; Watson and Sottile, 2008).

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 (Wright and Kelly, 1974; Bushway and Nash, 1977; Sierles and Hendrickx, 1980). Over forty years ago, the Carnegie Council Report (1979) made reference to a growing “ethical deterioration” in academic life with respect to the number of college students cheating to get their desired grades.

It is, however, with the Internet, and with access to digital documents and to networks of people willing to facilitate paid cheating, that issues of cheating have been highlighted over the past decade (Berkey and Halfond, 2015; Harper et al., 2021). Cheating in online examinations is becoming more prevalent, and has been explored in numerous studies, (Harmon and Lambrinos, 2008; Grijalva et al., 2006; Watson and Sottile, 2008).

There has been considerable research focusing on the “vulnerability” of online tests, how online tests may be made more secure and cheating might be prevented – see Corrigan-Gibbs et al. (2015) for an extensive discussion regarding cheating and academic dishonesty. Nonetheless, cheating should not be seen purely as an issue related to online tests. As mentioned, cheating has always taken place with traditional examinations. Indeed, in order to counteract cheating, it has been argued (see, e.g., Rose, 2009; Watson and Sottile, 2008) that with adequate protocols in place, online tests may be as secure, if not more secure, than traditional face-to-face tests.

Data

The data in the larger worldwide study (WS) into online proctoring (see Coniam et al., 2021) involves a survey administered to past candidates of LanguageCert's International ESOL suite of English language tests aligned to the CEFR levels, A1 – C2. While there are six tests in the IESOL suite, due to language constraints, examinations in OLP mode are only available for candidates at B1 level and above.

The Survey

Following extensive development and trialling, the survey was administered via the Internet in early 2021. 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.

All items were presented on a 6-point scale, to avoid choosing a mid-point, with '1' indicating a negative response or disagreement, and '6' a positive response or agreement.

In line with data protection legislation, only candidates who had previously agreed to being contacted were approached regarding participation in the survey. The survey was responded to by 920 of the 2,917 who opened the link. The response rate of 31.5% quite closely approximates the average responses reported by Nulty (2008) and may therefore be considered acceptable.

The analysis of the ten attitudinal items on the survey via Cronbach's alpha returned a figure of 0.89. Since 0.8 is generally recommended as desirable in a questionnaire (e.g., Tavakol and Dennick, 2011) the construction of the survey may be seen to be acceptable.

The number of respondents indicating Chinese to be their mother tongue was 64, 7% of the total cohort. Such a sample size is adequate for inferential analysis, with a sample size of 30 being taken as the threshold for conducting statistical analysis (Ramsey, 1980).

Two research questions were pursued in the current study.

RQ 1 investigated whether Attitudes and opinions to OLP on the survey will indicate a positive uptake and acceptance of OLP. This will be measured by item means being above 4.5 out of 6.

RQ 2 investigated whether responses to items will show no effect related to background demographics. This will be measured by no significance emerging on chi-square tests against the demographic variables.

Data Analysis

In the analyses discussed below – unless specified otherwise as being related to the worldwide whole group (WG) (see Coniam et al., 2021) – results and discussion directly relate to the analysis of the China cohort data. Where possible, responses for the Chinese mother tongue cohort are matched against respondents in the whole group and 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 IESOL B1-C2 test candidates. The IESOL test registration form asks candidates for details of gender, age, and mother tongue. Since not all candidates supply these details, there is, consequently, a degree of missing data in the IESOL whole test figures. In the survey, however all respondents provided this demographic data. Table 3 presents a comparison of candidate demographics of both cohorts: the China survey cohort, and all 15,000+ IESOL test B1-C2 candidates for the period 2017-2021

Table 3: Demographics (as percentages)

	Survey: China cohort [N=64]	Whole IESOL cohort
Test level		
B1	4.0%	20.7%
B2	50.0%	38.3%
C1	38.0%	26.6%
C2	8.0%	14.3%
Gender		
Female	68.0%	53.5%
Male	32.0%	38.5%
Age		
<21	16.0%	35.1%
21-30	66.0%	36.0%
31-40	14.0%	16.6%
41-50	4.0%	7.9%
>50	-	4.4%

As can be seen from Table 3, in terms of the distribution of China candidates by test level, the picture was broadly comparable with the typical IESOL profile. There were fewer candidates at B1, although this is 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. Concerning age, the whole test cohort showed a skew towards the younger age bracket. This skew was even more accentuated in the China cohort, especially in the 21–30-year range.

Attitudinal Items

To highlight key differences, in the current study, where a '6' indicated a positive and '1' a negative response, "strong positive responses" (see Coniam, 2013) are defined as those above '4.5'. Table 4 elaborates.

Table 4: Survey item and means: China cohort

Survey item	Means (out of 6)
12. How anxious were you before the OLP test?	3.3
06. Assessment of personal computer literacy	4.9
'institutional' items	
14. How straightforward was the OLP setup process?	4.7
15. How was the online connection with the interlocutor?	4.2
16. How was the interaction with the interlocutor?	5.0
'personal' items	
17. How was the overall OLP experience?	4.8
18. Preference for tests by traditional means (1) or tests by OLP (6)?	4.5
19. "Taking tests by OLP is a more personal experience"	4.5
20. "Taking tests by OLP is more efficient"	4.7
21. Your score: better on traditional (1) or OLP (6) tests?	4.5

The general pattern of responses of the China group generally mirrors those of the whole group (WG) – see Coniam et al. (2021) for a discussion and analysis. To avoid confusion, the discussion below only reports the results produced from the responses of the China cohort.

Item 12, test anxiety, had the lowest mean score, just below the mid-point of 3.5. This is perhaps unsurprising, given that for many candidates, this was the first time they had taken an examination via OLP.

Item 06 – 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.

Despite the anxiety experienced by many of the China cohort, responses to the attitudinal items were all very positive – 4.5 being the benchmark for strong endorsement. The positive nature of the responses may be seen by the fact that the majority of the 'institutional' items have means in the high 4's or above 5. For the majority of the China 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. Online connection was the only item (apart from anxiety) where the China cohort mean did not reach 4.5.

Responses to the 'personal' items were also, on the whole, very positive, with all items being rated above 4.5. Respondents showed a clear preference for taking tests by OLP as opposed to traditional means. To what extent, preference for OLP is a sign of the times, or was solely the result of the COVID-19 pandemic, will only be revealed by future research after the pandemic is classed as being over. Looking ahead, on the issue of preference for tests via OLP (6/6) or by traditional means (1/6), a mean of 4.5/6 emerged, indicative of very positive

acceptance of OLP and strong future uptake of OLP as a means of test delivery. One noteworthy finding was that the China cohort respondents.

In addition to the responses of the China cohort being quite consistent with those of the whole group, no incidences of significance emerged in chi square analyses.

Conclusion

This chapter has explored reactions to and perceptions of OLP by China candidates who had taken an English language examination via online proctoring. Of 920 respondents to a survey sent out to all past candidates of LanguageCert IESOL examinations, 64 (7%) were from China, and it is their responses which have been analysed in the current chapter. LanguageCert is experiencing rapid expansion in take-up of its examinations in China. It is hence both instructive and encouraging to see the reactions of China candidates who have taken examinations via OLP, in order that major issues may be identified and where possible addressed.

Demographically, the China cohort was broadly comparable to the cohorts who have taken LanguageCert examinations over the two-year period of data collection. There were more females than males. There were more candidates at levels B2 and C1 in China and in terms of age there was a skew towards the younger end of the spectrum in China, especially in the 21-30-year age range.

The first RQ investigated whether the attitudes and opinions towards OLP would indicate a positive uptake and acceptance of OLP. Virtually all item means were above 4.5 / 6. This reaction is very encouraging and becomes a springboard for encouraging researchers to delve even further into these issues so that stakeholders' concerns can be addressed and, hopefully, alleviated.

The second RQ investigated whether any significance would emerge in inferential analysis of the items against background demographic variables. The lack of significance against the background variables indicates that respondents were in agreement with items irrespective of gender, age, grade obtained, the level of exam taken, or test skill – Speaking or Listening, Reading, Writing, test – had been taken. This finding too is encouraging. It means that factors such as those listed above, that might skew results were found not to be significant.

Pre-exam anxiety was the only item which had a mean score below the mid-point of 3.5, although this effect was identical with the response of the whole group.

China respondents assessed their own personal computer literacy quite highly, in line with the whole group. High computer literacy was possibly a reason why China respondents were positive about setting up and interacting with the interlocutor. Online connection was the item with the lowest mean, although at 4.2/6, this was still positive.

Regarding preference for exams by traditional means or via OLP, a strong endorsement of OLP was recorded by the China cohort. Respondents felt that OLP was a more personal and efficient way of taking an exam – possibly an effect of exam 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 examination 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.
- Ahlawat, V., Pareek, A. & Singh, S. K. (2014). Online invigilation: A holistic approach: Process for automated online invigilation. *International Journal of Computer Applications*, 90(17), 31– 35.
- 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 Board of Higher Education, July 20, 2015.
- Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology*, 54(2), 199-231.
- Bushway, A., & Nash, W. R. (1977). School cheating behavior. *Review of Educational Research*, 47(4), 623-632.
- 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.
- Chaturvedi, S., Purohit, S., & Verma, M. (2021). Effective Teaching Practices for Success During COVID 19 Pandemic: Towards Phygital Learning. *Frontiers in Education*, 6, 646557, 1-10. doi: 10.3389/educ.
- 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.
- Coniam, D. (2013). The increasing acceptance of onscreen marking–The ‘tablet computer’ effect. *Journal of Educational Technology & Society*, 16(3), 119-129.
- Coniam, D., Lampropoulou, L., & Cheilari, A. (2021). Online proctoring of high-stakes examinations: A survey of past candidates’ attitudes and perceptions. *English Language Teaching*, 14(8), 58-72.
- Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E., & Thies, W. (2015). Deterring cheating in online environments. *ACM Transactions on Computer-Human Interaction*, 22(6), 1-23.
- 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(5).
- Gil-Jaurena, I., & Softic, S. K. (2016). Aligning learning outcomes and assessment methods: A web tool for e-learning courses. *International Journal of Educational Technology in Higher Education*, 13(1), 1-16.
- Grijalva, T. C., Kerkvliet, J., & Nowell, C. (2006). Academic honesty and online courses. *College Student Journal*, 40(1).

- Harmon, O. R., & Lambrinos, J. (2008). Are online exams an invitation to cheat? *The Journal of Economic Education*, 39(2), 116-125.
- Harper, R., Bretag, T., & Rundle, K. (2021). Detecting contract cheating: examining the role of assessment type. *Higher Education Research & Development*, 40(2), 263-278.
- 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., & Jawaid, M. (2020). Technology enhanced assessment (TEA) in COVID 19 pandemic. *Pakistan Journal of Medical Sciences*, 36(19), 108-110.
- Lim, C. P., & Graham, C. R. (Eds.). (2021). *Blended Learning for Inclusive and Quality Higher Education in Asia*. Singapore: Springer Nature.
- 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.
- 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.
- 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.
- 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.
- Rose, C. (2009). Virtual proctoring in distance education: An open-source solution. *American Journal of Business Education*, 2(2), 81-88.
- Sierles, F., & Hendrickx, I. (1980). Cheating in medical school. *Academic Medicine*, 55(2), 124-5.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55.
- Thompson, G. (2017). Computer adaptive testing, big data and algorithmic approaches to education. *British Journal of Sociology of Education*, 38(6), 827-840. DOI: 10.1080/01425692.2016.1158640
- 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>.
- 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).
- Wiesenberg, F. P., & Stacey, E. (2008). Teaching philosophy: Moving from face-to-face to online classrooms. *Canadian Journal of University Continuing Education*, 34(1), 63-79.
- Wright, J. C., & Kelly, R. (1974). Cheating: Student/faculty views and responsibilities. *Improving College and University Teaching*, 22(1), 31-34.

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
##09. How many Speaking Tests have you taken by OLP?	1 2 3 >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 10: The Delivery of Speaking Tests in Traditional or Online Proctored Mode: A Comparability Study

Michael Milanovic, Tony Lee and David Coniam

Abstract

This chapter reports on a study comparing test scores recorded for high-stakes English language Speaking Tests administered face-to-face in either a traditional centre-based mode (TM) or in an online proctored mode (OLP). The purpose of the investigation is to determine whether different modes (TM/OLP) produce different scores.

The data comprise a large sample of test takers taking English language Speaking Tests at four CEFR (the 'Common European Framework of Reference for Languages') levels – B1 to C2 – via TM or OLP. The data were analysed using descriptive statistics, effect size differences and equivalence tests. While a degree of difference in scores obtained between modes was apparent at C2 level, the differences were not found to be statistically significant.

The chapter concludes that whether Speaking Tests are delivered in online proctored mode or in traditional face-to-face mode, test takers receive similar scores. The study confirms the claim that mode of test delivery does not significantly affect test taker scores. The findings of the study described in this chapter echo those described in Chapter 8 where a cohort of Chinese candidates were investigated within the context of OLP.

