CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

Research and Validation at LANGUAGECERT

Volume 3

Edited by Leda Lampropoulou and David Coniam

LANGUAGECERT®

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING
Research and Validation at LANGUAGECERT Volume 3

Leda Lampropoulou, PeopleCert, London, UK
David Coniam, PeopleCert, London, UK

Publisher details: LANGUAGECERT, London, UK
Date of Publication: February 2025
ISBN: 978-9925-34-960-9

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

Research and Validation at LANGUAGECERT

Volume 3

Edited by Leda Lampropoulou and David Coniam

Foreword

Byron Nicolaides, CEO, PeopleCert Group

2025 marks the 25th anniversary of PeopleCert. 25 years of growth, innovation and success shaping the future of learning. 25 years of enabling others to chase their dreams of a better life, a fulfilling career and a brighter tomorrow.

2025 also marks the tenth anniversary of LANGUAGECERT. A major milestone, and an opportunity to reflect on, recognise and celebrate the steps we have taken and the progress we have made in our mission to provide high-quality, best-in-class language assessments. This third volume in our research series, 'Certifying Quality in Assessment and Learning', is a testament to and a record of the LANGUAGECERT commitment to research and validation.

The LANGUAGECERT mission is open-ended. It has a beginning, but we can never claim it has a final, culminating end point, because, like time, quality does not stand still. For any language assessment to be considered high quality it must be fit for purpose and meet the needs of different stakeholders. These needs and expectations change, technology changes and the world changes. This permanent state of change requires LANGUAGECERT to continuously innovate and improve, and one of the ways we do this is through our ongoing programme of research and validation.

Research is how we define what makes an assessment fit for purpose and validation is the process of ensuring a test is and continues to be fit for purpose. Our research looks at the present and future real-world language skills needs of test-takers and stakeholders such as institutes of higher education, employers and governments. Research is instrumental in developing and delivering tests that have a positive impact on test-takers' language acquisition and tests that integrate with pedagogical practice and school and national curricula.

All language tests must be fair and accessible to everyone and free of bias. Equality, diversity and inclusion (EDI) are foundational to the design, development and delivery

of LANGUAGECERT tests. All items (questions) are designed to elicit specific language skills, calibrated to the respective level and to be free of bias. After a test is taken, our validation processes use various statistical methods to analyse whether and how the test's respective items have performed as intended. This practice enables LANGUAGECERT to ensure that all our tests are fair and consistent across a diverse candidature and over time.

This volume showcases the breadth and depth of our research and validation output. It shows the steps we have taken on our mission, and points the way to the next ten years and beyond. LANGUAGECERT prides itself on its commitment to research and validation, and I am proud of the increasing contribution we are making to the body of wider assessment knowledge and the increasing role we are playing in the global assessment community.

I would like to thank all the contributors for their hard work on the design and development of our tests, and I would also like to thank our greatly valued network of partners worldwide who make test delivery possible.

Preface

Marios Molfetas, Chief Languages Officer and Series Editor

In his preface, Byron Nicolaides states: 'This third volume in our research series, 'Certifying Quality in Assessment and Learning', is a testament to and a record of the LANGUAGECERT commitment to research and validation'. In 2024, this commitment was recognised when LANGUAGECERT became a full member of the Association of Language Testers in Europe (ALTE).

ALTE was founded in 1989. Its membership consists of language test providers representing 26 European languages who work together to promote the fair and accurate assessment of linguistic ability across Europe and beyond.

To become a full member, the LANGUAGECERT Test of English (LTE) had to pass a comprehensive audit to attain the ALTE Q-mark. According to the Association of Language Testers in Europe the mark *'is a quality indicator which member organisations can use to show that their exams have passed a rigorous ALTE audit and meet all the core requirements of ALTE's 18 quality standards. The Q-mark demonstrates that ALTE member organisations aspire to consistent standards of quality and excellence in their exams'*.

One of the ways LANGAGECERT achieves consistent quality standards and excellence is by the ongoing alignment, through robust calibration and validation practices, of LANGUAGECERT tests to international standards — particularly the Common European Framework of Reference (CEFR). The following chapters on the LANGUAGECERT Global Scale demonstrate our commitment to maintaining test quality by external anchoring and alignment, with meticulous attention to the separate skills of Reading, Listening, Writing and Speaking across levels of communicative language ability.

Our alignment to the CEFR is confirmed by Ecctis, one of the world's leading providers of services in the recognition and evaluation of qualifications and skills: *'Independent review of LanguageCert tests against the Common European Framework of Reference for Languages (CEFR) has found that the LanguageCert General test provides a sound*

assessment of English language competence at CEFR levels A2-C1 and that the LanguageCert Academic test assesses CEFR levels B1-C2.' (Ecctis, 2023).

Two of the many features that make a test fit for purpose are accessibility and security. 2024 was a landmark year for both when we introduced an online, at-home option for LANGUAGECERT Academic: a four-skill multi-level English test offering choice to undergraduate and postgraduate students who need to prove their language proficiency for admission to higher education. The questions and tasks in the test are designed to elicit the language skills required for academic success. All four skills in the test focus explicitly on general academic English. The test experience is user-friendly and 'human', with a live proctor and speaking test examiner. LANGUAGECERT Academic is a high-stakes test, and we must maintain test security and integrity by countering identity fraud and impostors. For the online option LANGUAGECERT creates a secure, at-home test environment through the seamless combination of cameras, the human eye and software solutions.

Maintaining test security and integrity requires continuous innovation. The same is true for designing, developing and delivering quality assessments, but innovations in language testing must be built on firm foundations and hard evidence. The studies in this volume exemplify how the LANGAGECERT research-based approach relies on empirical evidence to fine-tune assessments for users' diverse needs, from general language proficiency to tests for migration and academic admission purposes.

I look forward to writing the introduction to the next volume, which will bring together our future original research, calibration, validation and empirical studies. Studies which will be instrumental to the next ten years of LANGUAGECERT designing, developing and delivering tests that are fit for purpose and meet the needs of our future stakeholders, test users and candidates.

References

The Association of Language Testers in Europe (ALTE). (n.d.) *The ALTE Q-Mark and Auditing System.* https://www.alte.org/Setting-Standards

Ecctis. (2023). *Executive Summary, LANGUAGECERT General and Academic Tests: Independent CEFR referencing*. Ecctis.

Contents

SECTION 2: ORIGINAL RESEARCH

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.

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

Corina Dourda is Assessment Development Manager at LANGUAGECERT. She coordinates the development of the LTE, IESOL Speaking, and Young Learners qualifications, focusing on designing fair, fit-for-purpose, and inclusive assessments. She holds a BA in Linguistics and Language Teaching, an MA in Curriculum Design, and a Trinity Diploma in TESOL. She has over fifteen years of experience in developing teaching and assessment materials across various educational settings and levels. Her main research interests include automated item generation and computer-adaptive testing.

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.

Catherine Jones is an Assessment Development Specialist at LANGUAGECERT. She has worked in the field of assessment for over twenty years with expertise in developing multi-level language curricula, tests and teaching materials for international organisations and governments. Catherine is particularly interested in the transformative potential of assessment and in examining the impact of high-stakes English language assessment on teaching and learning and student outcomes. She holds a BA in French and History of Art from University College London.

Leda Lampropoulou is Research Manager at LANGUAGECERT, with more than 15 years of experience in the fields of second language acquisition and language testing. In her role, she coordinates LANGUAGECERT's research programme and projects, while contributing to the maintenance of standards and the fitness-for-purpose of the LANGUAGECERT exam portfolio. She holds a BA in English language and Philosophy from the University of London and a MA in Language Testing from Lancaster university. She has published in peer reviewed journals, focusing primarily on test reliability and speaking skills assessment.

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.

Vlasis Megaritis leads the AI Engineering Team at PeopleCert, applying machine learning to a broad spectrum of real-world problems. His current work encompasses a variety of projects, including anomaly detection for fraud prevention and the development of automated scoring systems. He holds a Bachelor's Degree in Physics and an MSc in Energy Technology.

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 20 years' experience in the fields of English-medium Higher Education, Qualification Development and Language Assessment. At PeopleCert, he holds the position of Language Assessment & Quality Director. He has been responsible for the development of the LANGUAGECERT exams portfolio since 2015.

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.

Irene Stoukou is Research Associate at LANGUAGECERT. She is responsible for the analysis and monitoring of examiner performance, ensuring marking consistency and accuracy. She holds a PhD in English Literature and Culture. She has extensive EFL teaching experience in diverse educational settings. She is a member of IATEFL, UKALTA and is also affiliated with the European Association for the Study of English (ESSE), where she serves as a national correspondent for the Gender Studies Network. She is a Postdoctoral Researcher and Adjunct Lecturer in the School of English Language and Literature at Aristotle University of Thessaloniki.

Common Abbreviations

AES	Automated Essay Scoring
ARG	Accuracy and Range of Grammar
ARV	Accuracy and Range of Vocabulary
ALTE	Association of Language Testers in Europe
ANOVA	Analysis of Variance
CAT	Computerized Adaptive Test
CAF	Complexity, Accuracy, Fluency
CEFR	Common European Framework of Reference
CELST	Computerized English Listening and Speaking Test
CFA	Confirmatory Factor Analysis
CI	Confidence Interval
CP	Computer Processed
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
ENIC	European Network of Information Centers
ERA	Externally Referenced Anchoring
ESOL	English to Speakers of Other Languages
FOR	Frame of Reference
GMAT	Graduate Management Admission Test
IC	Interactional Competence
IEA	Intelligent Essay Assessor
IELTS	International English Language Testing System

IESOL	International English for Speakers of Other Languages
IF	Item Facility
IRT	Item Response Theory
HW	Hand Written
LC	LANGUAGECERT
LCA	LANGUAGECERT Academic
LCG	LANGUAGECERT General
LID	LANGUAGECERT Item Difficulty scale
LST	LANGUAGECERT SELT Test
LTE	LANGUAGECERT Test of English
L & R	Listening and Reading
MFRA	Multi-faceted Rasch Analysis
NARIC	National Recognition Information Centre
Ofqual	Office of Qualifications and Examinations Regulation
OLP	Online Proctoring
PB	Paper-Based
PEG	Project Essay Grade
PIF	Pronunciation, Intonation and Fluency
PPM	Pearson Product-Moment Correlation
PTME	Point Measure Correlation
RQ	Research Question
SD	Standard Deviation
SELT	Secure English Language Test
SEM	Standard Error of Measurement
SID	Similarity Detector
TBLT	Task Based Language Teaching
TF	Task Fulfilment
TF-IDF	Term Frequency - Inverse Dense Frequency
TLU	Target Language Use
TM	Traditional Mode
TOEFL	Test of English as a Foreign Language
TOST	Two One Sided Tests
UKVI	UK Visas and Immigration

Introduction

Leda Lampropoulou

This volume, the third in LANGUAGECERT's research series following the 2022 and 2023 publications, continues our commitment to a research-led, quality-focused approach to language assessment. The two sections of this volume comprise eleven chapters, each delving into diverse topics within assessment, including calibration, validation, and original research studies that underpin LANGUAGECERT's rigorous standards.

Each chapter is designed to be accessible as a standalone piece, which may result in some repetition, particularly in the descriptions of examination structures. Each chapter concludes with its own reference list.

Section 1: Calibration & Validation Studies

Section 1, "Validation and Calibration Studies", offers a comprehensive examination of the methods and research underlying LANGUAGECERT's approach to ensuring consistency, reliability, and alignment across its language assessments. The six chapters presented in this section showcase the ongoing efforts to align LANGUAGECERT tests with international standards, particularly the Common European Framework of Reference (CEFR), through robust calibration and validation practices. By detailing the development of the LANGUAGECERT Global Scale, the studies in this section demonstrate a commitment to maintaining test quality through external anchoring and alignment across levels, with meticulous attention to each test component—Reading, Listening, Writing, and Speaking.

Each chapter within this section investigates critical aspects of scale creation, item difficulty, and measurement consistency, underscoring LANGUAGECERT's reliance on empirical evidence to fine-tune assessments for diverse purposes, from academic and migration-oriented tests to general language proficiency. Together, these studies reinforce LANGUAGECERT's commitment to creating assessments that are both technically sound and aligned with user needs, highlighting the intersection of research-based methodologies with practical, high-stakes applications.

The opening chapter, "The LANGUAGECERT Global Scale", explores the development and evolution of LANGUAGECERT's measurement scale, aligned with the Common European Framework of Reference (CEFR). This scale, refined through data gathered since 2017, supports alignment across LANGUAGECERT exams and other scales as needed. The chapter traces the transition from the original Item Difficulty (LID) scale to the more encompassing Global Scale, detailing its calibration, alignment processes, and ongoing validation efforts to meet diverse user requirements.

Chapter 2, "The Development of LANGUAGECERT General", introduces LANGUAGECERT General, a qualification designed for the migration and employment sector as a counterpart to the academic-focused LANGUAGECERT Academic test. Building on the LANGUAGECERT IESOL B2 test, this four-skill, multi-level assessment aligns with a standardized measurement scale and incorporates pretested, calibrated content backed by validation research. The chapter discusses test construction and purpose, proficiency levels, content selection, delivery, assessment criteria, and results within a secure framework designed to meet high-stakes migration requirements.

Chapter 3, "Recalibrating and Extending the Analysis of the LANGUAGECERT Test of English", consolidates findings from multiple validation studies on the LTE, focusing on the scale development and calibration that affirm its reliability as both a linear and adaptive test. Highlighting work on the widely used adaptive LTE, the chapter presents analyses based on an extensive item bank of over 800 items, demonstrating how these insights support the robustness and precision of the LTE system.

Chapter 4, "Similarity Detection in Writing Test Scripts at LANGUAGECERT", addresses plagiarism challenges in LANGUAGECERT Writing Tests, including issues of collusion, copying, and reusing previous work. The chapter categorises common types of plagiarism and reviews statistical and computational tools for text similarity detection. It introduces SiD, LANGUAGECERT's custom similarity detection tool, developed to rigorously analyse writing test submissions for integrity. Examples illustrate SiD's operation and similarity metrics, while a growing corpus of exam scripts supports

ongoing detection. This tool is part of LANGUAGECERT's broader commitment to fairness and exam integrity.

Chapter 5, "SELT IESOL Writing Test Quality", presents findings from a study evaluating the test quality of LANGUAGECERT's SELT Writing Tests at CEFR B1 and B2 levels. Analysing data from over 11,000 candidates, 60 examiners, and 18 tasks administered between 2021 and 2022, the study uses Many-Facet Rasch Analysis (MFRA) to assess marking consistency across examiner, task, and rating scale dimensions. Results indicate strong fit to the Rasch model, with minimal examiner misfit and acceptable task and scale severity ranges, concluding that the SELT Writing Tests are robust and appropriate for their intended purpose.
Chapter 6, "Aligning LANGUAGECERT SELT Tests to the LANGUAGECERT Item Difficulty (LID) Scale," examines the alignment of SELT tests to the LID Scale, particularly for Listening and Reading components. Building on prior research using externally-referenced anchoring, this chapter confirms the robustness of SELT tests across CEFR levels B1 to C2. Findings indicate that while the tests generally align with designated CEFR levels, they include items that assess skills across adjacent levels, enhancing the breadth of proficiency measurement.

Chapter 7, "Externally-Referenced Anchoring of LANGUAGECERT SELT Tests", discusses the application of externally-referenced anchoring to vertically align SELT test forms to a calibrated midpoint. Using Rasch measurement and expert judgment, the analysis focuses on Listening and Reading tests at CEFR levels B1 to C1. Findings show strong alignment across test forms, the LANGUAGECERT Item Difficulty (LID) scale, and corresponding CEFR levels, reinforcing the accuracy of each SELT test's level positioning.

Section 2: Original Research

Section 2, "Original Research Studies", brings together five chapters that explore the ongoing innovation and empirical inquiry shaping LANGUAGECERT's approach to language assessment. Focused on addressing real-world testing challenges and ensuring fairness, these studies investigate essential aspects of test integrity, reliability,

and user experience. Each chapter contributes unique insights into specific components of the assessment process, from the handling of plagiarism and test bias to the comparability of different test delivery modes and response formats.

This section highlights LANGUAGECERT's commitment to refining its exams based on data-driven research and real-world applications, particularly in high-stakes contexts. By examining factors like bias detection, test mode comparability, and the impact of preparation methods on performance, these studies underscore LANGUAGECERT's proactive approach to maintaining fairness, inclusivity, and adaptability across its assessments. Collectively, this original research reinforces LANGUAGECERT's focus on evolving its tests to meet the diverse needs of test-takers while upholding high standards of validity and security.

Chapter 8, "Exploring Test Gender Bias in the LANGUAGECERT SELT IESOL Speaking and Listening Tests", investigates potential gender bias within the SELT IESOL Speaking and Listening tests. Covering test data from 2020 to 2023, the chapter presents an analysis using differential item functioning (DIF) to assess gender bias, finding negligible-to-no bias and affirming the tests' robustness and reliability. The chapter contextualizes these two-skills tests within the broader field of language assessment and confirms their consistency with CEFR standards.

Chapter 9, "The Use and Impact of Pre-task Planning Time in the Monologic Task of LANGUAGECERT Speaking Tests", examines how note-making during pre-task planning affects performance in the LANGUAGECERT IESOL B2 Speaking Test monologue. Focusing on whether note-making correlates with higher scores, the study compares test-takers who used this strategy with those who did not, also exploring their perceptions of planning time. Findings indicate that note-making did not enhance performance scores, though most test-takers used planning time to organize their main points.

Chapter 10, "A Comparability Study of Handwritten versus Typed Responses in High-Stakes English Language Writing Tests", explores the fairness of writing test scores between handwritten and computer-typed responses across CEFR levels B1 to C2. Analysing data from 2019 to 2022, the study uses effect size and equivalence testing, finding minimal score differences between the two formats at B1, B2, and C1, with a moderate advantage for typed responses at C2. Overall, results suggest that candidates achieve comparable scores regardless of writing mode, allowing them to choose their preferred format without risk of scoring bias.

Chapter 11, "The Delivery of Speaking Tests in Traditional or Online Proctored Mode: A Comparability Study", assesses the consistency of scores in high-stakes English language Speaking Tests administered either in traditional test centres or through online proctoring across CEFR levels B1 to C2. Using descriptive statistics, effect size, and equivalence testing, the study finds minor differences between delivery modes at the C2 level, though these differences are not statistically significant. The findings confirm that test delivery mode—whether online or face-to-face—does not meaningfully impact test-taker scores.

SECTION 1: CALIBRATION/ VALIDATION STUDIES

Chapter 1: The LANGUAGECERT Global Scale

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

Abstract

The LANGUAGECERT system is based on a measurement scale that is aligned to the Common European Framework of Reference (CEFR) and can, in turn, be aligned to other scales as required. The scale has been in development for some years (since 2017) and as data is gathered from the range of LANGUAGECERT assessments, the scale is subjected to on-going validation. As the requirements of users have become better defined, the nature of the underlying measurement scale has also developed and will progressively embrace the full range of LANGUAGECERT exams.

This chapter reports on the development of this measurement scale through a number of phases, culminating in what is now referred to as the LANGUAGECERT Global Scale. We first provide background to the original LANGUAGECERT scale – the LANGUAGECERT Item Difficulty (LID) scale. We describe its development, implementation and calibration. Discussion then moves to the nature and purpose of the Global Scale and on to the transition from the LID to the Global Scale in terms of calibration and alignment.

Keywords: measurement scale, calibration, alignment

Background to the LID Scale

The initial LANGUAGECERT Item Difficulty (LID) scale was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement. LID scale difficulty values range from CEFR Pre-A1 through to high C2 level. The scale ranges and midpoints are presented in Table 1 below.

Table 1: LID Scale

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

As mentioned, the LID scale was developed using both expert judgement and item analysis such that 20 points separated each CEFR level. In 2017, eight expert consultants, each of whom had 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 in the LANGUAGECERT item banks measured against them, thereby giving each item in these tests a difficulty value on the LID scale. An in-depth analysis was conducted on all anchor items at this stage and a small number were eliminated from further use as anchors, as they were not measuring as predicted. In the following sections, we summarise five studies that describe – against the backdrop of the LTE adaptive item bank – the validation of the LID scale.

Study 1: Initial Calibration of Paper-based Tests (2020)

One of the LANGUAGECERT item banks is devoted to the LANGUAGECERT Test of English (LTE). This test provided a very useful set of data in that it offers both linear and adaptive tests, measuring on the same scale. The bank used in these studies in 2020 contained, at the time, over 1000 items and was used to generate an adaptive test and linear tests.

To validate the expert judgements used to generate the original LID scale, a calibration exercise involving Rasch measurement was undertaken in 2020, with the focus on LTE. This version of the LTE is an English 'for work' exam intended for people over 18 in or about to enter the workplace, as well as those in higher or further education. It has been accredited by the UK's Office of Qualifications and Examinations Regulation (Ofqual). The LTE is available in three versions described in Table 2 below.

Test version	**CEFR levels aimed at**
(1) paper-based (PB) test measuring from A1-B1	beginner to intermediate CEFR levels
(2) PB test measuring from A1-C2	candidates at all CEFR levels
(3) adaptive test measuring from A1-C2	candidates at all CEFR levels

All three versions of the LTE are produced from the same LTE item bank. At the time of analysis (2020), the LTE item bank that was to be analysed contained around 1,600 items. Currently, it contains over 3,500 items and continues to grow. From this item bank, both paper-based and adaptive tests were produced, utilising in total approximately 1,600 items (827 in the adaptive test and more than1,000 in the PB tests) with many common items between the CAT and the PB tests, and between different versions of the PB tests for cross-calibration purposes.

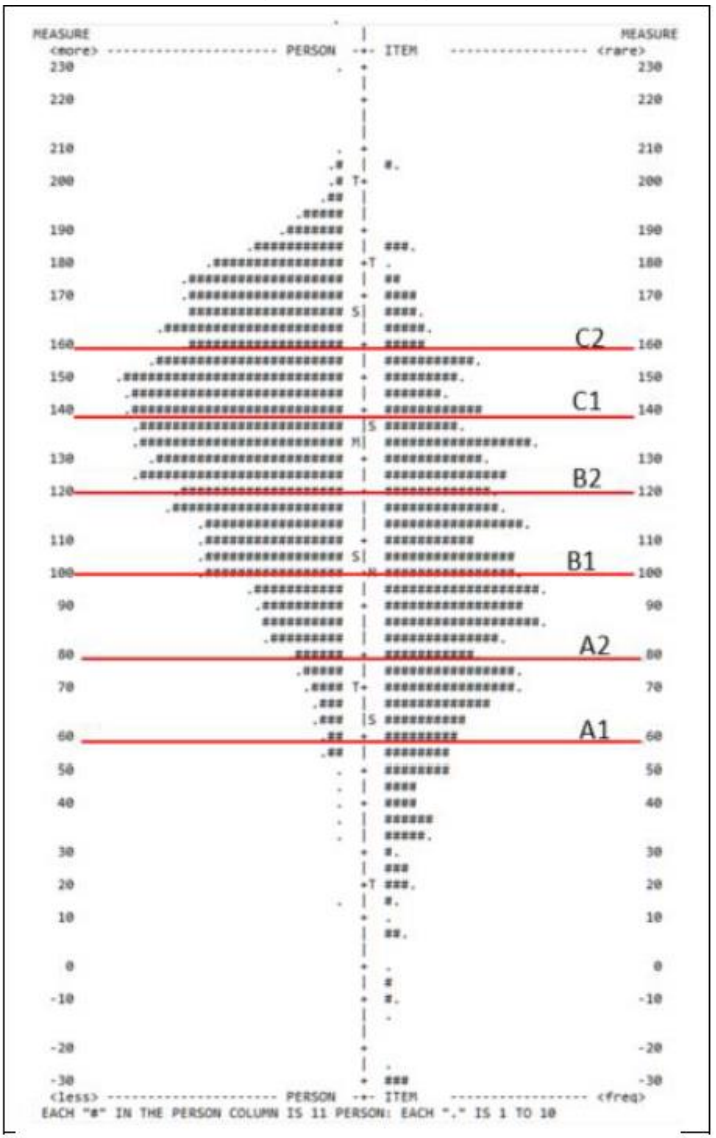
The first study explored four paper-based (PB) tests with a view to establishing an initial set of anchor items, by which the entire item bank might be subsequently calibrated. The initial sample comprised a total of 282 discrete items in the four-test database which had been administered to 2,112 candidates. The fit of the items to the Rasch model was good and reliability was high. With all four tests calibrated to a single scale, the calibrated scale was rescaled to a mid-point of 100 with a spacing factor of 20 in order to align the calibrated Rasch scale and the original LID scale. The rescaling of the Rasch scale in this manner produced a comparable alignment between the two scales although some differences were detected at the A level which required further exploration.

Study 2: Calibration of Adaptive Test Item Bank (2021)

The initial calibrated scale that emerged from the set of paper-based tests demonstrated that the paper-based tests were robust and consistent with the data. This provided a basis for the further validation of the LID scale through data generated by the adaptive test and was the second major calibration study.

The calibration conducted in 2021, based on the LTE item bank incorporated the 827 items in the LTE adaptive test which had been administered to 5,800 candidates (with each candidate having taken approximately 60 items). The dataset incorporated the 282 calibrated items from the paper-based tests in Study 1. These items formed anchors in the Rasch measurement calibration. The Rasch person-item map in Figure 1 below shows the fit of candidates (to the left-hand side of the map) and items (to the right).

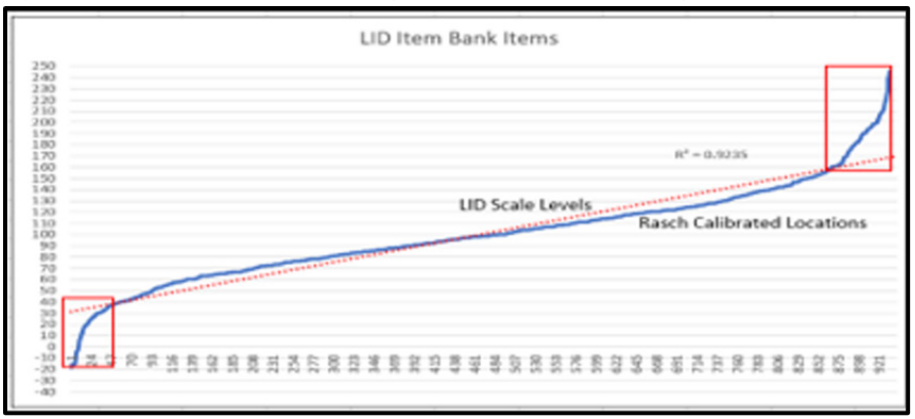
Figure 1: Person-Item map from LTE adaptive item bank calibration



As Figure 1 illustrates, person and item distributions extended approximately 120 points, or six logits – the rule-of-thumb operational range (Bond et al., 2020). Candidates generally matched with items. The person distribution is dependent upon the nature of the test population, and a considerable number of high ability candidates were known to be in the sample.