Keywords: test score comparability, English language, Speaking tests, CEFR, online proctoring

Introduction

Since the late 2010s, and more recently due in considerable part to the Covid-19 pandemic, many examinations have moved from face-to-face to online delivery. The current study was conducted in order to determine the extent to which mode of delivery might affect performance and in turn, therefore, affect Speaking test scores. Focusing on English language Speaking tests at CEFR levels B2 to C2, this chapter examines the comparability of scores achieved by test takers taking examinations administered in traditional face-to-face mode (TM) with those administered by online proctored mode (OLP).

The chapter first reviews different approaches to the increasingly-common online delivery of learning and teaching. This is followed by a review of the less common online delivery of examinations. A brief consideration of the assessment of speaking and the challenges of conducting communicative speaking tests is then provided. The chapter then examines studies which have compared the two modes of delivery.

Following the background section, data of a large sample of test takers taking English language Speaking Tests at CEFR levels B1 to C2 via TM and OLP is then presented and analysed for statistical difference.

Background

This section presents a background to the online delivery of learning and teaching, especially in the face of the Covid-19 pandemic. Issues in the delivery of online assessment – the benefits and drawbacks to taking tests in OLP mode – are then examined. A brief exploration of the assessment of speaking, and the particularly difficult challenges associated with assessing spoken communicative skills is provided. This is followed by a discussion of the increasingly vexed issue of the online assessment of speaking.

Online Delivery of Teaching and Assessment

In the face of the Covid pandemic, the common practice of learning and teaching being conducted by a teacher at the front of an actual class has undergone immense and rapid change (Hodges et al., 2020). Augmented by developments in technology, the acceptance of online learning has grown exponentially over the past two years (Lim and Wang, 2016), with the ‘traditional’ mode of delivery being rethought (Hodges et al., 2020). Todd (2020), for example, outlines how Covid was a strong mover in the adoption of online teaching.

Nonetheless, while the mindset has changed in terms of teaching content being delivered online, examinations continue to be viewed as an activity which occurs in a more traditional face-to-face situation (Coniam et al., 2021). There has been some take-up of technology in the area of assessment, but rather less than has been the case with online teaching (Gardner, 2020; Mays, 2021).

Assessment – and high-stakes assessment in particular outside certain public school systems where online testing is common – is generally viewed as something to be conducted in pen-and-paper mode, in front of an examiner/invigilator, in a physical test centre. While online learning technologies have permitted relatively

effective delivery of learning and teaching, the delivery of assessment in online mode has seen a mixture of advantages, problems and challenges: e.g., a reduction in cheating, connectivity issues etc (Sarrayrih and Ilyas, 2013, Hussein et al., 2020; Berrada et al., 2021).

Khan and Jawaid (2020), reporting on online assessment in Pakistan during the Covid pandemic, discuss how learning, teaching and assessment in particular need to be equally embraced in terms of access and delivery, stressing the need for attitudinal changes in the online delivery of assessment where both administrators and test-takers lose their fear of newly developed technology in economically developing nations.

García-Peñalvo et al. (2021), in the context of how Spanish universities responded to the Covid pandemic, provide a number of recommendations concerning online assessment. In addition to increased continuous assessment, they also suggest that technologies which support face-to-face teaching – such as teleconferencing – should be used to deliver assessment, in order to develop teacher and student readiness for and confidence in the “new context of online assessment” (p. 87). They stress that any marking schemes must be made known to students before any assessment takes place. García-Peñalvo et al. (2021) recommend that specifically designed online assessment methods be developed for the subject or group of students concerned when “complex subjects with a large number of students” (p. 88) are involved.

There are both benefits and drawbacks to taking tests in OLP mode for the test taker and the examining body as noted by Weiner and Henderson (2022). On the positive side, test takers may take an online-proctored exam in the comfort (and safety) of their own home, an important factor in times of a pandemic where movement is restricted or for test takers with a disability who find access to a remote testing centre challenging at the best of times let alone during a pandemic. In addition, the speed of test delivery and issuance of results may represent the benefit of exams taken in an OLP mode.

Online teaching has a rather longer history of accepted practices and expectations than online assessment. Online teaching, which has produced over a decade’s worth of research, stresses collaborative principles, such as discussion, peer support, learning that is tailored to individuals, self-regulated learning, encouraging students to set their own goals, and planning, monitoring and controlling their cognition (Boekaerts and Corno, 2005). In contrast, the online assessment record is shorter. There, expectations of assessment (and in particular high-stakes assessment) remain more traditional and, until relatively recently, have typically been the product of one test taker. While speaking tests administered by certain examination bodies involve group work, many examination bodies’ speaking tests, as with LanguageCert, involve a one-to-one interaction with an examiner. Furthermore, when it comes to test delivery, traditional views of comparability (and hence reliability), generally require that the same assessment be delivered to all test takers at the same time. However, in an online world, where the traditional approach to large-scale assessment is difficult, such a requirement potentially creates issues around security, honesty and fairness.

Regarding OLP examinations, there has been extensive discussion of security, the “vulnerability” of online tests and academic dishonesty (see Corrigan-Gibbs et al., 2015; Coniam et al., 2021). Such issues are very important, especially when examinations, often high stake, are taken in a remote location such as a test taker’s home.

Nonetheless, Foster and Layman (2013) describe how levels of security may be put in place which make the online proctoring of examinations viable. Indeed, there have been studies which report how exam security may even be more effective as a result of the technologies associated with monitoring of online examinations rather than in traditional face-to-face settings (Watson and Sottile, 2008; Rose, 2009).

Technical factors may also need some consideration. In their evaluation of OLP examinations, Giller et al. (2021) report a number of problematic issues, such as login failure and other technical issues (pp. 36-37). Such issues are not, however, the focus of the current study.

Despite such concerns, OLP remains a potentially important delivery method going forward. The current study explores the comparability and hence interchangeability of OLP assessment of speaking with traditional methods.

A brief summary of key issues surrounding the assessment of the speaking skill and assessing the skill remotely will now be provided.

Assessing Speaking

Speaking has long been considered the most complex of the four macro skills to assess. Some 40 years ago, Madsen (1983) outlined some of the reasons why speaking is challenging to assess. Apart from background construct issues such as defining the actual nature of the speaking skill and devising criteria to properly assess speaking in a communicative age, factors such as ability, tone, reasoning etc. as well as the reluctance of some test takers to even speak (p. 147) had to be dealt with.

Luoma (2004) reiterates how speaking is the most difficult language skill to assess reliably. This is especially the case when speaking is assessed by a human assessor in a face-to-face interaction, when assessments can be influenced by a number of factors such as features of spoken language, the test taker's language level, gender, the nature of the interaction, the tasks and topics driving the interactions, as well as the opportunities that the test taker has to demonstrate their ability. (2004: ix-x).

Sujanal (2016) echoes many of the above points in their discussion of the complexity of the aspects involved in testing oral proficiency, noting that many teachers almost avoid assessing speaking.

Assessing Speaking Online

Assessing speaking involves various 'complications', as mentioned above. To overcome some of these complexities, various educators and researchers have recommended moving the assessment of speaking to an online mode, which, they argue, affords advantages over a face-to-face mode. Fall et al. (2007), for example, describe a machine mediated Simulated Oral Proficiency Interview (SOPI) which renders large-scale assessment of test takers speaking proficiency on the ACTFL Oral Proficiency Scale comparatively easy to administer and rate. Regarding computer-mediated tests, Wagner comments, with reference to the Duolingo computerised

Speaking test, that such tests (i.e., those that are completely computer-mediated in that they have no human rater) tend to lack validity in that they generally assess constructs that are amenable to being assessed by computer. This results in a lack of real-world constructs being assessed.

It should be noted that LanguageCert Speaking Tests are administered face to face by human raters, albeit much of them via LanguageCert's OLP facility. Ockley et al. (2019) describe a speaking test administered online via the interactive video facility Skype. With test taker samples in the USA and China, Ockley et al. report comparative success with the assessment of oral abilities in interactive video.

Against the backdrop of the Covid pandemic, assessment of all forms moved, with differing degrees of success (Ali and Dmour, 2021), to various online modes. As might be expected – following the discussion above of the complexities of assessing speaking – it was indeed the assessment of students' oral proficiency that emerged as most challenging for many educators. Forrester (2020) elaborates the challenges of assessing speaking online in the time of the Covid pandemic. These issues apply to all forms of assessing oral proficiency, not just in formal examinations.

Comparability of Results from Exams Taken via OLP / TM

There has been considerable research into assessment conducted online with and without invigilation, although few studies have directly compared high-stakes tests conducted in OLP versus those conducted in traditional centre-based face-to-face mode. The following section briefly examines the research into these two related, if different, areas.

Examinations Conducted with and without Invigilation

Much of the research conducted on different modes of invigilation has been in higher education settings. Outside higher education and in the field of organisational psychology, Tippins (2015) discusses how new technology has led to "changes in the assumptions made about good testing practices" and the need "to confront new problems that are created by technological enhancements." She also provides examples of how technology is being used in assessments in realistic ways. In general, studies have reported, perhaps unsurprisingly, that students who sat tests without any invigilation – remote or otherwise – recorded higher grades than students who sat remote invigilated tests: Alessio et al., 2017; Goedl and Malla, 2020; Reisenwitz, 2020.

There have, however, been studies which reported no significant differences in the performance of students sitting tests with or without invigilation (see Castillo and Doe, 2017; Lee, 2020).

Examinations Conducted using Online Invigilation and in Traditional Centre-based Face-to-face Mode

Despite the increase in high-stakes assessments conducted online following the 2020-2022 Covid pandemic, as Weiner and Henderson (2022) observe, there has been little research into comparability of high-stakes test scores obtained from remotely-invigilated tests as opposed to tests invigilated face to face in testing centres. A summary of the limited amount of research in the area is presented below.

Weiner and Hurtz (2017) examined test taker performance in the context of licensing examinations in the USA, exploring the extent to which performance was equivalent regarding test takers sitting examinations in specially prepared computer-equipped 'kiosks' to test takers sitting the same examinations in physical test centres with human invigilators. No significant differences were found between performance in either proctoring mode. Hurtz and Wiener (2022) extended the scope of the above study following extended closures over the Covid pandemic. Their study reported no differences in test score due to proctoring mode.

Wuthisatian (2020) examined differences in performance between test takers taking high-stakes economics examinations using remote online proctoring versus those taken in traditional exam centres. Results suggested that test takers performed differently across the two proctoring methods: those who sat the examination at a centre obtained significantly higher scores than those test takers who were proctored online.

Cherry et al. (2021) examined professional licensure examinations in the USA, comparing outcomes for tests administered either by means of remote online proctoring or in test centres. While statistically significant differences were observed in results obtained between the two modes, no detectable pattern was observed in favour of either mode.

Morin et al. (2022) investigated a high-stakes national medical licensing examination in Canada taken via remote online proctoring or in exam centres. Despite some test takers reporting different examination experiences, Morin et al., report that test scores across the two proctoring modes – despite there being different examination question types – were broadly comparable.

Muckle et al.'s (2022) study explored scores on a study of North American pharmacy licensing examinations taken via the two proctoring modes following the Covid pandemic. Muckle et al. reported higher scores for examinations taken onsite by examinees. While they attribute some of the differences in results to the make-up of the sample, further research is clearly called for. Research conducted to gauge test taker reactions to LanguageCert's OLP delivery of tests (Coniam et al., 2021; Coniam, 2022) has thus far been generally positive – broadly echoing the results reported by Muckle et al. (2022) in their study.

The Study

The data in the current study are drawn from LanguageCert's International ESOL (IESOL) suite of Speaking tests administered between 2019-2021, with each test in the suite aligned to a CEFR level. The LanguageCert Speaking qualifications involve a comprehensive test of spoken English, with the tasks in the examinations

designed to test the use of English in real-life situations. The qualifications are suitable for non-native speakers of English worldwide; young people or adults attending an English course either in the UK or overseas; students learning English as part of their school or college curriculum; people applying to come to the UK for work purposes.

All Speaking tests comprise four tasks – of increasing complexity as test takers move through the test - and last from 12 minutes for the B1 examination to 17 minutes for the C2 examination. There are four rating scales, each of which has four score levels. The Speaking tests are conducted with a live interlocutor (whether face to face or via remote proctoring), with all examinations recorded for later grading and for use in possible appeals. All Speaking tests are scored against four rating scales. The maximum score is 50 with the following grades: Fail - below 50%, Pass for scores of 50%-74% and High Pass for scores of 75% and above. See <https://www.languagecert.org/en/language-exams/english/languagecert-international-esol>.

All examinations are assessed by a closed group of markers at LanguageCert, who are regularly standardised through training to ensure consistency and objectivity for assessments that are benchmarked against the CEFR (see Papargyris and Yan, 2022). A number of different test forms are available for each level of test with new test forms continually being added to the test pool.

To enhance security, not only are different test forms used randomly, but the four task types which comprise a test form are also randomised.

Table 1 below presents the number of test forms available for the 2018-2022 tests that were delivered, and the test taker sample for the analysis presented in the current study.