Concerning the items, there was nonetheless generally a good match between Rasch calibrated locations and LID scale expert-defined levels. Figure 2 illustrates.

Figure 2: Calibrated locations and expert-defined level fit in adaptive bank test items



While the R^2 figure was good at 0.9, there was a degree of misfit at the bottom (the 'lag' phase of the Sigmoid or 'S' curve [Handy, 1995]) and top (the 'steady state') ends of the scale, which should ideally be flat with small gradients to indicate slower rates of item difficulty increase. As can been in Figure 2, there are sudden rises and falls in item difficulty levels at the extreme ends of the scale, with items being either too easy (at the bottom end) or too difficult (at the top end). It was felt that the sharp downturn at the bottom end of the distribution was due for the most part to the fact that there were very few A1 and pre-A1 candidates in the dataset. The sharp upturn at the top of the distribution may have related to the inclusion of a small number of very difficult items.

However, the conclusion drawn from Study 2 was that the LID scale could be considered to be a comprehensive and robust scale.

Studies 3 and 4: Simulations (2022)

The next stage in the validation process was to consider the stability of the bank. With a coherent LID LTE scale developed, and when the adaptive test cohort surpassed 10,000 candidates, two linked studies exploring the stability of the 827 items in the adaptive test were conducted. Study 3 explored item bank stability through a simulated 'full' dataset generated through model-based imputation (i.e., whereby the parameter values of persons, items and thresholds from the current analysis were used to generate simulated data according to the probabilistic distributions defined by the Rasch model and generating Rasch parameters). Results pointed to item bank stability, indicating that items making up the adaptive item bank were of high quality both in terms of content and statistical stability. Potential future stability was confirmed by results obtained from a Bayesian ANOVA.

A linked follow-up study (Study 4) involved submitting the items to a 'real-world' test by which three (paper-based) tests were compiled from the calibrated items in the adaptive test and was administered to a sample of test takers. In the analysis of the three tests, good fit statistics emerged, with high correlations between the tests – an indicator of robust joint calibration and further evidence as to the stability of the item bank.

Study 5: Finalising the Calibration (2022)

As of mid-2022, the LTE adaptive test used in these studies comprised 827 items and had been administered to over 48,000 candidates. A recalibration was then performed. Figure 3 below summarises the recalibration results.

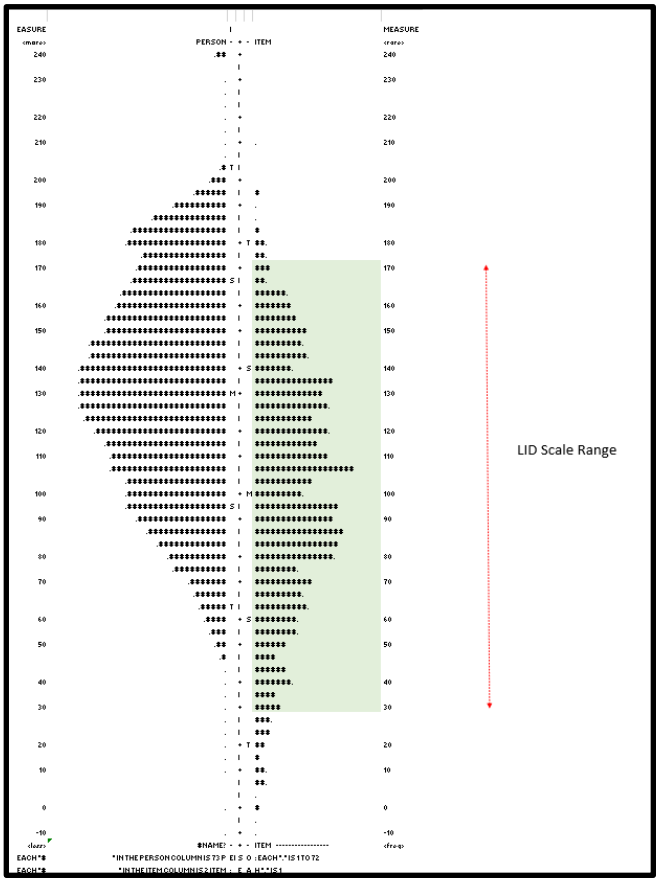
Figure 3: Summary of Rasch analysis

PERSON	48722 INPUT	48054 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	34.9	56.7	131.80	6.68	1.00	.0	1.01	.0
P.SD	5.8	2.8	35.14	1.09	.12	.9	.42	.9
REAL RMSE	6.77	TRUE SD	34.48	SEPARATION	5.10	PERSON RELIABILITY		.96
ITEM	923 INPUT	827 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	2027.8	3295.1	110.12	3.17	1.00	-.6	1.01	-.6
P.SD	1614.8	2122.6	59.04	7.96	.09	4.9	.20	5.1
REAL RMSE	8.57	TRUE SD	58.41	SEPARATION	6.81	ITEM RELIABILITY		.98

Measurement error (RMSE) was 8.57 (less than half a scale level against the 20-point LID scale); the separation index (an index pointing to construct validity) of 6.81 was well above the customary decision level of 2.0 for good separation, indicating clearly distinguishable item locations with little chance of overlaps due to measurement error. Reliability was very high at 0.98.

To finalise the calibration, additional external calibration was conducted on A1 and C2 level items, with the results subsequently incorporated into the overall LID scale. The item-person map incorporating all levels is presented in Figure 4 below.

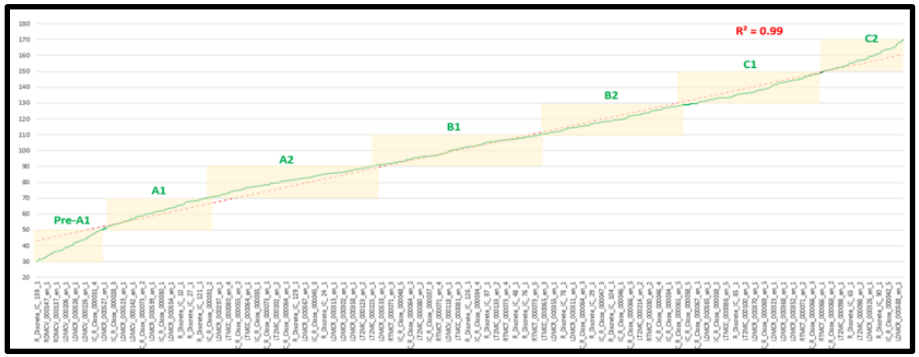
Figure 4: Person-Item map



As can be seen, LID scale item difficulties range from 30 (Pre-A1) to 170 (C2).

The finalised LID scale after calibration is presented in Figure 5.

Figure 5: Finalised LID scale calibration



As can be seen, there is a generally linear gradation from pre-A1 up to C2.

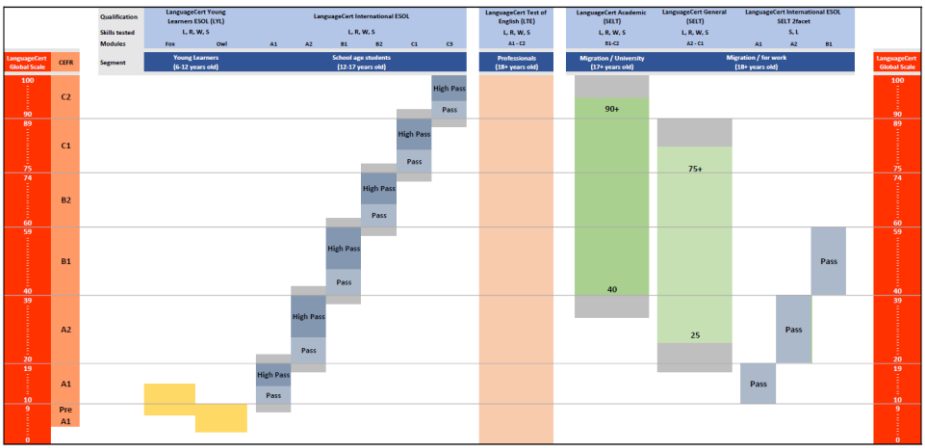
The results above provided confidence that the LID is measuring as has been claimed, with the 800+ LTE items in the adaptive test version forming the bedrock, the base, against which future items and tests may be calibrated.

The underpinning of the LID scale now allows for the transition to the Global Scale to be outlined, to which the discussion now turns.

Transitioning from the LID to the Global Scale

The LID scale was intended as an internal item difficulty scale in the item bank system. Feedback from external users of the LID scale suggested that the effective scale range of 50-170 was not sufficiently intuitive. Therefore, in order to make the scale easier to work with, following consultation, it was decided to recalibrate the scale to 0-100 and use this as a basis for mapping all LANGUAGECERT tests. It was also felt that 'LID' was not very transparent as a name. As a consequence the scale was renamed the Global Scale. It links directly to the LID scale and thereby the CEFR levels. Performance on LANGUAGECERT tests can then be mapped to other English language testing organisations' examinations such as IELTS and Cambridge Advanced. Figure 6 illustrates an initial representation of the Global Scale and how it reports against CEFR levels.

Figure 6: The LANGUAGECERT Global Scale



The figure above illustrates how the LANGUAGECERT System reports scores on the LANGUAGECERT Global Scale of 0-100 and applies across all the tests in the LANGUAGECERT System. The Global Scale provides candidates, employers, education institutions and government agencies an easy-to-understand results 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.

Mapping the LID Scale to the Global Scale

Figure 6 presented a visualisation of the Global Scale in relation to the full range of assessments offered by the LANGUAGECERT system. Before full implementation, however, it must be demonstrated that the LID scale, to which all LANGUAGECERT tests have been aligned, maps cleanly and clearly to the Global Scale. The discussion below outlines how this issue has been addressed.

Two methods were considered regarding mapping the two scales to each other. The first method was to simply divide the active 120-point LID scale into 100 with 1.2 LID points per Global Scale level point. While this method might appear intuitive, the realigning of a 120-point to a 100-point level would be mathematically fraught in terms of actual administration. More importantly, such a realignment would result in an ordinal scale which progresses in integer steps, omitting in-between step differences, and hence possibly obscuring between-level differences, potentially misrepresenting scores. An interval scale, in contrast, is continuous and permits in-between step differences.

Having therefore discounted the simplistic calculation of 120 to 100 by 1.2 scale points, the methodology adopted was to calibrate the LTE item bank such that a scale mid-point and a logit value yielded a scale with 100 as total. The aim is therefore to shift the scale to a lower mid-point and with a narrower logit range in order to transition to the Global Scale.

After several iterations, the scale mid-point of 50.5 and a logit of 15 were found to yield a good approximation, with a Pearson correlation of 1.0 between the LID and Global scales. The Global Scale (GS) that emerged in presented in Table 3 below.

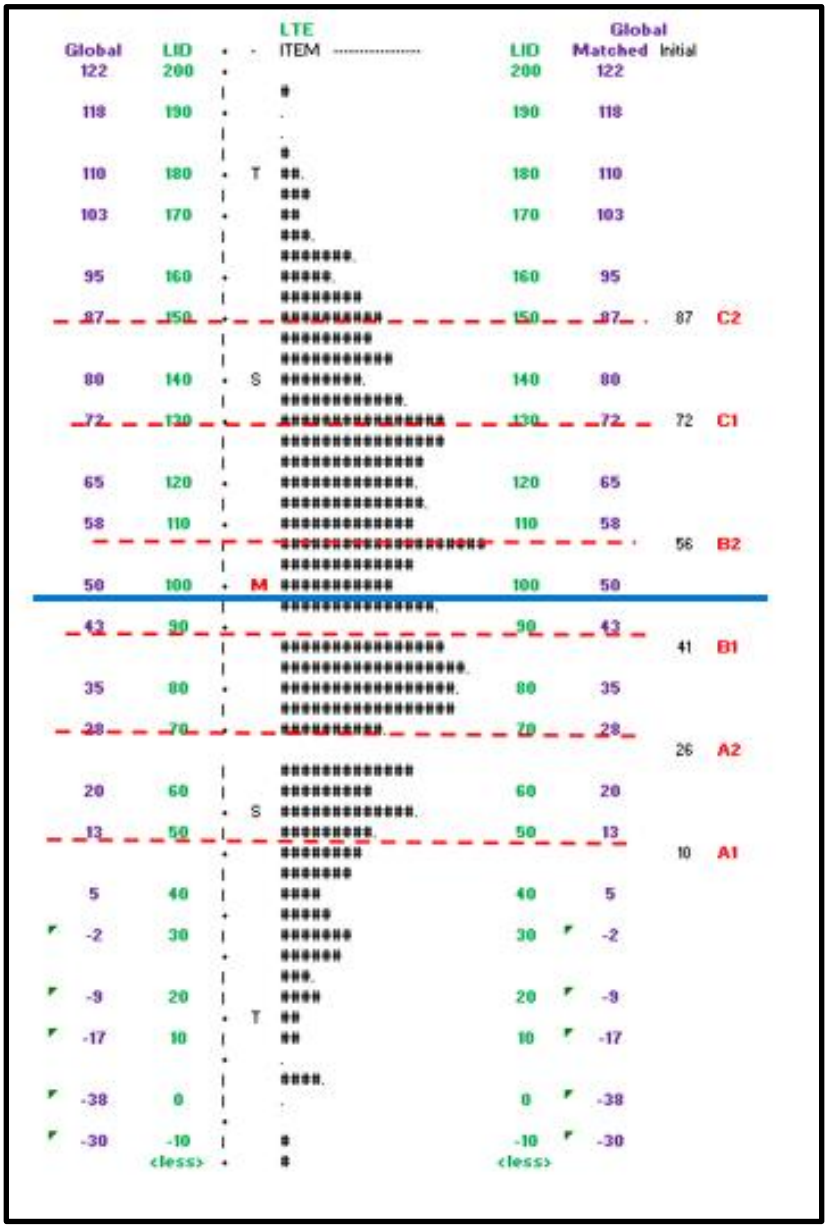
Table 3: LID and Global Scale fits

CEFR	LID Scale upper and lower ranges	Global Scale upper and lower ranges	GS cut score point	GS point range
C2	170.08	100		
C2	150.25	87.03	87	
C1	149.82	86.33		
C1	130.67	72.13	72	15
B2	129.93	71.40		
B2	110.06	56.36	56	16
B1	109.75	55.57		
B1	90.05	41.63	41	15
A2	89.79	40.83		
A2	70.01	26.05	26	15
A1	69.90	25.68		
A1	50.02	10.62	10	16

As can be seen, the Global Scale level widths are not uniformly equal. A1 begins at 10; there are 15 or 16 points up until C1, and then C2 comes in above 87.

With the new midpoint and spacing factor, the items in the LTE item bank needed to be recalibrated. The visual mapping of the two scales on the item map is presented in Figure 7 below.

Figure 7: LID and Global Scale item map



Future LANGUAGECERT Assessment Products

For the recently introduced LANGUAGECERT Academic and General tests (see Jones, 2023), results are reported against the CEFR levels and on the LANGUAGECERT Global Scale. 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.

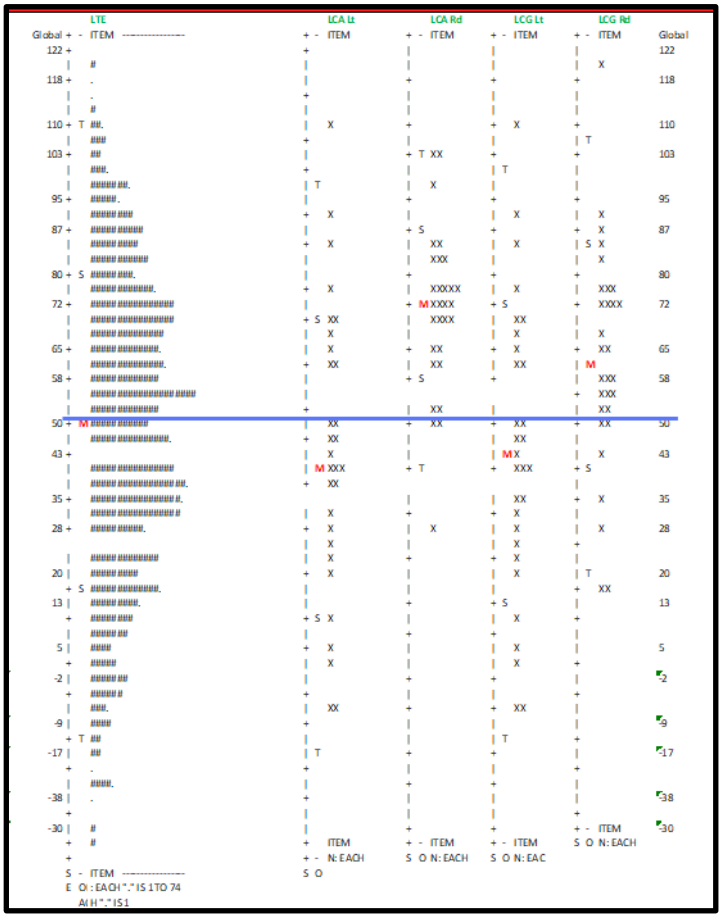
[1]

LANGUAGECERT Academic and General Tests

The Listening and Reading tests in the LCA and LCG series of tests were calibrated with the calibrated Rasch results of the LID item bank as the calibration standard. This was achieved by linking the LCA and LCG tests via common items in the LTE item bank. In order not to dislocate established calibrated Rasch values in the LID scale, all items in the LID scale were anchored before calibration with the LCA and LCG tests appended to the dataset. Since the initial LID scale values have now been matched with the 100-point Global Scale, the combined LID LCA / LCG data were calibrated using the 100-point Global Scale as the reference scale. It should be noted that the process described will be the general process adopted as future LANGUAGECERT tests are matched to the Global Scale.

Figure 8 below presents the calibrated LCA L&R, LCG L&R tests against the (re-calibrated) LTE item bank.

Figure 8: Global Scale showing re-calibrated LTE item bank and LCA L&R, LCG L&R tests



Estimating Reliability

A widely adopted approach to derive overall scores for language tests comprising two or more of the four skills involves summing total or average component scores. Such an approach assumes that the component tests have equal weighting, an assumption that needs to be verified if the resulting summary score is to reflect the relative importance of the component tests. To estimate the relative prominence of the LCA and LCG Listening and Reading tests, McDonald's Omega reliability was used to estimate, via confirmatory factor analysis (CFA) loadings, the relative weighting of the two component tests (see Hayes & Coutts, 2020). Table 4 below reports the results.

Table 4: CFA Standardised loadings of LCA and LCG tests

Test	**Standardized loading**
LCA_L	0.943
LCA_R	0.942
Test	**Standardized loading**
LCG_L	0.957
LCG_R	0.957

It can be seen that the two LCA and the two LCG tests have near equal loadings, indicating equal prominence. In such a case, summing up or averaging listening and reading in LCA and LCG to derive overall scores is justified. It is recommended that averaging be used to keep overall scores within the 100-point Global Scale. Using averages involves computing the mean of the component tests – two in the case of LCA Listening and Reading in Table 4 above. If the component tests do not have near equal weights, the mean would advantage the test/s with lower weight and disadvantage those with high weights, resulting in inaccuracies in the overall scores and leading to, in extreme cases, possible candidate appeals.

External Triangulation: Comparison with IELTS

In order to establish the extent to which results on the LCA and LCG related to those on another internationally used exam, around 500 candidates took one of the two LANGUAGECERT tests (LCA or LCG) and the IELTS equivalent.

The statistical procedure appropriate in such situations is the multinominal test. This test estimates the equivalence of CEFR levels obtained by candidates in the LCA/LCG tests and the CEFR levels obtained on the IELTS test.

The Bayesian version of multinomial tests further estimates the range of variation in the two sets of rankings in the population, known as the 95% credible interval (CI). If and when the two sets of rankings fall within the CI, they are deemed to be equivalent. Understandably, the multinomial test is sample size sensitive. Given the relatively small sample size, six sets of model-based simulated data were generated for the LCA and LCG data, extending the initial total sample of 500 for both tests to a large sample of over 3,000. Bayesian multinomial test credible intervals (CI) were calculated for the comparative distributions to provide an indication of future distributions. Table 5 presents the results.

Table 5: LCA/LCG and IELTS comparative distributions

			95% Credible Interval (CI)	
CEFR	LCA/LCG	IELTS	Lower	Upper
A2	1%	1%	0%	1%
B1	8%	8%	7%	9%
B2	32%	32%	30%	33%
C1	30%	29%	28%	31%
C2	30%	30%	29%	32%

Both projected LCA/LCG as well as IELTS sample totals were within lower-upper CI ranges. To exemplify, it was projected that 30% of the LCA/LCG and 29% of the IELTS sample would obtain C1. The 95% credible interval for the percent obtaining C1 was projected to be between a lower bound of 28% and an upper bound of 31%, which was indeed the case. These are very compelling findings.

Global Scale / Raw Score Conversions

With the Global Scale is place, it is now possible to produce a range of indicators or metrics which relate directly to the GS. One of these, which will need to be test-specific, is the concept of the raw score conversion table. For a given test, the raw score conversion table maps the raw score to the Rasch-calibrated 100-point scale. Table 6 presents a sample of the LCA test mapped on to the Global Scale.

Table 6: LCA test mapped on to the Global Scale

Score	Listening	Reading	
30	100	95.81	
29	100	95.47	
28	98.38	92.16	
27	92.41	91.29	
26	89.71	90.7	C2
25	84.53	86.13	
24	81.24	82.47	
23	80.92	79.01	
22	80.4	77.28	
21	78.07	77.04	
20	75.53	76.75	
19	74.84	76.73	
18	73.57	76	
17	73.12	75.78	
16	72.54	74.32	C1
15	69.76	73.32	
14	68.36	72.96	
13	67.65	72.42	
12	65.03	71.88	
11	64.78	70.6	
10	62.29	69.68	
9	59.33	69.65	
8	57.48	64.14	
7	57.46	61.13	
6	56.33	60.08	B2
5	55.48	58.32	
4	51.36	55.46	
3	50.6	55.01	
2	46.54	52.44	B1
1	40.95	41.49	A2

While score correspondences between skills may be expected to be close, they will not necessarily be exactly the same. This may be seen in Table 6 above. At C2, both Listening and Reading have a lower bound cut score at the same point. At C1 and below, however, Reading is somewhat offset to Listening with slightly lower cut score points. However, while some variance between test forms is inevitable, it important to reduce such variance to a minimum, and it is in this context that the methodology described here is important.

Global Scale Score Report

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 'C1 ability', for example, but a more precise level of detail is provided on test taker's performance. The Score 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. Appendix 1 presents a sample of a certificate for LANGUAGECERT Academic reporting Global Scale scores.

In Closing

This chapter traced the development of the LANGUAGECERT Global Scale from the original LID scale. The process began with the establishment of a set of Rasch-calibrated item locations for the LANGUAGECERT Test of English (LTE) test items. The LID scale was then calibrated, and the precision and stability of the scale established on the basis of overall reliability and construct validity. The LID scale was found to be sufficiently robust and after calibration aligned well to the Global Scale with an appropriate mean and logit value. The Global Scale was then used to calibrate and map the LCA and LCG Listening and Reading tests. An alignment with IELTS-based CEFR levels of candidates in the LCA and LCG tests with CEFR levels specified within the Global Scale resulted in a remarkably close match.

The LANGUAGECERT Global Scale may, it can be seen, be taken as appropriately established with a strong developmental background and rigorous validation procedures. External cross validation established via correspondences with IELTS underscores the robustness of the LANGUAGECERT Global Scale, illustrating its clear links to the CEFR.

To conclude, the detail outlined about the development of the LANGUAGECERT Global Scale illustrates how the Scale forms a solid foundation for all LANGUAGECERT tests to expand into a language test pool to assess most language ability areas with good assessment quality and a stable standard mapped to the CEFR.

References

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Milton Park, UK: Routledge.

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

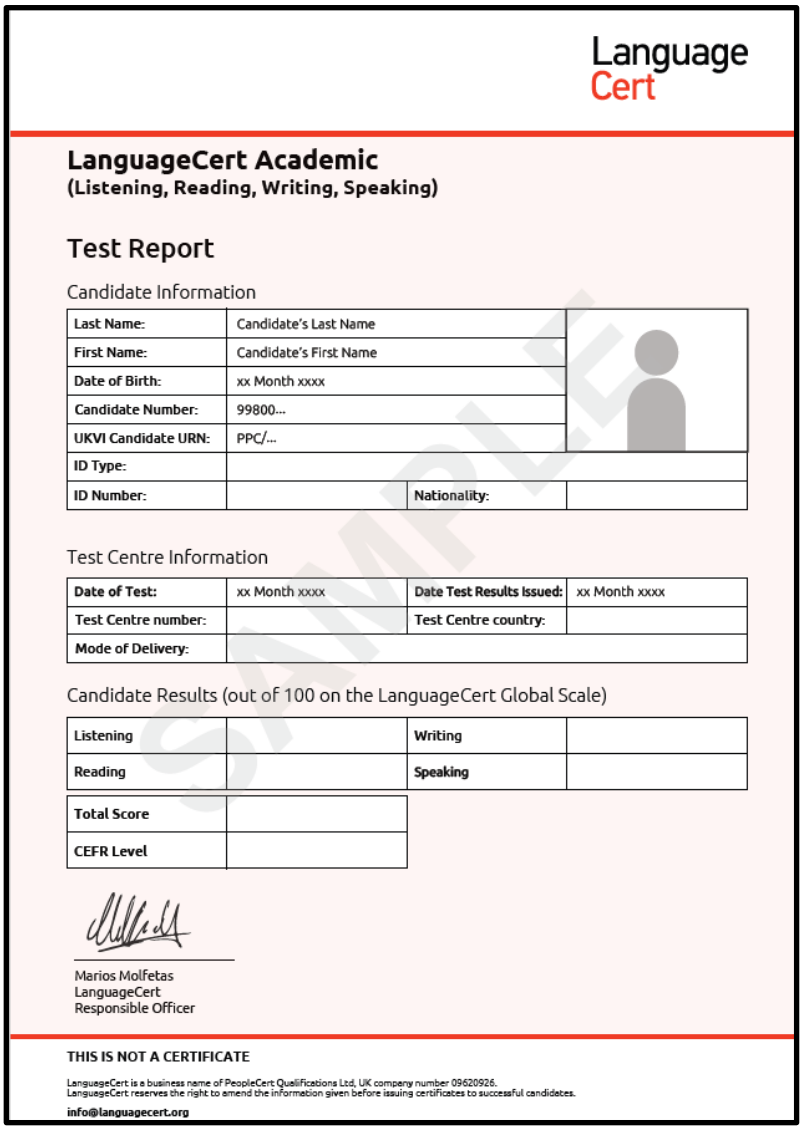
Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021b). *Validating the LanguageCert Test of English scale: The adaptive test*. London, UK: LanguageCert.