Table 1. Sample size

CEFR Level	Test Taker Sample Size	Different Test Forms
B1	19,745	30
B2	21,154	30
C1	7,943	29
C2	3,438	19

LanguageCert operates OLP internationally, with tests delivered in over 70 countries throughout the world. Consequently, all aspects of the assessment process by which OLP is conducted – logging on, security checks, connections and voice quality checks etc – are administered through the medium of English. In the face of potential English language constraints for lower-level proficiency test takers, the administration of tests in the IESOL suite by OLP principally takes place from B1 upwards. The dataset below for Speaking is therefore presented only for CEFR levels B1 to C2.

Since Speaking Test scores are obtained via the four rating scales, test reliability cannot be estimated via item- or rater-based estimation methods. It is, however, possible to estimate reliability by uni-dimensional factor analysis calculating McDonald's omega via the raw totals obtained for the four macroskills, i.e., Reading, Listening, Writing and Speaking, together with the CEFR grade awarded. Table 2 presents the reliability

estimates, including 95% confidence interval (CI) lower and upper bounds. (For brevity's sake, results are only reported for the Speaking Test.)

Table 2. Reliability estimates via McDonald's omega

CEFR Level	Speaking Test Score	Omega
B1	Posterior mean	0.64
	95% CI lower and upper bounds	0.64-0.65
B2	Posterior mean	0.62
	95% CI lower and upper bounds	0.62-0.63
C1	Posterior mean	0.65
	95% CI lower and upper bounds	0.64-0.66
C2	Posterior mean	0.72
	95% CI lower and upper bounds	0.71-0.74

McDonald's omega estimates may be interpreted in a similar manner to the Cronbach alpha, with 0.6 being acceptable. Table 3 below reports the McDonald's omega factor loadings for the Speaking Test.

Table 3. Single-factor model standardised loadings

CEFR Level	Factor	Standardised Loadings
B1	Grade	0.90
	Speaking test	0.96
B2	Grade	0.91
	Speaking test	0.96
C1	Grade	0.91
	Speaking test	0.97
C2	Grade	0.92
	Speaking test	0.96

As can be seen, loadings for Speaking tests and grades awarded at all CEFR levels are 0.90 and above, indicating that the Speaking tests exhibit a high degree of reliability.

Two sets of data are now presented below. One, descriptive statistics: means, standard deviations and effect size differences; two equivalence independent samples t-tests ("equivalence tests").

The equivalence independent samples t-test permit users to test the null hypothesis that the population means of two independent groups fall inside a user-defined interval, i.e., the equivalence region. The proce-

dure of using two-one-sided tests (TOST) permits significance to be observed via specified upper and lower bounds, as opposed to standard t-tests which report a single t score. As Lakens (2017) states:

Adopting equivalence tests will prevent the common misinterpretations of nonsignificant p values as the absence of an effect and nudge researchers toward specifying which effects they find worthwhile (p. 360)

The upper and lower bounds represent the extent of variation of t values regarding the two populations of the two samples being tested. If the t value of the equivalence test is within the estimated range, the two populations may be deemed to be equivalent.

Research Questions

The overarching hypothesis in the current study is that mean scores obtained between the two modes of test delivery – OLP and TM – will not be significantly different. Specifically, the following two hypotheses are pursued:

RQ 1: At worst, will only small effect size differences between the two modes be observed?

RQ 2: On equivalence tests, will significance emerge against specified upper and lower bounds for any given CEFR level?

Descriptive Statistics

Table 4 presents a summary of the effect size differences between the sets of means for the Speaking Test total score (maximum 50) for each mode using Cohen's d. Cohen's d indicates standardised differences between two means, sharpening comparisons between two means. In general, a small effect is taken as 0.2, a medium effect as 0.5, and a large effect as 0.8 (Glen, 2021).

Table 4. Effect size differences between mode means

Level	Mode	Number	Mean	Score Difference	SD	Cohen's d
B1	TM	17998	37.52	+1.04 (2.08%)	8.56	0.07
	OLP	1747	38.56		9.88	
B2	TM	11046	37.82	-0.58 (1.16%)	8.1	0.06
	OLP	10108	37.24		9.38	
C1	TM	2284	35.18	+0.14 (0.28%)	8.92	0.01
	OLP	5659	35.32		9.44	
C2	TM	1234	31.18	+4.12 (8.24%)	8.3	0.45
	OLP	2204	35.30		9.92	

As can be seen from Table 4, effect sizes are negligible for levels, B1 to C1. It is only at C2 level where the score difference between the two modes is greater than 5%, and where there is a notable small-to-medium effect size difference of 0.45.

Equivalence Tests

Tables 5 to 8 below present equivalence test results comparing OLP and TM.

Upper and lower bounds have been set at +/- 0.05 (i.e., the 95% interval) of the raw score (see Lakens, 2017). These bounds may be construed as representing 95% confidence intervals; however, as TOST consists of two one-sided tests, it makes more precise sense to refer to the upper and lower ends of the confidence intervals. The critical decision on equivalence, as stated earlier, is whether the estimated t value (labelled T-Test in the tables below) is between the upper and lower bound. The p values for the t values (Upper bound, T-Test and Lower bound) indicate significant T-Test values where these go beyond the specified bounds.

Table 5. B1 Equivalence test results

Statistic	t	df	p
Upper bound	-5.26	19743	< .001
T-Test	-4.80	19743	< .001
Lower bound	-4.34	19743	1.00

Table 6. B2 Equivalence test results

Statistic	t	df	p
Upper bound	3.97	21152	1.00
T-Test	4.80	21152	< .001
Lower bound	5.63	21152	< .001

Table 7. C1 Equivalence test results

Statistic	t	df	p
Upper bound	-1.07	7941	0.14
T-Test	-0.63	7941	0.53
Lower bound	-0.20	7941	0.58

Table 8. C2 Equivalence test results

Statistic	t	df	p
Upper bound	-12.78	3436	< .001
T-Test	-12.48	3436	< .001
Lower bound	-12.18	3436	1.00

At none of the four levels was significance observed at both lower and upper bounds. This indicates that although there is not a perfect match, the two modes of Speaking Test administration can be considered broadly equivalent for all the CEFR levels in the study. That said, there would appear to be an issue with the C2 level test, where more investigation is clearly called for.

Discussion and Conclusion

This study has explored the comparability of scores obtained by test takers of LanguageCert's IESOL English language Speaking Tests at CEFR levels B1 to C2 via traditional face-to-face mode (TM) versus online proctored mode (OLP).

The key hypothesis in the study was that mean scores and hence performance obtained in the OLP and TM modes of test delivery would not be significantly different. Specifically, two research questions were being investigated.

The first RQ was that, at worst, only small effect size differences between the two modes would be observed. While negligible effect sizes were observed for levels B1 to C1, a small-to-medium effect size was observed for C2.

The second RQ was that, on equivalence tests, significance would not emerge against specified upper and lower bounds for any given CEFR level. As significance was not observed for both bounds in any of the test levels, it was determined that the two modes of test administration may be considered equivalent broadly for the four CEFR levels examined. Nevertheless, at the highest level of ability (CEFR level C2), test takers scored considerably higher in online proctored mode than in face-to-face mode.

There are two possible reasons for such a discrepancy. One relates to the profile of the C2 test taker cohort. C2 level test takers tend to be professionals in their 30s and 40s, whereas at the lower levels, many test takers are younger school children who are more accustomed to traditional face-to-face centre-based assessments. In this light, C2 test takers are also more comfortable with extensive use of technology, a fact which may account for them being more at ease in the online proctored environment. The second issue is possibly that of malpractice. In this regard, however, stringent security checks to guard against issues such as impersonation are conducted before Speaking Tests take place. Speaking Test materials are, as mentioned, randomised to forestall possible pre-arranged sets of answers. Further, the Speaking Test is an oral performance test conducted in real time, which makes cheating much more difficult to carry out from a test taker's point of view.

To conclude, it would appear that results obtained from taking LanguageCert IESOL Speaking Tests at the lower CEFR levels indicate that similar results are obtained irrespective of whether tests are taken in traditional face-to-face mode or in online proctored mode. Nonetheless, the fact that C2 test takers score in an online mode higher does require further investigations at this level.

One limitation of the current study is that only one skill has been investigated – speaking. The skill of speaking is generally viewed as the most difficult to administer and assess, with difficulties in online delivery exacer-

bated rather more than with the more 'static' (in the sense that they do not require direct interaction with an interlocutor) skills of listening, reading and writing. A follow-up study analysing the other skills – listening, reading and writing – is underway.

The findings both in this chapter and in the previous chapter, Chapter 8, promise that greater reliance on OLP, when carried out meticulously, not only provides reassurance and evidence that the mode of proctoring is not significant, but indicate that OLP can provide even greater security in terms of avoiding interference or cheating in high-stakes examinations. These studies will, of course, continue in the ongoing battle against cheating and the growth of assessment security concerns as technological innovations, including those brought about by AI Bots, threaten traditional forms of examination conduct. LanguageCert will continue to work at these concerns to maintain the quality of its examinations.

A further consideration occurs in the next chapter, Chapter 12, when Lampropoulou investigates the increasingly important factor of Interactional Competence. In it she discusses the strengths and weaknesses of OLP when Interactional Competence is assessed.

References

- Alessio, H. M., Malay, N., Maurer, K., Bailer, A. J., & Rubin, B. (2017). Examining the effect of proctoring on online test scores. *Online Learning*, 21(1), 146-161.
- Ali, L., & Dmour, N. A. H. H. A. (2021). The shift to online assessment due to COVID-19: An empirical study of university students, behaviour and performance, in the region of UAE. *International Journal of Information and Education Technology*, 11(5), 220-228.
- Berrada, K., Ahmad, H. A. S., Margoum, S., EL Kharki, K., Machwate, S., Bendaoud, R., & Burgos, D. (2021). From the paper textbook to the online screen: A smart strategy to survive as an online learner. In *Radical Solutions for Education in a Crisis Context* (pp. 191-205). Springer, Singapore.
- Castillo, M. S., & Doe, R. (2017). Mobile and Nonmobile Assessment in Organizations: Does Proctoring Make a Difference? *Psychology*, 8(06), 878.
- Cherry, G., O'Leary, M., Naumenko, O., Kuan, L. A., & Waters, L. (2021). Do outcomes from high stakes examinations taken in test centres and via live remote proctoring differ?. *Computers and Education Open*, 2, 100061.
- Coniam, D., Lampropoulou, L., & Cheilari, A. (2021). Online proctoring of high-stakes examinations: A survey of past test takers' attitudes and perceptions. *English Language Teaching*, 14(8), 58-72.
- Coniam, D. (2022). Online invigilation of English language examinations: A survey of past China test takers' attitudes and perceptions. *International Journal of TESOL Studies*, 4(1), 21-31.
- Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E., & Thies, W. (2015). Detering cheating in online environments. *ACM Transactions on Computer-Human Interaction*, 22(6), 1-23.
- Fall, T., Adair-Hauck, B., & Glisan, E. (2007). Assessing students' oral proficiency: A case for online testing. *Foreign Language Annals*, 40(3), 377-406.
- Forrester, A. (2020). Addressing the challenges of group speaking assessments in the time of the Coronavirus. *International Journal of TESOL Studies*, 2(2), 74-88.

- 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>.
- García-Peñalvo, F. J., Corell, A., Abella-García, V., & Grande-de-Prado, M. (2021). Recommendations for mandatory online assessment in higher education during the COVID-19 pandemic. In *Radical solutions for education in a crisis context* (pp. 85-98). Springer, Singapore.
- 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(5).
- Giller, P. (2021). E-proctoring in theory and practice: a review. Dublin, Ireland. Quality and Qualifications Ireland.
- Glen, S. (2021). Cohen's D: Definition, Examples, Formulas. <https://www.statisticshowto.com/>.
- Goedl, P. A., & Malla, G. B. (2020). A study of grade equivalency between proctored and unproctored exams in distance education. *American Journal of Distance Education*, 34(4), 280-289.
- Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. *EDUCAUSE Review*.
- Hurtz, G. M., & Weiner, J. A. (2022). Comparability and integrity of online remote vs. onsite proctored credentialing exams. *Journal of Applied Testing Technology*, 23, 36-45.
- Khan, R. A., & Jawaid, M. (2020). Technology enhanced assessment (TEA) in COVID 19 pandemic. *Pakistan Journal of Medical Sciences*, 36(19), 108-110.
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4), 355-362.
- Lee, J. W. (2020). Impact of proctoring environments on student performance: Online vs offline proctored exams. *The Journal of Asian Finance, Economics, and Business*, 7(8), 653-660.
- 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.
- Mays, T. J. (2021). Teaching the teachers. In *Radical Solutions for Education in a Crisis Context* (pp. 163-176). Springer, Singapore.
- Morin, M., Alves, C., & De Champlain, A. (2021). The show must go on: Lessons learned from using remote proctoring in a high-stakes medical licensing exam program in response to severe disruption. *Journal of Applied Testing Technology*, 23, 15-35.
- Muckle, T. J., Meng, Y., & Johnson, S. (2022). A Quantitative Evaluation of a Live Remote Proctoring Pilot. *Journal of Applied Testing Technology*, 23, 46-53.
- Ockey, G. J., Timpe-Laughlin, V., Davis, L., & Gu, L. (2019). Exploring the potential of a video-mediated interactive speaking assessment. *ETS Research Report Series*, 2019(1), 1-29.
- Papargyris, Y., & Yan, Z. (2022). Examiner quality and consistency across LanguageCert Writing Tests. *International Journal of TESOL Studies*, 4(1), 203-212..
- Reisenwitz, T. H. (2020). Examining the necessity of proctoring online exams. *Journal of Higher Education Theory and Practice*, 20(1), 118-124. doi.org/10.33423/jhetp.v20i1.2782.
- Rose, C. (2009). Virtual proctoring in distance education: An open-source solution. *American Journal of Business Education*, 2(2), 81-88. doi.org/10.19030/ajbe.v2i2.4039.
- Sarrayrih, M. A., & Ilyas, M. (2013). Challenges of online exam, performances and problems for online university exam. *International Journal of Computer Science Issues (IJCSI)*, 10(1), 439.

- Sujana, I. M. (2016). Assessing Oral Proficiency: Problems and Suggestions for Elicitation Techniques. <https://academia.edu>.
- Tippins, N. T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior*, 2, 551–582. <https://doi.org/10.1146/annurev-orgpsych-031413-091317>.
- Wagner, E. (2020). Duolingo English test, revised version July 2019. *Language Assessment Quarterly*, 17(3), 300-315.
- Watson, G., & Sottile, J. (2010). Cheating in the Digital Age: Do students cheat more in on-line courses? *Online Journal of Distance Learning Administration*, 13(1).
- Weiner, J. A., & Henderson, D. (2022). Online Remote Proctored Delivery of High Stakes Tests: Issues and Research. *Journal of Applied Testing Technology*, 23, 1-4.
- 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.
- Wuthisatian, R. (2020). Student exam performance in different proctored environments: Evidence from an online economics course. *International Review of Economics Education*, 35, 100196.
- 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.

Chapter 11: Interactional Competence and the Role Roleplay Plays: The LanguageCert Perspective

Leda Lampropoulou

Abstract

This chapter describes the importance of including the construct of interactional competence in speaking assessments, drawing mainly from the literature in the field of language testing. The co-construction of meaning and the shared nature of the interaction are seen to be operationalised in an optimal manner using the roleplay task. The effect of the task is explored through the perspective of the LanguageCert International ESOL Speaking exams, which are used as examples to demonstrate the issues of scalability, discriminability, score separability, and the so-called interlocutor effect. Further research and technological innovations will assist in defining and scrutinising the aspects of interactional competence that can be reliably measured.

Keywords: Interactional competence, role play, Speaking tests, oral assessment

Introduction

This chapter critically summarises the research conducted in the field of interactional competence (IC) in order to describe how the construct of IC has been operationalised in Oral Proficiency Interviews (OPIs) in language testing in LanguageCert examinations. More specifically, the chapter focuses on describing how the research findings have influenced the format of OPIs, and on explaining the issues and the challenges which have been identified, as well as the issues addressed through the inclusion of roleplay tasks. The operationalisation of the assessment of IC through the prism of a specific test is considered, through the LanguageCert International ESOL Speaking suite of exams, specifically through the roleplay task these use as part of the test format.

Background and Definitions

Two decades have passed since Young (2000) described interactional competence (IC) as “a relatively new theory of spoken language use in face-to-face communication” (p.3). It was three decades before then that Hymes had used the term communicative competence to account for sociocultural variation in language use and acquisition, to challenge Chomsky’s dichotomy between competence and performance, contending that grammar rules cannot exist alone and, therefore lack meaningfulness unless they are considered together with the rules for their functional use (Hymes, 1972). Hymes’ ideas were further developed by Canale and Swain (1980) into an applied linguistics theory which suggested that an individual’s competence includes linguistic competence, discourse competence, pragmatic competence, and strategic competence. L2 teaching practices were strongly influenced by this theory of communicative competence, and its effect soon extended into language assessment, through Bachman (1990) and Bachman and Palmer (1996), and their observations of the assessment of communicative language ability.

Kramsch (1986) built on Hymes’ theories to develop the construct of what she coined interactional competence, and in defining it she explained that:

[S]uccessful interaction presupposes not only a shared knowledge of the world, the reference to a common external context of communication, but also the construction of a shared internal context or “sphere of inter-subjectivity” that is built through the collaborative efforts of the interactional partners. (p.367)

The interpretation of test taker’s speaking performance from this perspective could perhaps alleviate McNamara’s (1997) concern that language assessment based on previous theories considered the test taker’s performance in an unrealistically detached manner, and that the test taker was viewed as the sole person liable for the development of the performance, without considering that they were not the only one participating in it.

The IC construct was also explored by Hall (1995), who focused on interactive practices for which she saw a socially cohesive role for a community, developing through speech acts. Her considerations link pragmatic competence with communicative competence and interactive competence. This link, together with the idea that context is central to the speaking construct, begs the question for a distinction between IC and pragmatics. Young (2011) attempted to answer this by contending that they are interconnected but still distinct competencies. Plough et al. (2018) also identified similarities between IC and pragmatic competence in that they both make use of other competencies, such as grammatical and textual competencies. These are used in parallel as tools to achieve the communication of the intended message, yet IC is highlighted as the skill necessary for “building and maintaining relationships, an aspect of the co-constructed nature of speech” (p.442). The distinction is made even clearer through the understanding that IC emphasises the element of being almost equally constructed by all participants in a discursive practice and is specific to that practice in particular. (Young, 2019).

More importantly, perhaps, Young (2008) notes that IC is not to be found within the individual’s skillset or cognitive ability. Young asserts that since participants accomplish the interaction task jointly, the skills described in the theory are distributed among all participants in the interaction. It would, therefore, be inaccurate to claim interactional competence as a skill that a person exercises outside an interaction (He and Young, 1998).

Lam, in contrast, asserts that such a skill can only be showcased in the context of a multi-participant interaction which will also rely on the co-participants' performance (Lam, 2018)

It becomes clear that, to the question featuring in the title of McNamara's (1997) article "'Interaction' in second language performance assessment: Whose performance?", the answer can only be, the co-participants in the interaction.

Interactional Competence in Language Proficiency Interviews

The assessment of the speaking construct through language proficiency interviews and the extent of operationalisation of IC in different types of oral tests has led to what Galaczi and Taylor (2018) describe as two important strands in theoretical and empirical research, the debates on authenticity and variability.

Roever and Kasper (2018) see a similar 'tug-of-war' between: the conceptualisation of the construct from a primarily psycholinguistic-individualist perspective; and a primarily sociolinguistic-interactional perspective. It is clear that test developers face a dilemma, in which opting for the former perspective focuses on the individual and allows the elicitation of rateable amounts of language samples but can be considered invalid by failing to support inferences on the test taker's ability in typical, real-life interactions. It can, however, be hypothesised that such inferences can be validly supported by speaking tests designed to engage test takers in meaningful, interactive, social situations.

Even before Roever and Kasper's work, language proficiency interviews, such as the Oral Proficiency Interview (OPI), had been castigated for failing to recreate the co-constructed nature of interaction realistically and authentically, and for the absence of salient features of natural conversation caused by the asymmetric relationship between the interlocutors (Young and Milanovic, 1992; Johnson and Tyler, 1998; Johnson, 2001). When speaking tests are based solely on interview-like tasks and conducted in an interview setting, an unequal interaction will occur which will prevent the test from measuring conversational competence in an appropriate manner (Kormos, 1999).

The emerging picture is that if a speaking language test cares to make claims about measuring speaking performances which can be indicative of and generalisable to interactive social contexts, this can only happen through a broadened construct that includes interactional competence operationalised through tasks in which the co-participants jointly engage in conversation. This is the work that LanguageCert are currently developing and trialling.

The Role Roleplay Plays

Paired (and grouped) speaking tests, by nature, allow test takers to interact and co-construct discourse, a strength which, among other reasons, has made the paired format a common choice, not only for class-

room-based assessment, but also for high-stakes exams (May, 2011). Moreover, Ockey et al. (2015) suggest that even monologic speaking tests can measure interactional competence with the inclusion of dedicated tasks. Consequently, the *onus probandi* (burden of proof) appears to fall on task design.

On the one hand, Plough et al. (2018) claim that a unanimous verdict has yet to be reached regarding the extent to which the optimum operationalisation of IC relates to specific speaking task types. On the other hand, the roleplay task seems to have won the battle between the choice of tasks, as suggested by the findings of several studies.

Kormos (1999) compared non-scripted interviews and guided roleplay activities in oral assessments using discourse analysis and found that in roleplay “the conversational interaction is more symmetrical” (p.1). Moreover, she established that roleplay tasks can imitate aspects of conversations in an authentic and realistic manner and found that they can be useful in measuring conversational competence as exhibited in the test takers’ performance, while also concluding that, in terms of measuring conversation management, roleplay activities can better elicit the manifestation of IC features. Okada’s (2010) findings align with Kormos’ (1999) conclusions. In his study, which discusses roleplay in OPIs in terms of its construct validity, he describes the competencies displayed in performing a roleplay activity as strongly resembling those observed in real-life conversations and he concludes by recognising roleplay as a valid assessment instrument. In a very recent study based on a conversation analysis (CA) of a corpus of roleplay interaction, Youn (2020) was able to confirm these findings, while maintaining that the language samples elicited through roleplay interactions, despite not being entirely authentic, can still showcase the test taker’s level of competence regarding how well they would perform in a similar interaction in real life. Hu (2015) also found that roleplay affords an easier access to IC features than other types of paired tasks.

Apart from the conversational characteristics of IC featuring realistically in roleplay, researchers were able to point to more reasons arguing for the inclusion of such tasks in oral proficiency interviews. As an example, in response to the debates on validity and authenticity, Kasper and Youn (2018) assert that roleplay can be used to generate performances with authentic interactional features, such as topic and turn taking management. In addition, attempts to sequence organization attempts, while affording testers the element of control required to make the interaction measurable, render roleplay valid in terms of construct representation. The potential of roleplay tasks to allow test takers to co-construct discourse is also noted by Galaczi and Taylor (2020).

It is important to note that roleplay tasks need not be limited to paired speaking test formats, however. OPI roleplays can be conducted with a trained examiner/ interlocutor assuming different roles (Ikeda, 2017; Youn, 2015, 2020). In these OPI roleplays, as in the LanguageCert International ESOL Speaking suite of exams (Appendix 1), a specific part of the speaking test is dedicated to a roleplay activity. During that part of the test, the examiner sets the context by informing the test taker of the scenario and the roles to be assumed (as also explained in Kasper and Youn, 2018).

In the case of LanguageCert, the examiner may assume different personas, which range in register formality, such as a colleague or a line manager, a neighbour or a stranger in the street, a doctor’s receptionist or a tour agent, thus enabling different levels of the Common European Framework of Reference for Languages (CEFR) (as analysed by the Council of Europe, 2017) to be measured. Unlike the interlocutor/examiners assuming dif-

ferent roles, the test taker, is not expected to take on a role other than their actual self in that interaction. In this context, a high degree of authenticity can be achieved, since the language the test taker will have to use can be expected to resemble the Target Language Use (TLU) domain of social interactions. This is because, in real life, they are likely to need to book a dentist's appointment or a hotel room, but it would be irrelevant for a non-specific test to assess how well the test takers can perform on the other end of the interaction and assume the role of the doctor or the receptionist, the kind of roles that they may never need to assume in real life.

In the interactions described above, which can either be brief or develop unscripted for a longer period depending on the targeted CEFR level and the test taker's ability, a wider range of functions can be elicited than the interviewer-structured interaction allows, such as expressing regret, sympathy, condolence, expressing surprise or lack of it, complaining, offering and accepting an apology, etc. (LanguageCert, 2020). The item writer aiming to elicit the demonstration of functional language relating, for instance, to an apology may choose to set the context of the test taker's late arrival for a meeting with a friend or to work.

Given the evidence above, we can assert with confidence that roleplay tasks can be considered as appropriately operationalising the construct of interactional competence (Grabowski 2013; Kasper and Youn, 2018; Walters 2007, 2013; Youn 2015).

Assessing interactional Competence

Assessment professionals who adopt a sociolinguistic-interactional perspective, foregrounded by the research in applied linguistics, which includes roleplay and other interactive tasks to operationalise the construct of interactional competence, are immediately faced with the challenge of having to assess it. This is a multi-faceted challenge. The reason for the challenge for assessors is that research has revealed two main problematising areas in the measurement of IC. These are, one, the need for differentiation at various levels (or "scalability"); and two, "discriminability" (Galaczi and Taylor, 2018, p. 230), where the separability of scores in the co-constructed performance, also sometimes referred to as "the interlocutor effect" (O'Sullivan, 2002), must deal with the issue of the feasibility of measuring non-verbal behaviour as part of the construct itself.

Descriptors and Scalability and Discriminability Issues

The Common European Framework (CEFR) has developed a descriptive scheme providing scaled descriptors for communicative language competences, which are classified into linguistic competences, sociolinguistic competences, and pragmatic competences (Council of Europe, 2018). The scheme – provided in Appendix 3 – can be read horizontally and vertically, with the horizontal dimension describing the different capabilities expected at the level, while the vertical one attempts to sequence an ascending series of learner proficiency. There is one scale specific to IC, Interaction. Table 1 elaborates.