Handy, C. (1995). The Sigmoid Curve. In *The empty raincoat: Making sense of the future* (pp. 50–57). London, UK: Arrow Books.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, 14(1), 1-24.

Jones, C. (2023). An exercise in evolution: Refocusing LanguageCert IESOL C1 to the academic context. In P. Falvey & D. Coniam (Eds.), *Certifying quality in assessment and learning: Research and validation at LanguageCert* (Vol. 2). London, UK: LanguageCert.

Appendix 1: Certificate Reporting LANGUAGECERT Academic Global Scale Scores



Language Cert

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...		
UKVI 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

SAMPLE

Chapter 2: The LANGUAGECERT General Test: Assessing Language In The Migration And Employment Domains

Cathy Jones

Abstract

Following Jones' (2023) chapter outlining the positioning of the *LANGUAGECERT Academic* IESOL C1 examination to an academic context, the current chapter describes the development of *LANGUAGECERT General*, a counterpart qualification to *LANGUAGECERT Academic*, which addresses the migration employment domain. *LANGUAGECERT General*, which is closely based on the pre-existing LANGUAGECERT IESOL B2 test, is a four-skill, multi-level test, aligned to a common underlying measurement scale, derived from a bank of pretested and calibrated assessment material and associated validation research based on an established candidature. The current chapter highlights underpinning research, evidence and best practice which have informed the development and definition of a high-stakes relevant, reliable and secure test for migration purposes. It covers test purpose and construct, proficiency levels, task selection, test content, assessment criteria, test delivery, results and an integrated learning ecosystem.

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

Background

As a leading provider of language examinations and qualifications recognised by universities, employers and governments around the world, LANGUAGECERT designs its examinations such that they 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 suite of examinations, ranging from levels A1 to C2 in the Common European Framework of Reference for Languages (CEFR) for both occupational and personal use. The portfolio also includes the LANGUAGECERT Test of English, a multi-level linear and adaptive test of English in the workplace, as well as a suite of secure level-specific IESOL SELT (Secure English Language Test) qualifications, using ESOL examination 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 (see https://www.gov.uk/guidance/prove-your-english-language-abilities-with-a-secure-english-language-test-selt).

As outlined in Jones (2023), in 2020, *LANGUAGECERT General* (LCG) and its counterpart qualification, *LANGUAGECERT Academic* (LCA) were conceived as a dynamic response to changing markets and stakeholder expectations. As a result, work began to extend the portfolio with two high-stakes tests: one for the academic sector and one for those wanting to migrate for work or training in an English-speaking environment. Both tests are derived from the LANGUAGECERT item bank and report scores across relevant levels on the same measurement scale that is used for all LANGUAGECERT – the Global Scale. Global Scale scores are reported for the four skills, Listening, Reading, Writing and Speaking; and the overall result. The focus of LANGUAGECERT General is general English language proficiency for adults. It is designed to measure various aspects of language proficiency to support language policymaking and decision-making by governmental institutions, authorities and employers. One of the main outcomes of the evolution of the existing IESOL B2 and C1 tests into the LANGUAGECERT General and LANGUAGECERT Academic is to enable domain-specific measurement and certification across a broader range of relevant language attainment levels. This meets growing demand from different stakeholders (candidates, recognising institutions, and educational and business authorities) for more breadth in the areas that single level examinations assess. The multilevel format offers practical advantages, particularly in the context of migrants, accommodating an inclusive range of candidates with varied language backgrounds and experiences.

A phased roll out of LCG and LCA began in 2022 to ensure that all issues related to the effective delivery of the examinations could be addressed. A gradual roll-out (Phase 1) was planned to ensure not only a smooth introduction of the revised examinations but also to avoid confusion with existing IESOL SELT examinations used for UK visas and immigration (UKVI). LCA and LCG have been designed to replace the four single-level tests, already in use by UKVI, before the end of 2023. Phase 2 of the rollout took place in early 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 methodology for refocusing LANGUAGECERT IESOL B2 responsively as part of the LANGUAGECERT continuous test development and review cycle. It also provides evidence for test users of how ongoing research informs best practice and how it can be applied to test development.

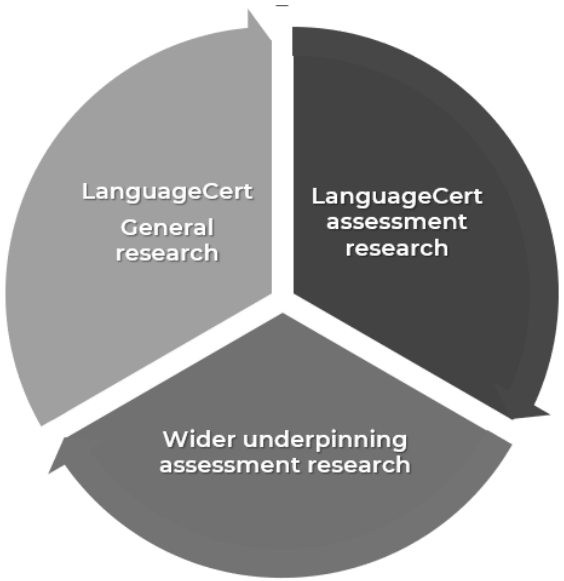
An Evidence-informed Approach

The LANGUAGECERT General test development references a portfolio of research and validation covering three main areas:

1. Wider underpinning research into assessment, learning and teaching – evidence which is referred to below.
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 LCG

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 LCG feeds back in turn to the wider assessment landscape and informs future LANGUAGECERT products as well as the 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



Summary of Underpinning Evidence

The LCG test is designed to measure the English language skills and abilities of individuals who migrate to an English-speaking country. It evaluates language skills for various purposes, including immigration, employment, education and social integration. The LANGUAGECERT B2 test has been widely used since 2017 and was fine-tuned in 2019 based on requirements set by the UK Home Office's Visas and Immigration authority.

In terms of underpinning evidence, for this development LANGUAGECERT drew mainly on the levelled specifications of language needs that complement the Common European Framework of Reference for Languages (Council of Europe 2001, 2018) and the CEFR itself. LANGUAGECERT's development team were also cognisant of publications such as Brindley and Burrows' (2000) *Studies in Immigrant English Language Assessment*.

Publications related to the CEFR used in the development of the LANGUAGECERT General test include:

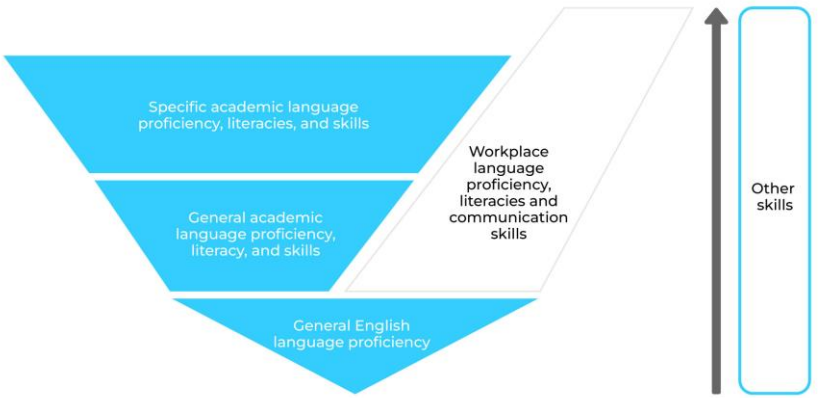
- The Common European Framework of Reference for Languages, Council of Europe (2001, 2018)
- A2: van Ek, J.A. and Trim, J.L.M. (1990b/1998b) Waystage 1990. Cambridge University Press.
- B1: van Ek, J.A. and Trim, J.L.M. (1990a/1998a) Threshold 1990. Cambridge University Press.
- B2 and above: van Ek, J.A. and Trim, J.L.M. (2001) Vantage. Cambridge University Press.

Consideration was also given to the performance descriptors as well as findings and recommendations of the ALTE reports *Linguistic integration of adult migrants: requirements and learning opportunities* (2018). as well as *Language tests for access, integration and citizenship: an outline for policy makers* (2016).

What is General English and Why is it Important?

Knoch (2021) proposes a useful model of what language proficiency entails in a post-secondary context. At its foundation, the model includes general English language proficiency (e.g. Bachman, 1990; Bachman and Palmer, 2010). Alongside these general and specific academic English proficiencies sits a component that Knoch (2021) labels workplace proficiency, literacies and communication skills.

Figure 2: Language proficiency in a post-secondary context (after Knoch, 2021)



The LCG is a test of the underpinning building blocks of general English language proficiency which include some elements of general workplace language, proficiency, literacies and communication skills as well as some elements of language for further study and training.

Defining the Target Language Use Domain

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

International migration takes place in many different contexts for many different reasons, from entry for work and study purposes, family reunion and entry as an asylum seeker or refugee. The linguistic requirements for migrants may vary, depending on the migrant "journey" – a journey whose stages are described by Saville (2009) in the context of assessment in the management of international migration.

In the context of migration, language skills are vitally important in securing some fundamental human rights: fostering social inclusion, access to education, employment, healthcare and housing. Defining the target language use domain involves establishing the real-life linguistic demands on migrants and deciding if and how these be measured in a valid test design.

Definition of the LANGUAGECERT General target language test domain must contain more detail than a test solely designed "for migration" and this is intrinsically linked to the test purpose. The LANGUAGECERT General test is for candidates seeking to migrate for work or vocational purposes. It can also be used to measure the language competence required for effective social interaction. As a high-stakes test, it can also be used to make decisions regarding immigration, right to remain and the acquisition of citizenship.

The test measures a range of skills and competences appropriate for personal, occupational and vocational contexts: reading and listening for gist and detailed understanding of a range of written and audio sources including adverts, articles, websites, diaries, radio programmes, and podcasts. The test includes writing in formal and informal registers, expressing viewpoints and taking part in role plays in real-life scenarios. In total, LANGUAGECERT General is tailored to those wanting to live, work, study or train in an English-speaking context.

The focus on domains, and the target language use within them, permeates all aspects of test design, development, and delivery. This includes how LANGUAGECERT ensures that candidates are supported with examination-specific practice tests and learning materials. LANGUAGECERT does this "by design", with all aspects of each qualification being fully integrated and aligned.

The LANGUAGECERT System of examinations has been developed using a range of language models addressing different language sub-skills and competencies. This includes the models from authors such as Bachman (1990), Canale and Swain (1980), and Weir (2005), as well as the model proposed by the CEFR (2018, 2019) which is the recognised international standard. These sources are used to ensure that LANGUAGECERT's tests are valid, reliable, and authentic for the targeted domains.

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 (2021) 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 candidates 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 candidates and their tutors not to prepare for the tests blind to the language skills necessary to succeed so that they will not be unclear as to 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 examinations. It also ensures that the skills learners practice for the tests have real-world validity and maximise learners’ quality of life in terms of personal, occupational, social and economic wellbeing.

English-language learning for migration to an English-speaking context leads to a wealth of individual benefits and societal advantages for test takers. These include increased access to opportunities and resources in terms of education, employment and personal growth and social services including education, healthcare and housing; improved ability to understand laws, customs and practices that influence interaction, communication and behaviour in the host context, smoothing integration into daily life and community activities; enhanced communication with a range of acquaintances, employers, colleagues and service providers. More effective communication skills lead to increased social interaction and foster a sense of belonging as well as mutual cultural exchange, understanding and appreciation.

An overarching intention is to contribute to understanding how assessment might be used to improve long-term outcomes. An underpinning principle to LANGUAGECERT's approach to test development is that if the test is not fit for purpose, it is understandable that teaching (or learning) to the test can constitute negative washback and a focus on skills or knowledge – nothing more than hurdles to clear in an examination scenario – which will not enable personal, occupational, social or economic success and wellbeing. If the test is designed consultatively to meet the specific needs of stakeholders – including migrants, employers, authorities and policy makers – then LCG 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.

Designing Tests that Measure Language Competence Across the Four Skills

This section outlines how language competence is measured across the four skills test.

Developing Domain Relevance in the Listening Tests

The LCG Listening test consists of 30 items across four parts. The range of content types are appropriate for the targeted domain in terms of relevant task types underpinned by robust statistical measurement that allow candidates to focus on content rather than familiarity with too many different activity requirements.

In one part of the Listening Test, candidates hear a range of dialogues in a range of situations and contexts in which migrants might find themselves. An awareness of the appropriacy of language depending upon who the interaction is with - a formal interview, a boss at work, a co-worker, or a neighbour - enables successful communication, helps achieve desired outcomes and derive value from social networks, relationships and interactions within a community or society.

In another listening task, candidates take notes while listening to a monologue. The ability to listen to an extended monologue and take notes is an essential skill for a migrant in many scenarios in everyday life. Listening and taking accurate notes is important in social, educational and occupational scenarios. The ability to note factual details when presented with information is important, for example, when dealing with the administrative requirements associated with migration, e.g., noting required documentation or addresses of offices. In a personal context, migrants might wish to take notes to record medical arrangements or details provided by friends and acquaintances in a more social setting. Additionally, in an educational setting, the ability to take accurate notes is a prerequisite for supporting educational development not only in English, but across a complete range of professional, vocational, and academic study.