Table 1: CEFR Interaction scale across the CEFR levels

Level	INTERACTION
A1	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.
A2	Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.
B1	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.
B2	Can initiate discourse, take his/her turn when appropriate and end conversation when he / she needs to, though he /she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.
C1	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skillfully to those of other speakers.
C2	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turn taking, referencing, allusion making etc.

Note. Reprinted from <https://www.coe.int/en/web/common-european-framework-reference-languages/table-3-cefr-3.3-common-reference-levels-qualitative-aspects-of-spoken-language-use>

Table 1 above outlines the relevant IC features that learners can be expected to have acquired at each CEFR level, from A1 to C2. Oral proficiency tests mapped to the CEFR often use these as a reference tool to describe standard performance expected at each exam level. To use the same example as in the previous section with the roleplay task, the LanguageCert International ESOL Speaking test mark scheme describes an A1-A2 test taker as being expected to rely on the support of the interlocutor/ examiner. At B1 level, turn taking is expected to be mostly natural, whereas at B2 level, the test taker should be able to handle topic and turn management appropriately and independently, while not always elegantly. These areas are accounted for under Task Fulfilment and Coherence. There are also descriptors under the Pronunciation, Intonation and Fluency criterion which references the use of intonation to support meaning.

There are, however, no descriptors to cover non-verbal behaviour features, such as eye contact or posture (LanguageCert, 2021). Both have been identified as key features of IC. Such features might be difficult to include in standardised high-stakes exams at present, at least until more research and technological advances permit. However, there are more IC features that could be described and used to measure IC.

The importance an assessment developer places on IC can perhaps be detected by noting whether IC features are displayed under various criteria, informing them by being included in the descriptors, or whether IC is seen as a separate criterion, in a manner that also has a greater impact on the test taker's overall score. In the LanguageCert ESOL Speaking exam the former is the case, but given the prominence that IC is currently being given and the research evidencing its role in communication, it will be interesting to see whether in future revisions of the LanguageCert ESOL Speaking mark schemes IC features will be assigned to a criterion on their

own, as has happened with more recently developed LanguageCert exams, e.g., the LanguageCert SELT Speaking and Listening test, in which the criterion is referred to as Interactive Communication and Task Fulfilment.

The CEFR scales on interaction (Council of Europe, 2018) do include references to some IC features, yet not consistently nor at all levels. The need for clearer and more specific descriptors differentiating between performance levels has been highlighted across L2 assessment literature relating to IC (Galaczi, 2014; Galaczi and Taylor, 2018, 2020; Lam, 2018; Seedhouse, 2012). These descriptors will need to be developed further, before they can be of wider use to language assessment stakeholders. Furthermore, Galaczi and Taylor (2020), in listing the key features of IC, also refer to breakdown repair, interactive listening, and non-verbal behaviour, aspects analysed in a very limited way in the CEFR scales, even though research relating to rater studies and test taker discourse has noted that they are salient IC features (Galaczi, 2014; Gan, 2010; May, 2011; Orr, 2002).

Roever and Ikeda (2021) argue that “IC develops along a predictable trajectory” (p.3). They report that research in second language acquisition demonstrates that – as proficiency improves – learners’ IC expands in range and improves in appropriateness (Al-Gahtani and Roever, 2012, 2014, 2018; Cekaite, 2007; Pekarek Doehler, 2019). However, Roever and Ikeda (2021) still identify a challenge in drawing a clear distinction at the higher levels, where IC features may be harder to describe. This aspect of IC can apply both to L2 and to L1 speakers.

This challenge may be illustrated in the descriptors in the mark schemes used at the higher levels of the LanguageCert International ESOL speaking exam where the differentiation – albeit minimal – between the descriptions of turn-taking performance at the two higher CEFR levels exists. At CEFR C1, the criterion for a passing mark at Task Fulfilment and Coherence includes a descriptor of a performance where turn taking is naturally handled. Going up a mark at the same criterion, turn taking needs to be spontaneous, flexible and wholly natural. Looking at the highest level offered, CEFR C2, the expectation for a passing mark under the same criterion describes turn taking as naturally handled with a high degree of flexibility whereas for full marks the descriptor expects turn taking to be consistently spontaneous, flexible and wholly natural. It is seen that the differentiation between a passing performance and the one achieving full marks is made through assessing how consistently and flexibly the skill is demonstrated. However, although this is not uncommon for mark scheme descriptors aligned to the CEFR, it is perhaps indicative of the CEFR’s limitation pertaining to the vague differentiation between IC descriptors at the different levels, underlined by researchers just above. In addition, in marking examiners’ training, the differences in benchmarked performances can be used to standardise what a performance at the level entails, and this is much easier to achieve at a level-specific test, such as the LanguageCert IESOL, than at a multi-level one.

Researchers still maintain that scalability and discriminability are possible. It is understood, as mentioned above, that IC develops in parallel with the learners’ general L2 language ability and that as the learners’ cognitive processes rely on higher automaticity of conversation processes (Field, 2011), their working memory will afford them a more effective and collaborative participation in interactions (Galaczi, 2014). Roever and Kasper (2018) point to the sequential organization of speech events as a gradable characteristic that can be classified and rated. In their study, they suggest that certain interactional features, such as repair, could be induced by the examiner attempting to elicit this strategy. Galaczi and Taylor (2020) also advise in favour of supporting interaction at the lower CEFR levels with visual or verbal prompts, for reasons of scoring practicality and reliability. Lam (2018) looks at IC through the prism of interactive listening and notes that IC features need to be

accounted for as more than the sum of the test taker's responses, and that their appropriacy to the interaction need to be given prominence.

In discussing roleplay tasks, Youn (2019, 2020) provides evidence that interactional performances can be elicited so that differences can be measurable against rating criteria. This task type also seems to offer itself for appropriately accommodating highly specific professional contexts, such as the context of radiotelephony communication in aviation, where the need to include IC reference to elements of professional knowledge and role behaviour seems to be particularly critical (Kim, 2013). For example, Kim (2013) suggests that the success of the communication in the interaction between pilots and air traffic controllers is so important that the test taker's ability to effectively interact using the aviation radiotelephony conventions should form part of the construct of such an ESP assessment. In such an assessment, the roleplay would assign the test taker with the role they will be called on to operate in in their future job, whereas the examiner would take on the persona of the opposite role, to achieve an, as much as possible, authentic performance. The research literature that has been discussed above would appear to indicate that roleplay tasks are strong contenders for being judged the most effective means by which IC can be measured.

Interlocutor Effect and Score Separability Issues

O'Sullivan (2002) used the term 'interlocutor effect' to refer to the sociolinguistics concept of the influence asserted in the interaction by the participants' identities and characteristics. From an assessment perspective, where the focus is traditionally on the individual, the idea and perhaps even the name of interactional competence could be enough to raise concern over standardisation and, consequently, validity. At the same time, the co-construction of meaning between the interlocutors perplexes this further, as the test takers' contributions and their performance are seen as shared, interwoven, and linked (Brown, 2003; May, 2011; McNamara, 1996; Roever and Kasper, 2018).

L2 assessment research exploring the different interlocutor variables such as gender, cultural background, acquaintanceship (O'Sullivan, 2002), and extroversion (Nakatsuhara, 2013), did find such characteristics exerting an influence. However, Brown and McNamara (2004) concluded that "the magnitude or direction of that influence is less clear and not directly predictable" (as cited in Galaczi and Taylor, 2020, p.343). More importantly, it is the construct definition that should determine whether this variability is irrelevant and undesired, or whether it is actually part of the construct itself (Galaczi & Taylor, 2020).

Even so, the paired speaking test format can be criticised for (mis)matching test takers of different abilities, causing an observed shared performance that is unrepresentative of the true capabilities of the individual participants in the interaction. Hu (2015) claims that a more proficient speaker will be disadvantaged if paired with a substantially less proficient speaker.

On the debate on the (in)separability of test takers' scores, May's (2001) suggestion that shared scores could be awarded in response to what raters perceive as a mutually achieved performance has not yet been widely accepted, and although it is a tempting prospect, in high-stakes testing especially, it appears that there is a long way to go before this can be done, if ever.

For now, the safest path seems to include the challenge of having to overcome the perception of test takers' contributions to the interactions as entangled (Fulcher, 2010) and of training raters to isolate what the individual test taker brings to the paired task. Under this light, raters might be facilitated by a paired task performed between the individual test taker and the interlocutor/examiner, instead of between a pair of test takers. The test takers' contributions can become even more distinguishable and measurable in a roleplay task, where opportunities for a more symmetrical interaction can be afforded, and the interlocutor/examiner can be trained to elicit specific IC resources the tester is interested in examining.

Non-verbal Behaviour

A third issue which needs to be mentioned as a problematising area in assessing IC is non-verbal behaviour, even though it has been considerably less researched in L2 assessment literature, both in terms of its conceptualisation as part of the IC construct, and its operationalisation. Features such as eye contact, facial expression, and posture have been included by Galaczi and Taylor (2020) as denoting non-verbal behaviour pertaining to IC. Researchers have indicated that raters perceive and note non-verbal behaviour even if it is not described in the rating scales (May, 2011; Nakatsuhara et al., 2018; Vo, 2019). Nonetheless, it is still seen as too complex a model to attempt to assess. Oksaar (1990), one of the first explorers of the concept, who was also able to provide insight from multilingual contexts, defined IC aspects with reference to "cultureme and behaviourme" (p.530), which include paralinguistic features as well as sociocultural norms, which, if testers are to include them in the construct, will also need to answer the pertinent question: 'whose culture?'

To conclude, integrating IC scales into speaking assessments would appear to enable a wider and more accurate representation of the construct as well as allowing valid inferences about real-world speaking competences, despite the issues which remain under investigation (Roever and Kasper, 2018).

Developing Research Areas

L2 development of interactional competence (IC) has been widely explored in the literature and continues to offer a fertile field for research, while L2 assessment literature has been growing exponentially, and can be expected to continue in a similar manner. The construct of IC is far from having been completely researched, and areas of future research involve both older and newer developments in language testing in general. Plough et al. (2018), see future research targeting four main areas. The first two pertain to issues already touched upon in this literature review, namely the link between task type and elicited evidence of the IC construct, and the role of the 'behaviourme'. The other two involve technology-related issues, as in the effect of the mode of speaking test delivery on IC affordances to test-takers, and the extent to which IC inferences can be drawn using computer-delivered tests.

On the first pointer, Youn (2020) argues for the usefulness of CA contributions in recognising various interactional devices in speaking assessment discourse emerging from interactional performances, to inform L2 learning and assessment.

Roever and Kasper's suggestion incorporates the second and the fourth points, and combines the visual access allowed by computer-assisted testing with using the methodological tools multimodal CA provides, to drive research on non-verbal behaviour such as gaze, gesture, and head movements as part of IC.

On the comparability of the IC construct through different modes of delivery, Nakatsuhara et al. (2017) looked at video-conferencing meetings and how these were distinguished from face-to-face meetings in the use of back channels and the management of turn taking. The researchers noticed differences in the interactions which point to the question of whether IC could support different operationalisations for different delivery modes.

In computer-delivered tests May (2011) also sees the potential for isolating test-takers contributions to co-constructed interactions through a standardised prompt. To these areas, Lam (2018) adds the need for research to support the creation and development of more accurate IC rating scales.

Conclusion

This chapter has provided a critical overview of the literature looking at interactional competence as a skill and construct, and its conceptualisation and operationalisation in Oral Proficiency Interviews in language testing using the roleplay activity as a task type. It has become possible to recognise interactional competence as an important construct pertaining to spoken ability, one that is highly relevant to real-life social contexts. More specifically, IC features such as turn management (e.g., interrupting), interactive listening (e.g., backchanneling), or non-verbal behaviour (e.g., laughter) are seen as key concepts in measuring interactional competence (Galaczi and Taylor, 2020). These have had a varying degree of uptake from assessment developers as some seem easier than others to integrate into assessment tasks, such as turn and topic management. Others, however, seem to require further research or innovative technology before they can be accepted by testers and test stakeholders as measurable and construct-relevant. Non-verbal behaviour or interactive listening are two such areas that require further research.

The literature relating the measurement of interactional competence with a specific task has found the roleplay activity to be a realistic and authentic task type, able to tap into most of the IC characteristics testers would wish to elicit and measure. However, as there is no overall comparison of all possible tasks, as with a lot of issues in assessment, there are no definite solutions without considering the test purpose and the TLU domain. Nevertheless, the roleplay task has been found to afford a less unequal interaction than other types of tasks, like non-scripted interviews (Kormos, 1999) and through appropriately designed roleplay situations the power imbalance can be authentically created and simulated, as in a situation between an employee and their manager, or a patient and their doctor.

To better illustrate the roleplay task's effectiveness in assessing IC, the LanguageCert International ESOL test has been used. It appears to be able to operationalise the IC features that assessment developers aim to elicit, in addition to overcoming the challenge of the inseparability of scores, since the performance is shared between the test taker and the interlocutor. The issues identified in the literature that also seem pertinent to

the specific assessment of IC relate to the inclusion of more IC features in the mark scheme and the scalability of these, together with the issue of including relevant aspects of non-verbal behaviour.

These findings have contributed to the development of LanguageCert's Academic and General speaking tests. An awareness of the factors involved in IC enables examiners to be more sensitive and cognisant of interpersonal discourse especially when different roles are used in roleplay assessment task.

Looking to the future, more research and empirical studies will allow a stronger integration of IC features in tests measuring speaking constructs. This can only be further facilitated through technological innovations which will accelerate and enhance assessment design and delivery, and allow a fuller exploration and conceptualisation of interactional competence.

References

- Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics*, 33(1), 42–65.
- Al-Gahtani, S., & Roever, C. (2014). Preference structure in L2 Arabic requests. *Intercultural Pragmatics*, 11(4), 619–643.
- Al-Gahtani, S., & Roever, C. (2018). Proficiency and preference organization in second language refusals. *Journal of Pragmatics*, 129, 140–153.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. & Palmer, A. (1996). *Language Testing in Practice*. Oxford University Press.
- Barraja-Rohan, A. (2011). Using conversation analysis in the second language classroom to teach interactional competence. *Language Teaching Research* 15(4) 479–507.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- Cekaite, A. (2007). A child's development of interactional competence in a Swedish L2 classroom. *The Modern Language Journal*, 91(1), 45–62.
- Council of Europe (2018). *Common European Framework of Reference for languages: Learning, teaching, assessment: Companion volume with new descriptors*. Council of Europe.
- Ducasse, A., & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing*, 26(3), 423–443.
- Field, J. (2011). Cognitive validity. In L. Taylor (ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 65–111). UCLES/Cambridge University Press.
- Fulcher, G. (2010). *Practical language testing*. Routledge.
- Galaczi, E., & Taylor, L. (2018). Interactional Competence: Conceptualisations, Operationalisations, and Outstanding Questions. *Language Assessment Quarterly*, 15:3, 219–236.
- Galaczi, E., & Taylor, L. (2020). Measuring Interactional Competence in Taylor Winke, P., & Brunfaut, T. (Eds.). *The Routledge handbook of second language acquisition and language testing*.

- Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, 27(4), 585–602.
- Grabowski, K. (2013). Investigating the construct validity of a role play test designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In: Ross S.J., Kasper G. (eds) *Assessing Second Language Pragmatics*. Palgrave Advances in Language and Linguistics. Palgrave Macmillan.
- Hall, J.K. (1995). Aw, man, where you goin? Classroom interaction and the development of L2 interactional competence. *Issues in Applied Linguistics*, 6, 37–62.
- Hall, J. K., & Pekarek Doehler, S. (2011). L2 interactional competence and development. In J. K. Hall, J. Hellermann, & S. Pekarek Doehler (Eds.), *L2 interactional competence and development* (pp. 1-18). *Multilingual Matters*.
- He, A., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. E. Young & A. He (eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24).
- Hu, J. (2015). Interaction in Assessment-Oriented Role Play: A Conversation Analytic Approach. *Chinese Journal of Applied Linguistics*, 38(4), 472-489.
- Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth.
- Ikeda, N. (2017). Measuring L2 oral pragmatic abilities for use in social contexts: Development and validation of an assessment instrument for L2 pragmatics performance in university settings [Unpublished doctoral dissertation]. University of Melbourne, Australia.
- Johnson, M. (2001). *The art of nonconversation*. Yale University Press.
- Johnson, M., & Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? In R. Young & A. He (eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 27–51). John Benjamins.
- Kasper, G. (2006). Beyond repair: Conversation analysis as an approach to SLA. *AILA Review*, 19, 83-99.
- Kasper, G. & Youn, S. (2018). Transforming instruction to activity: Roleplay in language assessment. *Applied Linguistics Review*, 9(4), 589-616.
- Kim, H. (2012). Exploring the construct of radiotelephony communication: A critique of the ICAO English testing policy from the perspective of Korean aviation experts. *Language Testing and Assessment*, 2 (2), 103-110.
- Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing*, 16(2), 163–188.
- Kramsch, C. (1986). From Language Proficiency to Interactional Competence. *The Modern Language Journal*, 70(4), 366-372.
- Lam, D. M. K. (2018). What counts as “responding”? Contingency on previous speaker contribution as a feature of interactional competence. *Language Testing*, 35(3), 377–401.
- LanguageCert, (2020). *LanguageCert International ESOL Qualification Handbook (Speaking)*. <https://www.languagecert.org/en/preparation/practice-material/languagecert-international-esol>
- LanguageCert, (2021). *Assessing Speaking Performance*. <https://www.languagecert.org/en/preparation/practice-material/languagecert-international-esol>
- Markee, N. (2019). Some Theoretical Reflections on the Construct of Interactional Competence.

- May, M. (2011) Interactional Competence in a Paired Speaking Test: Features Salient to Raters, *Language Assessment Quarterly*, 8:2, 127-145.
- McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, 18(4), 446-466.
- Nakatsuhara, F. (2014). The Co-construction of Conversation in Group Oral Tests. Peter Lang
- Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the Use of Video-Conferencing Technology in the Assessment of Spoken Language: A Mixed-Methods Study. *Language Assessment Quarterly*, 14, 1 - 18.
- Nakatsuhara, F., May, L., Lam, D., & Galaczi, E. D. (2018). Learning oriented feedback in the development and assessment of interactional competence. *Research Notes*, 70.
- Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, 32(1), 39–62.
- Okada, Y. (2010). Role play in oral proficiency interviews: Interactive footing and interactional competencies. *Journal of Pragmatics*, 42, 1647–1668.
- Oksaar, E. (1990). Language contact and culture contact: Towards an integrative approach in second language acquisition. In H. Dechert (Ed.), *Current trends in European second language acquisition research* (pp. 230–43). Clevedon: Multilingual Matters.
- Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System*, 30(2), 143–154.
- O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277-295.
- Pekarek Doehler, S. (2019). On the nature and the development of L2 interactional competence: State of the art and implications for praxis. In M. R. Salaberry & S. Kunitz (Eds.), *Teaching and testing L2 interactional competence: Bridging theory and practice* (25–59). Routledge.
- Plough, I. (2018). Revisiting the speaking construct: The question of interactional competence. *Language Testing*, 35(3), 325–329.
- Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, 35(3), 427-445.
- Roever, C., & Ikeda, N. (2021). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing*.
- Roever, C. & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking.
- Seedhouse, P. (2012). What kind of interaction receives high and low ratings in oral proficiency interviews? *English Profile Journal*, 3, 1–24.
- Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, 23(4), 411–440.
- Vo, S. T. (2019). Effects of task types on interactional competence in oral communication assessment. (Unpublished doctoral dissertation). Iowa State University, United States.
- Walters, S. F. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing* 27(2). 155–183.
- Walters, S. F. (2013). Interfaces between a discourse completion test and a conversation analysis informed test of L2 pragmatic competence. In Steven J. Ross & Gabriele Kasper (eds.), *Assessing second language pragmatics*, 172–195. Basingstoke: Palgrave Macmillan.

- Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing* 32, 199–225.
- Youn, S. J. (2019). Interactional Features of L2 Pragmatic Interaction in Role play Speaking Assessment.
- Youn, S. J. (2020). Managing proposal sequences in role play assessments.
- Young, R. F. (2000). Interactional competence: Challenges for validity. Paper presented at the annual meeting of the American Association of Applied Linguistics, Vancouver, Canada.
- Young, R. F. (2008). *Language and interaction*. Routledge.
- Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (2) 426–443). Routledge.
- Young, R. F. (2019). Interactional competence and L2 pragmatics. *The Routledge handbook of second language acquisition and pragmatics*, 93, 110.
- Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14, 403–424.
- content.

Appendix 1: Practice Paper 2 of LanguageCert International ESOL (Speaking), B1 level, Part 2

LanguageCert Achiever B1

PART 2 (3 minutes)

I: Now, Part Two. We are going to role-play some situations. I want you to start or respond. First situation (*choose one situation from A*).

A

- We're friends. I start.
I'm training for a race. Do you want to join me?
- I work for the local government. I start.
Good morning. I'm doing some research. Are you happy with local leisure facilities?
- I was your teacher last year. I start.
Hello, nice to see you again. How's your new class?
- I'm a stranger in your town. I start.
Excuse me, do you know of a good place to eat near here?

C: (*Responds.*)

I: (*Role-play the situation with the candidate – approximately two turns each.*)

I: Second situation (*choose one situation from B*).

B

- I work at a library. You want some help. You start.
- We're neighbours. You want me to look after some things while you are away. You start.
- I work at the airline baggage desk. Your bag is missing. You start.
- I work in a shop. I've given you the wrong change. You start.

C: (*Initiates.*)

I: (*Role-play the situation with the candidate – approximately two turns each.*)

I: (*Role-play a third situation from A or B if time allows.*)

I: Thank you.

Note: Reprinted from <https://www.languagecert.org/en/preparation/practice-material/languagecert-international-esol>

Appendix 2: Practice Paper 6 of LanguageCert International ESOL (Speaking), C1 level, Part 2

LanguageCert Expert C1

PART 2 (3 minutes)

I: Now, Part Two. We are going to role-play some situations. I want you to start or respond.
First situation (*choose one situation from A*).

A

- We're flatmates. I start.
Why on earth do you keep tidying up? I can't find my laptop yet again.
- We're classmates. I start.
What do you think of our new teacher? She seems a bit strict to me.
- We're neighbours. I live next door. I start.
Your child's football has made a mark on my fence again.
- You're visiting a museum. I'm the security guard. I start.
Excuse me. I've told you twice already – please don't stand so close to the paintings.

C: (*Responds.*)

I: (*Role-play the situation with the candidate – approximately two turns each.*)

I: Second situation (*choose one situation from B*).

B

- I'm your boss. You think you deserve a pay raise for your hard work. You start.
- We're flatmates. I haven't paid my share of the rent for the last three months. You start.
- I'm a plumber. I've recently done some work in your house and you're totally dissatisfied. You start.
- We're best friends. I've just told you I've been made redundant. You start.

C: (*Initiates.*)

I: (*Role-play the situation with the candidate – approximately two turns each.*)

I: (*Role-play a third situation from A or B if time allows.*)

I: Thank you.

Appendix 3: CEFR Interaction scale across the CEFR levels

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like "and" or "then".
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like "and," "but" and "because".
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he / she needs to, though he /she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skillfully to those of other speakers.	Can produce clear, smoothly-flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.

Note. Reprinted from <https://www.coe.int/en/web/common-european-framework-reference-languages/table-3-cefr-3.3-common-reference-levels-qualitative-aspects-of-spoken-language-use>



Chapter 12: Validating Communicative Tests of Reading and Language Use of Classical Greek

David Coniam, Polyxeni Poupounaki-Lappa and Tzortzina Peristeri

Abstract

This chapter builds on previous work by Poupounaki-Lappa et al. (2021), which described the development of a communicative test of Reading and Language Use of Classical Greek, calibrated to the Common European Framework of Reference (CEFR) at levels A1 and A2 (Council of Europe, 2001). In this chapter, the two tests of Classical Greek are calibrated both together and to the CEFR. In addition to describing the methodology for comparing the two separate tests of Classical Greek, the chapter is also designed to be of interest to educators of other classical languages. It is hoped that they may find it useful not only in facilitating robust test design, but also by demonstrating the methods by which tests can be linked together on a common scale (as with the CEFR) or by linking tests one to another (e.g., different end-of-year tests, at different points in time).

Key words: Classical Greek, reading and language use, assessment, Rasch, test linking

Introduction

This chapter builds on the groundwork presented in Poupounaki-Lappa et al. (2021), which described the development of a communicative test of Reading and Language Use of Classical Greek calibrated to the Common European Framework of Reference (CEFR) at levels A1 and A2 (Council of Europe, 2001) [Note 1].

The test outlined in Poupounaki-Lappa et al. (2021) discussed issues – in line with more ‘communicative’ approaches to language teaching (see Richards and Rodgers, 2014; Lloyd and Hunt, 2021) – related to the creation of a communicative testing system for Classical Greek, initially centring around Reading and Language use. Appendix 1 presents samples of the constructs assessed in terms of reading skills, grammar and syntax, and topics in the two tests.

The major focus of the chapter involves calibrating – both to one another and to the CEFR – two tests of Classical Greek. In this light, the assessment element of the chapter’s methodology may be seen to extend beyond its Classical Greek dimension. The study will hopefully be of interest to educators of other classical languages who may also find it useful not only regarding robust test design, but also regarding linking tests together on a common scale (as with the CEFR) or in linking tests one to another (e.g., different end-of-year tests, at different points in time).

The test development process described in Poupounaki-Lappa et al. (2021) involved the construction of two tests of Reading and Language Use, with each test consisting of four parts, each using distinct task types to assess specific sub-skills. Figure 1 elaborates.

Figure 1: Classical Greek task types

Part 1: 10 items: Multiple matching	(images and words)
Part 2: 10 items: True/False	(statements with visuals)
Part 3: 10 items: Multiple-choice items	(gapped text)
Part 4: 10 items: Multiple matching	(gapped text)

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, as well as those in Appendix 1, are drawn from this Qualification Handbook.