Candidates also have the opportunity to hear an extended conversation. Understanding and following an extended dialogue is an essential communication skill in a range of settings for a migrant, including occupational, social, and educational. There are multiple challenges when listening to extended conversations and discussions between colleagues, friends, acquaintances, or when listening to the news, or enjoying forms of entertainment for relaxation. Speed of delivery, lack of visual cues, extended range of vocabulary and topic knowledge are all in play. The skills a migrant needs to overcome potential barriers to communication include listening for key vocabulary and linguistic signposts to ascertain gist, what the speaker is talking about and why. Candidates also need to be able to focus on the details of extended discussion to identify opinion, purpose, agreement, disagreement, feelings, and emotions in range of social, cultural and economic contexts.

Range of Accents

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

The Listening Test includes a range of accents drawn from the UK national and regional and other English-speaking communities, including North American, Australian, 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 candidates study or migrate to. It also recognises where institutions reside. As the market is dynamic, this balance is continually reviewed and integrated within 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. As a global examination board, working with international teams of test developers and writers, LANGUAGECERT avoids a UK-centric bias (in terms of accent, topics, vocabulary, cultural context and socio-economic or educational bias), which could lead to advantages or disadvantages for certain groups of candidates.

Developing Domain Relevance in the Reading Tests

Reading skills enable migrants to access a range of information including services, employment opportunities and learning resources. For migrants with families, reading skills are important both for their children's education and to help with homework. The Reading Test consists 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 test. These include newspapers, websites and public notices. Two of the IESOL SELT content types are unchanged and two new content types have been included to target level and domain more effectively.

LCG includes a new (in relation to IESOL B2) Reading Part 1, divided into Part 1a and Part 1b, both of which are vocabulary tasks. Part 1a is a multiple-choice task in which candidates read six sentences and replace a highlighted word in each sentence with a synonym without changing the meaning. There are four options to replace each word. Part 1b is a multiple-choice cloze task in which candidates 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. In everyday life, migrants are likely to encounter unfamiliar vocabulary. The ability to deploy reading strategies to work out the meaning of unfamiliar words, for example by using the surrounding language, is essential to support understanding in reading. Migrants will need to understand unfamiliar words so they can interpret the overall meaning of a sentence, and thereby understand wider meaning of whole texts including books, magazines, work-related documentation, legal documentation, forms, newspapers, letters, and emails. Without understanding how to use these strategies, reading is disjointed, frustrating and unpleasurable (impacting negatively on confidence and reading for pleasure). In addition, communication is impaired.

The Reading Test includes a range of different genres. Understanding how texts are structured is an important skill for a migrant who needs to follow longer texts for a range of occupational, educational, and personal purposes, e.g., reports, instructions, articles, and training documents. Different types of writing are structured according to specific conventions using specific cohesive devices and the ability to identify and use these markers supports fluent and active reading that will also support the development of writing skills.

The Reading Test also includes opportunities to process information from a range of sources. Migrants will need to assimilate information from a range of texts on a related theme or for an overall purpose; e.g., a range of product reviews on a website, comments on a topical matter or workplace issue, or other written material to support understanding of an issue, instruction, or question. It is important and helpful for migrants to be able to identify meaning, opinions, facts, and attitudes in a range of texts and to be able to compare and contrast these in their reading.

Finally, in occupational, educational, and personal domains, migrants are required to read longer texts. The ability to read and understand longer texts is a foundation skill which empowers people to grow and succeed as employees, students, individuals in society and as prospective citizens. Understanding the key features and content of a range of longer texts will enable migrants to develop and consolidate new skills, learn, and grow their imagination, as well as improve their other English language skills.

Developing Domain Relevance in the Writing Tests

Writing skills allow migrants to express themselves, self-advocate and access opportunities. LCG contains two writing tasks set in contexts that are appropriate for the nature of the candidature and the desired outcomes of the test. Tasks revolve around neutral/formal and informal communicative writing for a specific purpose and intended audience. In the first task, candidates produce a short letter email or report in response to a short input text covering three required pieces of information. In the second task candidates compose an informal email, a narrative or descriptive text, or an article which addresses an experience, ideas on a topic, future plans or explaining feelings.

Developing Domain Relevance in the Speaking Tests

Migrants need to be able to engage in interactions giving personal information, opinions and describing feelings. Effective verbal communication skills enable migrants to actively engage in life in a new country. Speaking skills enable integration in terms of employment, education and social interaction. The LCG Speaking Test includes opportunities for candidates to engage in interaction about themselves and their opinions, exchange views and state advantages and disadvantages. Candidates also initiate and respond in role plays designed to replicate a range of workplace scenarios or situations in everyday life. In another part of the test, candidates read a short text aloud and answer some follow-up questions. Such a task is intended to replicate reading aloud in a workplace, study or social context. Follow-up questions require the ability to report, paraphrase and recommend. Finally, candidates prepare and deliver a short presentation on a given topic in an occupational or personal context. Candidates have an opportunity to express and justify their thoughts, view and opinions in the presentation.

Developing Domain Relevance in the Marking Criteria

LCG and LCA both use an analytical mark scheme for all tasks in the Speaking Test and individual task-based mark schemes for the two tasks in the Writing Test. In the Writing test, the two examiners use the same markschemes and the same analytical criteria. In Speaking, the interlocutor awards marks for 'Task Fulfilment and Communicative Effect' which is, in effect, a holistic 'global achievement' scale while the second examiner who listens to the recording retrospectively awards marks against analytical criteria. The application of the marking criteria to each respective domain reflects the nature of the domain-specific tasks in the examinations and outlined in this chapter. For example, under task fulfilment in an LCG writing task, examiners are looking for an appropriate genre and tone when candidates respond to a task requiring an email in a work context. This differs from the LCA test, where the writing tasks require the ability to present relevant information, develop arguments, 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 and rating is based on 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 candidate scripts which illustrate the features referred to above, and in the mark schemes.

Reliability and Scoring

LCG is a four-skill test that reports performance across multiple levels (the IESOL SELT tests are single-level). This extension in the reporting capability is in response to demands (from both candidates and recognising institutions) for a practical and effective multi-level test. LCG is focused on the B1 and B2 levels but also measures at A2 and C1. Compared to the original IESOL B2 test, LCG has an increased number of items (from 26 to 30) to facilitate a greater spread of 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, 2023). 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 by 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 a concordance study which is 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.

The LCG test is a multi-level assessment, as mentioned, and measures across levels. LanguageCert research (Lee et al., 2023) has shown that, while the IESOL SELT level-based tests assess at their target CEFR levels, they contain an appropriate number of items to allow assessment across levels. The IESOL SELT B2 examination, for example, has items which assess above and below B2. The ability to measure and report candidate ability across a range of levels is useful for candidates and stakeholders who make decisions informed by candidate results.

Lee et al.'s (2023) study explored the alignment of LanguageCert IESOL SELT tests in relation to the two objectively marked components of Listening and Reading. The use of externally referenced anchoring demonstrated the robustness of the CEFR test levels. For example, in the case of LanguageCert IESOL SELT B2 test, most accurate measurement was observed across two CEFR levels (B1 and B2) and reasonable measurement was observed at the lower end of C1 (see Lee et al., 2023).

The value and utility of a test that measures across multiple levels on a common scale are heightened in LCG (and LCA). 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 original IESOL tests supported 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 content types in the 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 Guidelines (IWGs). The materials are created by experienced LanguageCert writers and reviewers. Used in combination with calibrated anchor items, LanguageCert is confident that both tests assess across the stated ability range effectively. This is reinforced through ongoing internal and external validation 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 reports this both overall and for the individual Listening, Reading, Speaking and Writing skill tests.

Measurement Scale

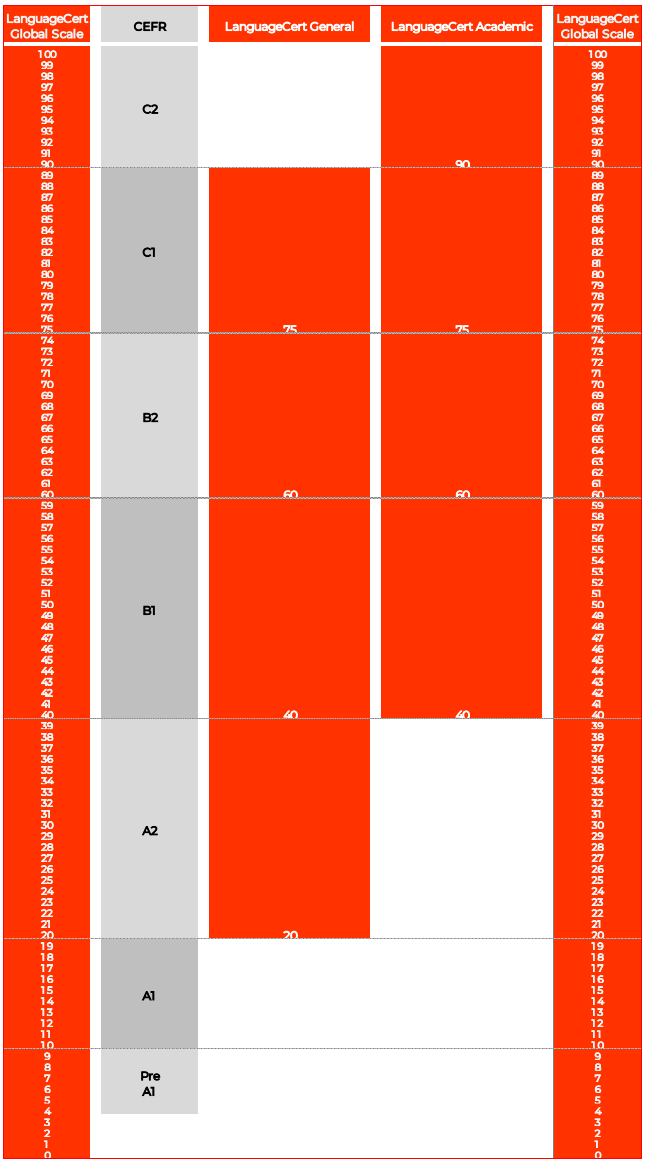
The Global Scale is used to measure each candidate's performance – see Milanovic et al. (2023). The Global Scale reports scores on a 0 to 100 scale. Candidates receive a score for each skill on the Global Scale, as well as a CEFR level based on the alignment of their total score with the Global Scale. The Global Scale corresponds directly to LANGUAGECERT's internal LID (LANGUAGECERT Difficulty) scale.

The LID scale has been in use since 2016. It is a scale of difficulty used for internal item banking and test construction purposes. The LID scale was developed using a combination of expert judgement and statistical analyses. Up to ten expert consultants, each of whom had over 20 years' experience 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 a difficulty value on the LID scale (see Lee et al, 2023).

An in-depth analysis was conducted on all anchor items and adjustments made where necessary. Rasch and Classical Statistics analyses were then conducted on all live and trial tests, leading to the majority of test items in the bank now considered as being fully calibrated. Research and validation studies in this area are provided 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 may be seen to be directly comparable to examinations provided by other English language testing organisations, such as IELTS, Cambridge Advanced and the China Standards of English (CSE) scale. Figure 3 illustrates how the Global Scale reports against the CEFR levels. These findings are under ongoing review in the LANGUAGECERT concordance study which is currently underway.

Figure 3: The LANGUAGECERT Global Scale



The three-page candidate report (Appendix 1) reveals how the LANGUAGECERT Global Scale is operationalised.

The Global Scale allows ease of interpretation for test users and provides a finely- tuned results service across all language skills. As shown, performance can be separated both by each skill and overall, so that a candidate is not only described as having "B2 ability", but a more precise level of detail is provided on a candidate'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. The LTE has been successfully administered to tens of thousands of candidates worldwide, and the Global Scale has received good customer feedback in terms of its simplicity, clarity, and ease of use.

Items in the LCG Reading and Listening Tests range in difficulty from CEFR level A2 to C1, with the majority of items focusing on the B1 and B2 levels (Intermediate and Upper Intermediate). Item difficulty is established through pre-testing and live test calibration using Rasch and Classical Statistics. All Reading and Listening items are calibrated to 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 Statistics are described in Falvey and Coniam (2023) and reveal that this method of calibration is demonstrably robust.

Each LCG Reading and Listening Test is designed to cover a wide range of the B1/B2 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.

For the LCG Writing and Speaking Tests, detailed mark schemes are used by examiners. In Writing, candidates complete two writing tasks. Task 1 requires candidates to produce a short letter, email or report of approximately 100 to150 words covering three required pieces of information in response to a short input text. In Task 2, candidates need to produce a slightly longer piece of informal writing – either an informal email, a narrative or descriptive text or article of 150 to 200 words which addresses an experience, ideas on a topic, future plans or explaining feelings.

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

1. Task Fulfilment
2. Accuracy and Range of Grammar
3. Accuracy and Range of Vocabulary
4. Organisation and Coherence

The use of separate criteria to measure different aspects of Writing performance allows the LCG test to deliver rich feedback to both candidates and receiving organisations and provides indications as to where further development is needed by the candidate. The marking criteria have been adapted from the LANGUAGECERT IESOL B2 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 fine-tune the wording of the criteria so that examiners find them easy to use, so that they reflect candidate output, and so that the key features expected from candidates in the examination at each CEFR level are considered.

The criteria have been extended to measure performance across a broader range of ability (from A2 to C1) 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 the marks awarded, the script is passed to a third (more senior) examiner whose decision is final. It is intended, that in the medium to longer term, auto-marking by computer will be introduced as part of a hybrid scoring approach.