Piloting

The two tests were piloted in mid-2021, administered to groups of test takers who were judged to be at the intended level of language proficiency by the subjects’ teachers. This chapter presents details of the tests, and their match with the supposed target levels of A1 and A2.

Key test qualities are validity and reliability (Bachman and Palmer, 2010). With regards to validity, central issues include how well the different parts of a test reflect what a test taker can do, and how well test scores provide an indication of test taker communicative ability (Messick, 1989; Bachman and Palmer, 2010). The Classical Greek tests assess what test takers will be expected to have control over at particular levels of ability (i.e., in relation to the CEFR). Against such a backdrop, test content needs to match target test takers’ levels in terms of grammar, functions, vocabulary and topics.

As a starting point, Morrow (2012) outlines the rationale and aims of communicative language tests, stating that the aim of a communicative language test is to find out what a learner can do in the language. Moving to practicalities, if a communicative test is to be valid and reliable, it nonetheless needs to be well constructed. In addition to validity, some of the features of a 'good' test (see Hughes, 2003) are defined as tests being reliable and at an 'acceptable' level of difficulty.

Test difficulty needs to be considered from two perspectives. One, that it matches the ability level of the intended target group; but two, that it is sufficiently discriminating to permit the exam body (or teacher etc.) to be able to confidently make decisions about the extent to which test takers have met the language competencies required for the particular level, the pass mark and the grade they should be awarded.

Statistical Analysis

In the current study – to gauge test fitness for purpose, and to link two different tests to a common scale – two types of statistical analysis have been performed. The first of these involves classical test statistics, reporting test mean and test reliability. The second involves the use of Rasch measurement which serves the purpose of calibrating the two tests together.

Detail on classical test statistics can be found in the Glossary of statistical terms and techniques. In terms of test reliability – where levels of reliability are associated with test length (Ebel, 1965) – expected reliability with a 40-item test is in the region of 0.67.

An overview of the methodology surrounding Rasch is also provided in the Glossary, along with an outline of the infit and outfit mean square statistics which are key to the interpretation of Rasch results in the context of data 'fit'.

In the current Classical Greek study, the tests developed for A1 and A2 needed to be linked to one another, so that items could be placed on a common scale. The two tests have therefore been linked via a set of common items (the cloze passage in Part 3) (Bond, Yan and Heene, 2020).

To adequately validate a test (or tests) nonetheless requires some form of external triangulation or confirmation beyond the test. To this end, the single scale produced through Rasch measurement in the current project has been validated by a number of test takers who completed a set of Can-do statements (see Appendix 2) ranging from Pre-A1 to B1+ levels. These self-assessments were then regressed against test scores, providing evidence for the validity of the test constructs through the test takers' judgements of their own abilities. Detail on the Can-do statements and the validation procedure is reported below.

Data and Analysis

As mentioned, two tests of 40 items were constructed with a 10-item MC cloze passage common to both tests. One test, at intended A1 level, and another, at intended A2 level, were administered in spring 2021. It had been hoped that about 150 subjects from a variety of first language backgrounds would take each pilot test. Due to Covid-19 pandemic restrictions, however, sample sizes were consequently smaller, with most subjects being first language speakers of modern Greek. Sample sizes were, however, large enough for statistical analyses to be performed.

Classical Test Statistics

This section briefly describes key classical test statistics.

Table 1 first presents test means and reliabilities.

Table 1: Test means and reliabilities

	A1	A2
Test takers	74	89
Mean	28.9 (72%)	30.3 (76%)
Standard deviation	4.4 (11.0%)	6.2 (15.5%)
Reliability	0.72	0.88

For 40 items, both test means were in the desirable range – in the 70 percent range. This suggests that the tests broadly fit the target population, and that most test takers finished the test and had given it their best shot. Test reliability for both tests was above 0.67 indicating that the tests may be assumed to have been well constructed. The spread of ability (indicated by the standard deviation) was narrower in the A1 cohort of test takers.

Table 2 now presents the picture of means in each of the four subtests. Each subtest comprised 10 items, with Part 3 the common section in both tests.

Table 2. Subtest means (Max=10 on each subtest)

Part	A1	A2	Subtest type	Note
1	94%	86%	Matching words to pictures	
2	82%	81%	True/False statements with visuals	
3	59%	76%	Multiple-choice cloze passage	Common section
4	47%	60%	Multiple matching gapped text	
O'all mean	72%	76%		

As had been intended, a cline of difficulty emerged, with subtests increasing in difficulty from one subtest to the next. Part 1, Matching words to pictures emerged as very easy, with means close to or above 90%. Part 2, True/False statements with visuals emerged as comparatively easy, with means in the 80% range. Part 3, the Multiple-choice (MC) Cloze passage common to both tests emerged with a mean of 59% for the A1 cohort and 76% for the A2 cohort – an indication that the two groups were of differing ability. Part 4, the Multiple matching gapped text exercise, emerged as the most difficult. The cline of difficulty may also be seen as a reflection of intended functional demands. Part 1 involved vocabulary and items were discrete. Part 4 required that test takers operate at the more complex text level – constructing a text by matching the two lists of ten possibilities.

Raw scores are a baseline indication of how ‘good’ a test may be deemed. If comparisons are to be made across tests, however, or if tests are to be calibrated together, Rasch measurement needs to be employed, and it is to this statistic that the discussion now turns.

Rasch Measurement

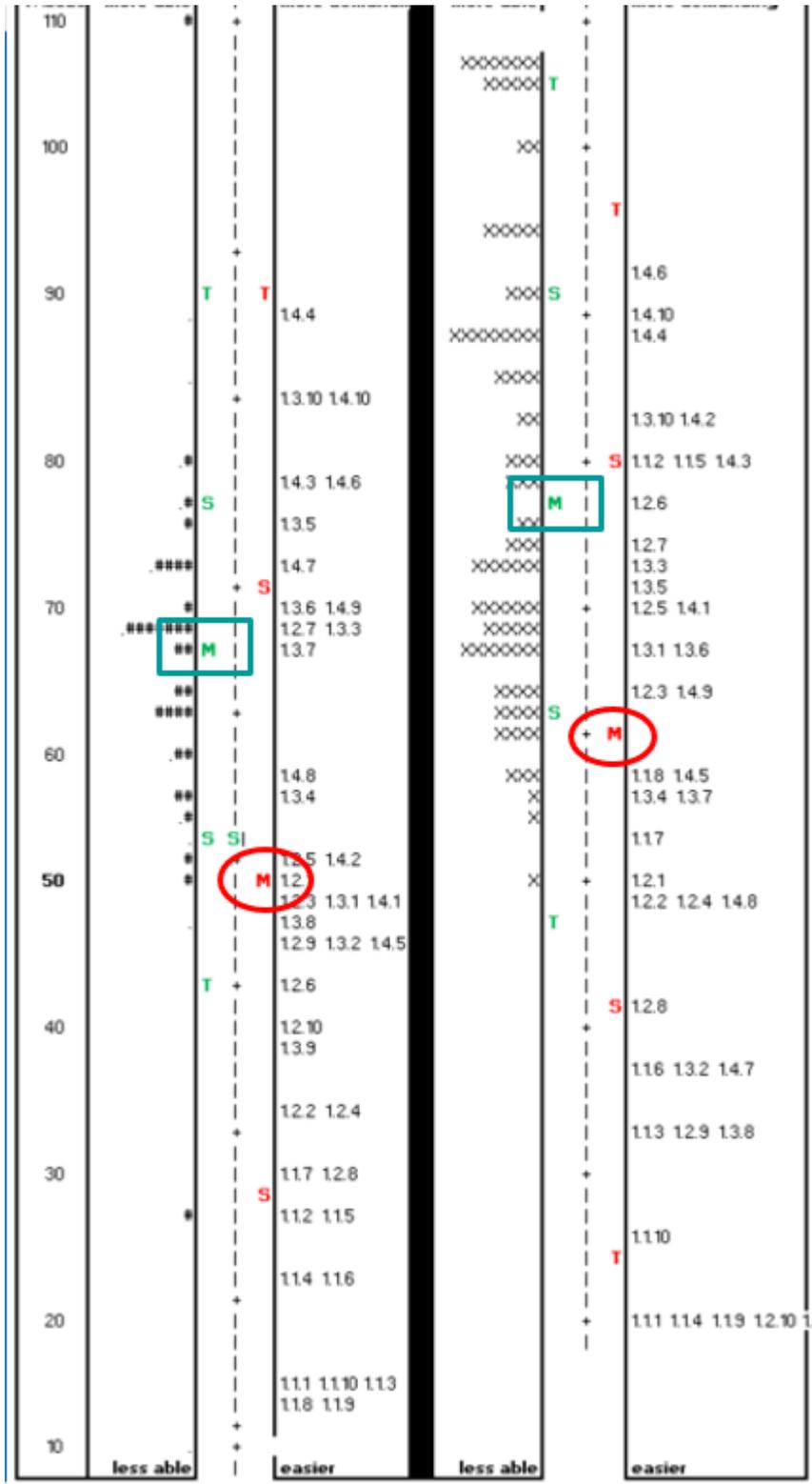
In interpreting Rasch, the key statistic involves the ‘fit’ of the data in terms of how well obtained values match expected values (Bond et al., 2020). A perfect fit of 1.0 indicates that obtained mean square values match expected values one hundred percent [Note 2]. Acceptable ranges of tolerance for fit range from 0.5 through 1.0 to 1.5 (Lunz and Stahl, 1990). The outcomes from the Rasch analysis confirm, from a different perspective, the classical test statistics results that have been presented above. Both sets of analyses underscore and add to an appreciation of the baseline robustness of the two tests.

To provide an overview of the Rasch measurement technique, the vertical ruler (the ‘facet map’) produced in the Rasch output is presented below. The facet map is a visual representation of where the facets of test takers and items are located on the Rasch scale. Figure 2 below presents the map for tests A1 and A2 as seen when calibrated together. The results for the A1 test takers and items appear to the left-hand side of the Figure, while the A2 test takers and items appear to the right-hand side. Test takers are represented as asterisks or crosses, and items are represented by the item numbers.

The map should be interpreted as follows. For each test, the top left-hand side of the map indicates more able test takers; in a similar manner, the top right-hand side represents more difficult items. Conversely, less able test takers appear to the bottom left-hand side of the map, and easier items to the bottom right-hand side. The green rectangles indicate the test taker midpoints, while the red ovals indicate the item midpoints.

To ease interpretation, Rasch measures (‘logits’) in the study have been rescaled to a mean of 50 with a standard deviation of 10.

Figure 2: Facet map



As can be seen from Figure 2, the midpoint for the A1 items is 50, whereas the midpoint for the A2 items is 62. This indicates one logit of difference (10 points) between the items; the A2 items, as had been intended, have emerged as more demanding. Turning to test takers, the A1 test takers have a mean of 67, while the A2 test takers have a mean of 78. This clearly indicates that the A2 test takers are more able than the A1 cohort, again by one logit, or 10 points.

The bottom line has thus been satisfied in two respects: a) the test items differentiate between tests; and b) test taker cohorts may be seen to be of increasing ability. The item / test taker match, however, is less than optimal.

Test takers are in a comparatively narrow range. Ignoring outliers, the A1 test takers are in a three-logit range from 50 to 80; the A2 cohort show a rather wider range from 60 to 105, a 4.5 logit range. While the A1 items cover a wide difficulty range, many items – as can be seen from the map – are below 50, the bottom end of test taker ability. These items are too easy since they do not match with any test taker abilities, and consequently return no useful assessment ‘information’ on test takers. In future live tests, an attempt will be made to address this situation: the number of very easy items will be reduced, with a view to working towards a closer test taker ability / item difficulty match.

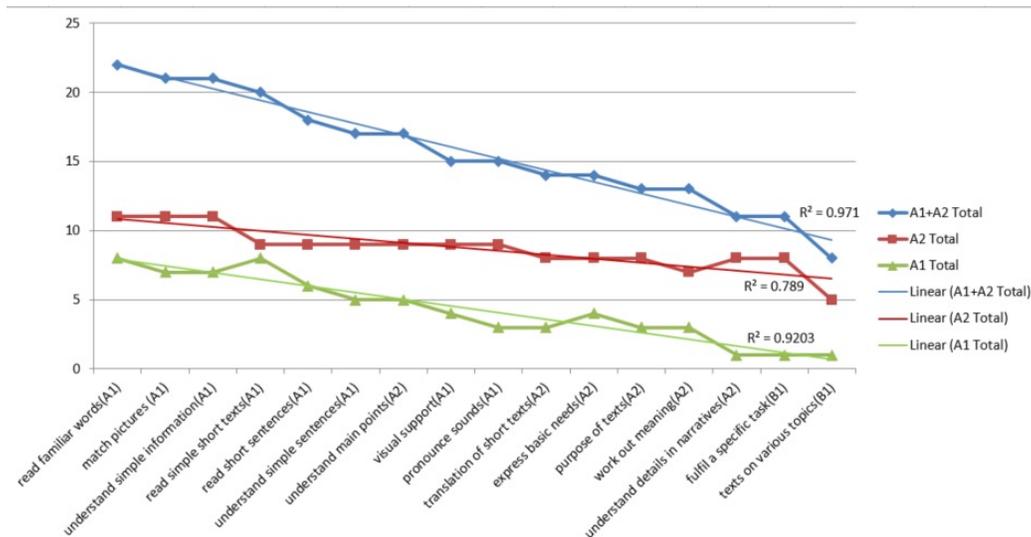
A similar, although less exacerbated, situation exists with the A2 test, where there is still a number of very easy items towards the bottom right-hand end of the map. Similar attention will be paid in order to redress the imbalance in this test.