For Speaking, the test is split into four parts. Part 1 involves responding to transactional questions across a range of topics. In Part 2, candidates take part in role-plays which are set in a range of real-life scenarios relevant to a migrant living, working or studying in an English-speaking context. In Part 3, candidates read aloud a short passage of around 80 words in length on a topical issue and answer follow-up questions selected from a list by the interlocutor. In Part 4, candidates talk about a topic selected by the interlocutor for up to two minutes. The candidate has preparation time, and after giving their talk they then answer follow-up questions selected from a list held by the interlocutor.

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

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

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

The criteria have been adapted from the IESOL B2 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, 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 candidate output, and so that the key features expected from candidates at each CEFR level are considered.

The criteria have been extended to measure performance across a broader range of ability (from A2 to C1).

Currently, candidate 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 recording. The first criterion *Task Fulfilment and Communicative Effect* is marked by the interlocutor and provides a general impression score that contains elements of the more analytical criteria used by the second examiner. The second examiner marks the other analytical criteria. The interlocutor general impression mark is then double-weighted. If there is a significant difference in marks awarded by the two examiners, 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 approach. A hybrid assessment model will garner the proven benefits of both human and machine marking (see e.g., Babitha et al., 2022).

Test Development Process and Quality Assurance

LANGUAGECERT's Assessment Development department contains academics as well as professional linguists and assessors, who publish research on all aspects of the 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 by high-calibre test developers operating to clear guidelines, workflows, and quality assurance protocols which include layers of reviews, editing, statistical analyses, and vetting. The LANGUAGECERT proprietary item bank is used to manage all LANGUAGECERT tests, with strict access protocols, and robust workflows for process compliance. LANGUAGECERT's team of examiners includes expert Chief Examiners as well as Examiners and their Team Leaders. All undergo stringent training before marking live papers. A defined marking process operates within the PeopleCert proprietary marking application, which standardises, and quality assures the process and its outputs. All candidate 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 candidate.

To explore whether any subgroup of candidates sitting a test is being unfairly disadvantaged, LANGUAGECERT addresses the challenge at a number of levels. The process starts with comprehensive item writer guidelines and item writer training. This is then supplemented by the detailed vetting and editing of test materials with a focus, amongst other things, on whether there is a risk of candidates of specific backgrounds being disadvantaged. In addition, differential item functioning (DIF) analyses – the key to investigating and dealing with test bias – are conducted. Coniam and Lee (2021) describe DIF analysis conducted on IESOL examinations delivered from 2018 to 2021 (with some of the populations involving IESOL examinations delivered for the UKVI scheme). With gender, typically a key variable in the exploration of DIF, there was a very low incidence of 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 examinations.

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 candidates 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 candidates, and their cultural contexts.

Both LANGUAGECERT's Item Writer Guidelines and the training process stress bias awareness, and the requirement to produce materials which will not favour or discriminate against certain candidates. This entails ensuring test materials are as free from specific regional or national cultures as possible, and that topics are universal. Writers have a list of taboo topics to aid in this. These taboo topics include areas which may cause distress or distraction to candidates or relate to unfortunate experiences they may 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 candidate or beyond their life experience. The LANGUAGECERT team also take care to avoid introducing 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 LCG. 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 LANGUAGECERT General and Language Cert Academic research will become a regular feature in this publication as the qualification roll-out is widened.

Conclusion

This chapter has described how an examination evolves to ensure the target language use domain is covered and provides a valid, fair, inclusive and reliable assessment tool.

The chapter has provided the rationale for the evolution of the LCG test, 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. LCG is closely based on the LANGUAGECERT IESOL B2. Its development, and the guiding body of research, has informed the ongoing review and evolution of that examination.

It has been outlined how the LANGUAGECERT General test focuses on general language requirements for use in the migrant employment target language domain. The test has been 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. All these factors underscore the care taken to employ the best research findings, methodology, and statistical procedures in order to develop and improve the quality the test.

References

ALTE (Association of Language Testers for Europe). (2016). *Language tests for access, integration and citizenship: An outline for policy makers*. Strasbourg, France: Council of Europe.

ALTE (Association of Language Testers for Europe). (2020). *ALTE linguistic integration of adult migrants: Requirements and learning opportunities*. Strasbourg, France: Council of Europe.

Babitha, M. M., Sushma, C., & Gudivada, V. K. (2022). Trends of artificial intelligence for online exams in education. *International Journal of Early Childhood Special Education, 14*(01), 2457-2463.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

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

Brindley, G., & Burrows, C. (2000). *Studies in immigrant English language assessment*. Sydney, Australia: NCELTR, Macquarie University.

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. (2021). Washback: Looking backward and forward. In *The Routledge handbook of language testing* (pp. 136-152). London, UK: Routledge.

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 adaptive test*. London, UK: LanguageCert.

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

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg, France: Council of Europe.

Falvey, P., & Coniam, D. (Eds.). (2023). *Certifying quality in assessment & learning: Research and validation at LanguageCert. Volume 2*. London, UK: LanguageCert.

Jones, C. (2023). *Orienting the LanguageCert IESOL C1 examination to an academic context*. London, UK: LanguageCert.

Knoch, U. (2021). *A guide to English language policy making in higher education*. International Education Association of Australia (IEAA). Retrieved from http://www.ieaa.org.au

Lee, T., Papargyris, Y., Milanovic, M., Pike, N., & Coniam, D. (2023). *Aligning LanguageCert SELT tests to the LanguageCert Item Difficulty (LID) scale*. London, UK: LanguageCert.

Milanovic, M., Pike, N., Papargyris, Y., Lee, T., & Coniam, D. (2023). *The LanguageCert Global Scale*. London, UK: LanguageCert.

Saville, N. (2009). Language assessment in the management of international migration: A framework for considering the issues. *Language Assessment Quarterly, 6*(1), 17-29.

van Ek, J. A., & Trim, J. L. M. (1990a/1998a). *Threshold 1990*. Cambridge, UK: Cambridge University Press.

van Ek, J. A., & Trim, J. L. M. (1990b/1998b). *Waystage 1990*. Cambridge, UK: Cambridge University Press.

van Ek, J. A., & Trim, J. L. M. (2001). *Vantage*. Cambridge, UK: Cambridge University Press.

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

Appendix 1: Candidate Report

Language Cert

LanguageCert General
(Listening, Reading, Writing, Speaking)

Test Report

Candidate Information

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

Test Centre Information

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

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

Listening	73	**Writing**	66
Reading	68	**Speaking**	71

Total Score	69
CEFR Level	B2

Signature LanguageCert Chairman

Name of LanguageCert Chairman

THIS IS NOT A CERTIFICATE

info@languagecert.org

Language Cert

Candidate Performance Feedback (Writing Part 1)

Task Fulfilment	You addressed all three points of the task and expanded some of your points further. You were able to communicate clearly. Your genre and tone were mostly appropriate. The reader was, on the whole, informed.
Accuracy and Range of Grammar	You used a range of simple grammatical forms accurately and you were able to attempt and have some success with a few complex forms. Some errors occurred which, at times, prevented understanding. Re-reading was sometimes needed.
Accuracy and Range of Vocabulary	Your use of everyday vocabulary was accurate and of an adequate range for the task. There was some misuse of less common words and phrases. A few spelling errors occurred. Any errors did not prevent comprehension, but some re-reading was needed.
Organisation and Coherence	Your text was often well-structured and clear. You made use of a variety of linking words and phrases, usually successfully. You used paragraphs mostly appropriately and your organisation was appropriate to the text type. There were some punctuation errors, but these did not prevent comprehension

Candidate Performance Feedback (Writing Part 2)

Task Fulfilment	You satisfied the demands of the task to a large extent and fully addressed both points. You could have expanded a few of your points further, but your communication was largely successful. Your genre and tone were highly appropriate. The reader was highly informed.
Accuracy and Range of Grammar	You used a range of simple grammatical forms accurately and you were able to attempt and have some success with more complex forms. A few errors occurred but these did not prevent comprehension. Some re-reading may have been needed.
Accuracy and Range of Vocabulary	You used a range of vocabulary, including less common words and phrases, effectively. A few errors in usage, spelling and word formation occurred but did not prevent comprehension. Minor re-reading may have been needed. There were very few spelling errors
Organisation and Coherence	Your text was often well-structured and clear. You made use of a variety of linking words and phrases, usually successfully. You used paragraphs mostly appropriately and your organisation was appropriate to the text type. There were some punctuation errors, but these did not prevent understanding

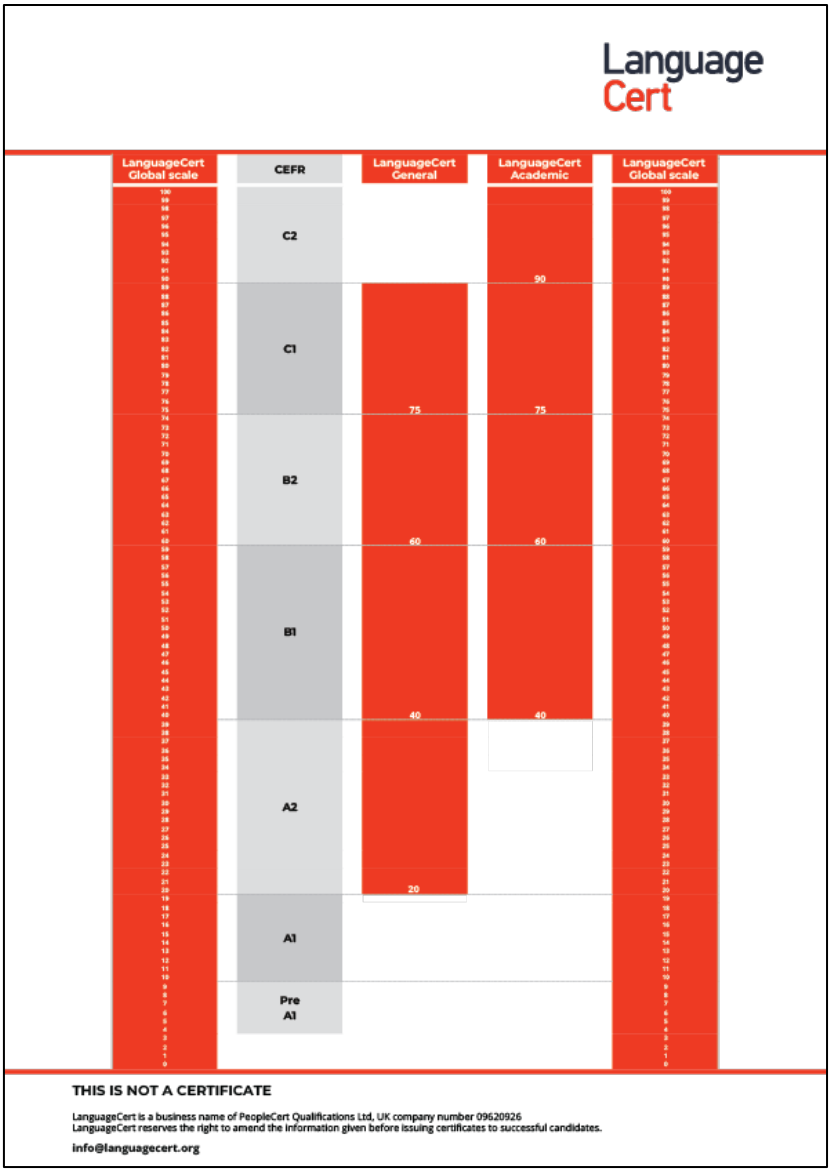
Candidate Performance Feedback (Speaking)

Task Fulfilment and Communicative Effect	You completed all tasks with ease and, at times, confidence. You communicated what you wanted to say mostly clearly and almost always in a natural manner. Misunderstandings were very rare. Your contributions were always relevant and often fully detailed.
Coherence	Your use of language was clear and often well-structured. You were able to speak at length; some hesitations may have caused coherence to breakdown. You used linking words and phrases almost always effectively.
Accuracy and Range of Grammar	You used a range of grammatical structures. Some errors occurred but were repaired or did not impact in a major way on meaning. You had very good control of basic structures. You attempted complex structures, but errors occurred, and restarts were necessary. Your meaning was almost always clear despite errors.
Accuracy and Range of Vocabulary	You used a very good range of vocabulary. Your use of vocabulary was clear. You were able to use some less common words and phrases and idiomatic language. Occasional errors in usage occurred but did not prevent understanding.
Pronunciation, Intonation and Fluency	You generally maintained a spontaneous flow of language; any hesitations did not strain the listener. During the read-aloud task, you spoke with a high level of naturalness and fluency. Your pronunciation was clear and easily understood. You used stress and intonation appropriately to convey meaning

Language Cert

CEFR Level	Scaled Score	Performance Descriptors (Listening, Reading, Speaking, Writing)
C2	90 - 100	• Can understand with ease any kind of spoken language, provided there is familiarity with the accent. • Can read with ease virtually all forms of the written language, including abstract or linguistically complex texts. • Can produce clear, smoothly flowing well-structured speech with an effective logical structure which helps the recipient to notice significant points. • 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.
C1	75 - 89	• Can understand an extended speech even when it is not clearly structured and when relationships are only implied. • Can read and understand long and complex texts, appreciating distinctions of style. • Can give clear, detailed presentations on complex subjects, integrating sub themes, developing points and rounding off with an appropriate conclusion. • Can write clear, well-structured texts on complex subjects, underlining relevant issues, expanding and supporting points of view with subsidiary points, reasons and examples, and rounding off with an appropriate conclusion.
B2	60 - 74	• Can understand extended speech and lectures and follow complex lines of argument provided the topic is reasonably familiar. • Can read and understand articles and reports in which the writers adopt particular attitudes or viewpoints. • Can give clear, detailed presentations on a range of subjects related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples. • Can write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options.
B1	40 - 59	• Can understand the main points of clear standard speech on familiar matters. • Can read and understand texts that mainly consist of high frequency everyday language. • Can reasonably fluently give a straightforward description of subjects within his/her field of interest, presenting it as a linear sequence of points. • Can write a text on a subject of personal interest, using simple language to list advantages and disadvantages and give his/her opinion.
A2	20 - 39	• Can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevanc • 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	• 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.