Triangulating Test Results

As mentioned, test takers who took the two tests were subsequently approached – by email, following their consent to be contacted – and asked to complete a survey. This consisted of a series of Self-assessment Can-do statements, adapted from CEFR material for other languages. The use of instruments such as Can-do statements in self-assessment has been validated in a number of studies (see Zhao and Coniam, Chapter 9 this volume). In the current study, there were 16 items, with intended difficulty levels ranging from low A1 (at the left-hand end of the figure) to high A2/low B1 (at the right-hand end of the figure) – see Appendix 2. Respondents were asked to rate themselves for each item on a six-point scale (‘6’ being high) in order to demonstrate whether they felt they could master the requisite skill.

The survey was completed by only a small number of test takers (12 for A1; 15 for A2), so results may only be seen as indicative. Figure 3 presents the results of regressing the different Can-do statements against test scores. In Figure 3, the blue diamonds are the responses of the A1 group, the red squares those of the A2 group, and the green triangles those of the combined groups. R2 indicates the amount of variance accounted for by the regression.

Figure 3. Self-assessment Can-do statements regressed against test score



Key: Linear = Regression line of best fit

Figure 3 illustrates a clear match between test takers' perceived abilities in reading and usage in Classical Greek, and their actual scores on the relevant test. The R^2 values to the bottom right of the regression line indicate a close fit between perceived abilities and test scores. Despite the small sample sizes, Figure 3 provides additional external validity evidence for the communicative traits underpinning the two tests.

Conclusion

This chapter is a sequel to that introduced in Poupounaki-Lappa et al. (2021) which described the development of a communicative test of Reading and Language Use of Classical Greek calibrated to the CEFR at levels A1 and A2. Both A1 and A2 level tests comprised four parts, with the four parts designed to produce a cline of difficulty, progressing from vocabulary recognition at the lowest level through to text-based exercises in the later part of the tests. For calibration purposes, one part (the MC cloze passage) was common to both tests.

In addition to reporting on the further development, administration and analysis of the two tests and the robustness of their validity and reliability, the chapter has described an important feature of test development: the calibration, separately and then together, of the two tests of Reading and Language Use produced for levels A1 and A2.

The first part – matching words to pictures – emerged as very easy, possibly too easy, on both tests. The second part – matching True/False statements to visuals was also comparatively easy. The third part, common to both tests, was a cloze passage requiring test takers to make lexical / grammatical / syntactic contextual fits. The fourth and final part, which proved to be the most demanding, was a multiple-matching gapped text exercise, which required test takers to make sense of a whole text.

The tests were administered to comparatively small cohorts of test takers who had been estimated by their teachers to be at the approximate level for whichever level of test that they took. Subsequent to taking the test, test takers were approached and asked to complete a self-assessment of CEFR-linked Can-do statements reflecting abilities from pre-A1 to B1+ level.

As a baseline, based on classical test statistics, both tests were reliable; and means scores were acceptable overall – even if some parts of the test were very easy. As a result, difficulty levels of some parts will need to be reconsidered. The cloze passage common to both tests indicated that the two cohorts were different and that the tests could be linked. Linking together was then achieved via the use of Rasch measurement, where fit statistics were good and the two tests were successfully calibrated and linked together on a common scale.

The regressing against test scores of test takers' self-assessments on the Can-do statements enabled a degree of triangulation to be conducted, providing an external validation of the fit of the test to the target population. While the A2 test was seen to be pitched at a higher level to the A1 test, test results suggest that the difference between the two tests needs to be extended when future tests are produced.

In addition to presenting the development and analysis of the trialling of the two tests, the main focus in the current chapter has centred around calibrating two tests of Classical Greek to each other and to the CEFR, an important issue in effective test development. Looking beyond the study's immediate Classical Greek focus, however, it is hoped that the methodologies outlined may also be considered useful from a more general perspective, and may interest educators and teachers of other classical languages who wish to consider developing good tests. Such a 'more general' assessment perspective may involve the construction of different tests which need to be linked to other tests – possibly via a common scale of ability – or simply that of different tests being analysed together so that direct comparisons may be made between different test taker cohorts, for example.

Limitations

Two limitations alluded to in the study will now be discussed, with both linked to the Covid pandemic and restrictions imposed on the administrators and test takers just as the piloting of the two tests was scheduled.

The first limitation related to sample size, which had been projected to be in the region of 150 or so for each test. Ultimately, a sample of only approximately half that number was achieved. The second limitation was mother tongue, with the sample originally projected as having an international perspective, comprising subjects with a range of mother tongues. Again, as a result of the Covid lockdown, this was not achieved, and the mother tongue of over 90% of test takers was modern Greek, with the test takers based in Greece.

The A1 and A2 tests have now gone live. As data becomes available with the administration of the live tests, it is anticipated that the analysis conducted in the current study will be revisited. Further, given that the Classical Greek Reading and Language Use test may be taken from anywhere worldwide via LanguageCert's Online Proctoring facility (<https://www.languagecert.org/en/welcome>), the sample will be further adapted to an in-

ternational audience as more test takers take the test over time. By this means, further data analysis will be conducted to see whether the findings of the original study are replicated and made more generalisable.

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>.

2. The mean square of a set of values is the mean of the squared differences between the mean and the values from which the mean is calculated. The reason for the squaring in the calculation is due to the fact that the sum of the actual differences between the mean and the values from which the mean is calculated would always be zero.

References

- Bachman, L. & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press: Oxford, UK.
- Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language testing*, 6(1), 14-29.
- Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Milton Park, UK: Routledge.
- Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals*, 47(2), 261-285.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). *Validating the LanguageCert Test of English scale: The paper-based tests*. London, UK: LanguageCert.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg Cedex, France: Council of Europe.
- Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, 11(2), 186-197.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press: Cambridge, UK.
- LanguageCert.org (2021) Exam Information. <https://www.languagecert.org/en/language-exams/classical-greek>.
- Lloyd, M. E., & Hunt, S. (eds.) (2021). *Communicative approaches for ancient languages*. London: Bloomsbury.
- Messick, S. (1989). Validity. In R. L. Linn (ed.) *Educational measurement*. 3rd ed. New York: Macmillan. 13-103.
- Morrow, K. (2012). Communicative language testing. *The Cambridge guide to second language assessment*, 140.
- 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).

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Summers, M. M., Cox, T. L., McMurry, B. L., & Dewey, D. P. (2019). Investigating the use of the ACTFL can-do statements in a self-assessment for student placement in an Intensive English Program. *System*, 80, 269-287.

Appendix 1: Assessment Construct Samples in Classical Greek A1 and A2 Reading and Language Use Tests

<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 	
<p>Grammatical and syntactic features</p> <ul style="list-style-type: none"> • alphabet • syllables – accentuation • parts of speech / syntax • verb forms • nouns • pronouns 	<ul style="list-style-type: none"> • prepositions • articles • adjectives • infinitives • participles
<p>Topics</p> <ul style="list-style-type: none"> • Personal identification • House and home, environment • Daily life • Free time, entertainment • Activities 	<ul style="list-style-type: none"> • Relations with other people • Health and bodycare • Food and drink • Places • Weather

Appendix 2: Self-assessment Can-do Statements for CEFR levels Pre-A1 to B1+ for Reading and Language Use in Classical Greek

Adapted Can-do Statements	CEFR Level
I can recognise and read familiar words.	Pre-A1
I can pronounce the sounds and the stress on simple, familiar words and phrases.	Pre-A1
I can understand simple sentences, if I read them slowly, several times.	Low A1
I can match words and sentences with pictures.	Low A1
I can understand simple information if there is visual support.	Mid A1
I can read simple short texts and understand familiar words and phrases.	Mid A1
I can read short simple texts and understand simple information.	High A1/low A2
I can read short sentences with familiar and unknown words.	High A1/low A2
I can understand the main points and locate specific information in short simple texts on familiar matters.	Mid A2
I can use simple language to provide an approximate translation of short texts on familiar topics that contain high frequency words.	Mid A2
I can adapt well-rehearsed memorised simple phrases changing a few words to express basic needs.	High A2
I can work out the probable meaning of unknown words from the context.	High A2
I can locate and understand details in narratives, descriptions, and informative texts on familiar topics.	B1
I can understand the purpose of different texts.	B1
I can understand a wide range of longer texts on various topics.	B1+
I can get information from different parts of a text, or from different texts in order to fulfil a specific task.	B1+

Glossary of Statistical Techniques Used in the Volume

Peter Falvey

Much of this section is adapted from Coniam and Falvey (2018: 125-156). 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 thirteen 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]).

Equivalence Independent Samples T-Test

The equivalence independent samples t-test permit users to test the null hypothesis that the population means of two independent groups fall inside a user-defined interval, i.e., the equivalence region. The procedure of using two-one-sided tests (TOST) permits significance to be observed via specified upper and lower bounds, as opposed to standard t-tests which report a single t score (see Lakens, 2017). The upper and lower bounds represent the extent of variation of t values regarding the two populations of the two samples being tested. If the t value of the equivalence test is within the estimated range, the two populations may be deemed to be equivalent.

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.

McDonald's Omega

McDonald's, or coefficient, omega is based on the estimated association between a unidimensional underlying or latent variable: within the context of a one-factor confirmatory factor model, and a group of assessment results of a sample of candidates. Unlike Cronbach's alpha, omega can be used to estimate reliability in situations where tests are not unidimensional and are not within the same frame of reference of measurement, (that is, they are not necessarily measuring the same latent trait) and where test items are not tau-equivalent (i.e., all items having equal covariance with the true score). The implementation of coefficient omega with a Bayesian perspective extrapolates the probability of the estimated reliability coefficient regarding its stability in the future.

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.

Kappa

Cohen's Kappa is a statistical measure for examining the agreement between two rated categories. It aids in determining the implementation of a given coding system.

Kappa helps to assess levels of agreement between two variables. According to Landis and Koch (1977), a level of 0.21 – 0.40 for kappa indicates 'fair agreement', 0.41 – 0.6 'moderate agreement', 0.61 – 0.8 'substantial agreement', and 0.8 or better 'strong' agreement.

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).

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). 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 and 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 and 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

A key statistic in the interpretation of Rasch results is that of data 'fit', which relates to how well obtained values match expected values (Bond et al., 2020). Broad criteria in assessing model fit are the Infit and Outfit mean square statistics (i.e., estimates of population variance, or standard error).

Infit is generally seen as the 'big picture' in that it scrutinises the internal structure of an item. High infit values indicate rather scattered information within an item, providing a confused picture about the placement of the item. Outfit gives a picture of 'outliers' – responses from items which appear to be out of line with where an item would expect to be located.

For both infit and outfit, 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 0.5 for the lower limit to 1.5 for the upper limit (Lunz and Stahl, 1990). 1.5 to 2.0 is considered just about acceptable, with figures beyond 2.0 unacceptable.

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.

Bayesian Statistics

Bayesian statistical methods describe the conditional probability of an event based on data as well as prior information or beliefs about the event, with probabilities computed and updated after obtaining new data – see Andraszewicz et al. (2015).

Since Bayesian statistics treat probability as a degree of belief, permitting inferences about future events to be estimated in a positive way – rather than simply of failure to reject the alternative hypothesis, as in standard statistical testing.

In Bayesian statistics, the critical statistic is the Bayes Factor (BF) – the ratio of likelihood between the null and the alternative hypothesis. Jeffreys (1961) proposes cutoff levels for interpreting the strength of Bayes Factors, recommending cutoff levels ranging from 1 (no evidence for the alternative hypothesis) to 10-30 (strong evidence), to 30-100 (very strong evidence), to > 100 (extreme evidence for the alternative hypothesis).

The credible interval is the Bayesian statistics version of the standard (“frequentist”) statistics confidence interval. The credible interval represents the spectrum in which a specified percentage, e.g., 95%, of cases

would fall. It has a direct interpretation as “the probability that p is in the specified interval” (Hoekstra et al., 2014).

References

- 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*. Second edition. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Coniam, D. & Falvey, P. (eds.) (2018). *High-stakes testing: The impact of the LPATE on English language teachers in Hong Kong*. Springer Nature: Singapore.
- 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, Katherine. (1999). Interpreting correlation coefficients. <http://acad.cl.uh.edu/itc/educ6032/course/resources/unit2/index.htm> (8 October 1999).
- 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).
- 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.
- Jeffreys, H. 1961. *Theory of probability* (3rd ed.). New York: Oxford University Press.
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4), 355-362.
- 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.
- Schulz, W., & Fraillon, J. (2011). *The analysis of measurement equivalence in international studies*.
- 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*. Houndmills, 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 and The Department for Instructional Science. Brigham Young University. <https://testing.byu.edu/handbooks/bettertests.pdf>.

ISBN: 978-9925-34-309-6



9 789925 343096