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.


Language Cert
LanguageCert Global scale
CEFR
LanguageCert General
LanguageCert Academic
LanguageCert Global scale
C2
C1
B2
B1
A2
A1
Pre A1
90
75
75
60
60
40
40
20
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

Chapter 3: Recalibrating and Extending the Analysis of the LANGUAGECERT Test of English

Nigel Pike, Yiannis Papargyris, Corina Dourda, Tony Lee and David Coniam

Abstract

This chapter summarises a number of validation research projects carried out on the LANGUAGECERT Test of English (LTE). Undertaken over a number of years, this work underpins the creation of the underlying LANGUAGECERT Item Difficulty and Global Scales and aims to provide a single source of confirmatory evidence that the LTE system is a robust measurement tool in both linear and adaptive forms.

The chapter extends and builds on analyses and calibration of the LTE system and, in particular, the adaptive test. This extensively-used test is drawn from the LTE item bank, which is also used to generate linear paper-based LTE tests. The particular adaptive test bank referenced in this chapter consists of over 800 items and provides the basis for the studies reported below.

Keywords: LANGUAGECERT Test of English (LTE), validation, item bank, adaptive test

Introduction

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 (at CEFR level B1, for example) but also when developing tests which measure across multiple levels from A1 to C2. The master LTE item bank contains thousands of items, calibrated in terms of their difficulty on the LID scale which runs from 50-170. Candidate results are reported against the CEFR (the Common European Framework of Reference for Languages) levels, as well as the LANGUAGECERT Global Scale which is aligned to the CEFR as laid out in Table 1 below. This scale is used with the full range of LANGUAGECERT tests and allows for level comparison between tests, where appropriate and alignment of the tests to the CEFR for ease of reference.

Table 1: CEFR level, Global Scale results reporting and the LID scale

CEFR level	Global Scale score	LID scale range (item difficulty)	LID scale midpoint
A1	11-19	51-70	60
A2	20-39	71-90	80
B1	40-59	91-110	100
B2	60-74	111-130	120
C1	75-89	131-150	140
C2	90-100	151-170	160

The LID scale was created between 2017-2019 on the basis of Classical Test Statistics (CTS) and expert judgement. Subsequent phases of measurement scale development for LTE build on the original LANGUAGECERT Item Difficulty scale using Rasch analysis – in addition to expert judgement and CTS. The enhanced LID scale forms the empirical basis for the alignment of all current and future LANGUAGECERT assessments to the same measurement scale that is itself aligned to the Common European Framework of Reference for Languages (CEFR).

All of the studies in this chapter have the objective of establishing the robustness of the items used in the LTE tests, and the candidate results which emerge from the administration of the adaptive test (for an overview of the functioning of the adaptive test, see Pike and Coniam, 2021). The first section below describes the initial calibration studies; the following section outlines two simulation studies aimed at evidencing the stability of this adaptive test. The first simulation study explored potential future item bank stability via imputing and analysing a larger dataset; the second involved constructing tests from the item bank, administering those then-live tests to target sets of candidates and analysing the outcomes, i.e., candidate and item performance. Both studies indicate a robust item bank.

In addition to the perspectives of robustness and stability as judged by item and test quality, two studies report on candidates and their backgrounds. The first provides a picture of the composition and background demographics of candidates who have taken the LTE over the three-year period 2020-2023. The second study explores potential bias among candidates in terms of whether any of the eight item types was unfairly disadvantaging any subgroup of candidates.

Initial Calibration Studies

With a view to providing background to the analysis conducted, this section reports on four related studies.

Phase 1 of the analysis (Coniam et al., 2021a) took place in early 2021, and involved an analysis of four level-agnostic (i.e., which generated results from A1 to C2) paper-based (PB) tests comprising 364 items which had been administered to over 2,000 candidates in a number of countries. This study established a baseline measurement scale. Having calibrated the four tests onto a single scale using Rasch measurement, the embryonic scale was then aligned to the original LID scale. Rescaling the calibrated scale from standard logit values to a mid-point of 100 with a spacing factor of 20 resulted in a scale which was comparable to the original LID/CEFR level scale.

The calibrated Rasch scale produced from the four LTE paper-based tests which were seen to be well aligned to the LID scale then provided the baseline for further integration of LANGUAGECERT products onto the common scale and validated the use of expert judgement and CTS in the original LID scale creation. All the items in the four paper-based tests are drawn from the overall LTE item bank and many of the items also feature in the adaptive test.

Phase 2 (Coniam et al., 2021b) involved an analysis of the adaptive test, which in mid 2021 consisted of over 800 items and 5,870 candidates. In the results, item and person reliabilities were both high. Rasch fit statistics – item and person infit and outfit mean squares – were well within acceptable ranges (i.e., 0.5 – 1.5), with the calibration statistics pointing to a test that could be viewed as sound.

It is worth noting that the calibration of the adaptive test – in terms of both item and candidate numbers – led to an improvement in the rigour of the LID scale with regard to percentile ranges and item distribution means. The scale mid-point (the 50th percentile) was 100 (99.92), closely matching the item distribution mean of 100.76. Following on, and everything else being equal, the mid-range ability group would be expected to occupy the major central region of the distribution while the higher and lower ability groups would be expected to occupy the upper and lower narrower range of ability. This indeed emerged to be the case: levels A1 and A2 fell below the 25th percentile, levels B1 and B2 between the 25th and 75th percentiles, and C1 and C2 in the top 25th percentile.

This positive picture notwithstanding, the sample size of 5,870 candidates was not considered to be sufficiently large to make definitive predictions about the robustness of the adaptive test. To this end, two approaches were seen as necessary. First, simulation studies (involving larger candidate sample sizes) would be conducted. Second, once the adaptive test had reached a comparatively large sample size (in the region of 50,000 candidates), the analyses in Phase 2 would be redone.

Confirming Item Bank Stability

With the purpose of examining the stability of the LTE adaptive test 1.0 from both statistical and operational perspectives, two simulation studies have been conducted with imputed large candidate sample sizes.

The first simulation study (Lee et al., 2022) was undertaken in late 2021, at which point, the adaptive test item bank comprising 827 calibrated items had been administered to over 13,000 candidates, each of whom had taken 58 items. In the study, performance in the 13,000-candidate live dataset was compared with a simulated much larger dataset generated using model-based imputation. Simulation regression lines showed a good match and Rasch fit statistics were also good: indicating that items comprising the adaptive test could be seen to be of high quality both in terms of content and statistical stability. Potential future stability was confirmed by results obtained from a Bayesian ANOVA.

The second simulation study (Coniam et al., 2022) built on the previous study, although with a different – real-world – focus, i.e., producing live tests from the LTE adaptive test, administering them to actual candidates and analysing the results. This process therefore involved submitting the adaptive test to a real-world test in that the quality of actual tests derived from the adaptive test was scrutinised. Three paper-based tests were compiled from the calibrated adaptive test and administered to target candidate groups. 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 adaptive test. The second simulation study concluded with the claim that the items comprising the adaptive test were well set, and that the master LTE item bank (in its entirety, that is) was sufficiently robust to be used as a clearing house from which many different tests could be constructed. The caveat nonetheless remained that the analysis needed to be redone once a large candidate sample size – in the region of 50,000 – had been reached.

Confirming Fitness for Purpose

As of mid 2022, the adaptive test (comprising 827 items) had been administered to over 48,000 candidates. The studies described below are designed to confirm that the measurement characteristics remain stable with high volumes of candidates and that the item types used are fit for purpose.

The first recalibration study reported below (recalibrating the LTE adaptive test) builds on the research and analysis reported above, with two studies reported upon. This study updates the mid 2021 initial calibration study, which comprised 827 items and 5,870 candidates.

The second study extends the scope of the analysis – from analysing all (827) items in the adaptive test as a single entity – to a more fine-grained analysis, exploring the relative difficulty of the four different listening and four reading item types in the adaptive bank.

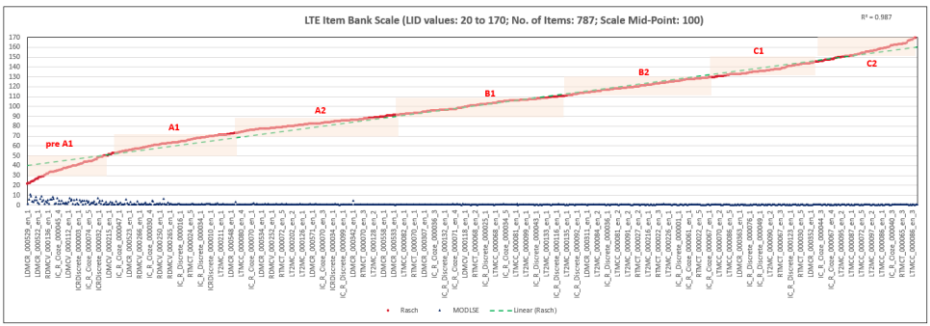
Adaptive Test (Re)calibration

As mentioned, as of mid 2022, over 48,000 candidates had each been administered 58 items via the adaptive test.

Following the methodology adopted in Lee et al. (2021), the 827 items were recalibrated using the midpoint of the scale (100, that is, B1) – in line with the previous calibration methodology. Of the 827 items, 21 were calibrated above 170 – the ceiling of the LID scale while 19 were calibrated below 20 – the bottom end of the LID scale. Those items were not included in the specification of the LTE scale presented below because the 21 items with values above 170 were too difficult for candidates while the 19 below 20 were too easy. Including such items in the final specification of the scale would have skewed distributions at both extreme ends. The final scale specification therefore currently has a total of 787 calibrated items.

Figure 1 below presents the picture the 787 items and their locations across the LID/CEFR levels.

Figure 1: Item distributions (N=787) across the item bank



The distribution of items, as presented in Figure 1, emerged at about 99% linear, especially in the A1 to C2 range. Such a distribution indicated a robust LTE scale with little distortion from the expected linear progression in an ability scale. Standard errors (SE) were minimal from A1 to C2. Even at pre-A1, where standard errors were highest, the largest SE was only 10 LID scale points, half a logit, a value commonly regarded as acceptable (Zwick, 1999).

Rasch Summary Statistics

Summary statistics for the dataset analysed via the Rasch measurement software Winsteps (Linacre, 2010) comprising 48,056 candidates and 827 items is presented in Table 2 below.

Table 2: LTE item bank summary statistics

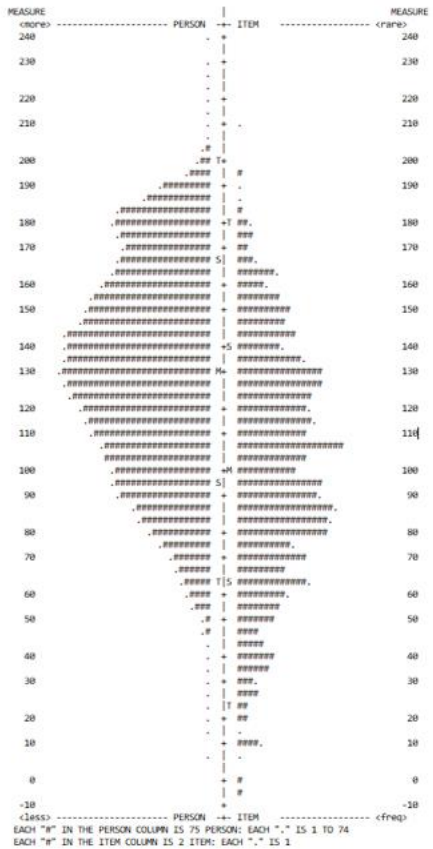
PERSON	48056	INPUT	48054	MEASURED		INFIT		OUTFIT	
	TOTAL		COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	36.3		56.7	131.31	6.52	1.00	.0	1.00	.0
P.SD	5.5		2.8	34.49	1.04	.12	.9	.39	.9
REAL RMSE	6.61	TRUE SD	33.85	SEPARATION	5.12	PERSON RELIABILITY		.96	
ITEM	827	INPUT	827	MEASURED		INFIT		OUTFIT	
	TOTAL		COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	2109.7		3295.1	100.00	1.32	1.00	-.6	1.01	-.6
P.SD	1556.9		2122.6	39.00	1.50	.09	4.8	.19	5.0
REAL RMSE	1.99	TRUE SD	38.95	SEPARATION	19.56	ITEM	RELIABILITY	1.00	

Focusing on the right-hand, blue side of the table, item reliability was high at 1.00, as was person reliability at 0.96, the latter being the equivalent of classical test theory reliability (Anselmi et al., 2019). Person infit mean-square (1.00) and outfit mean-square (1.00) fit statistics were both within the acceptable range of 0.5 to 1.5, suggesting that the calibration of persons may be taken as acceptable. By the same token, item infit mean-square (1.0) and item outfit mean-square (1.00) fit statistics were also acceptable. Overall summary calibration statistics pointed, therefore, to a test that may be viewed as sound.

Person / Item Map

Person / item maps give a useful visual representation of candidate / item distributions. In Figure 2 below person/item maps are laid out such that the candidate spread (in LID scale points) appears to the left-hand side of the ruler while the item spread appears to the right-hand side of the ruler. More able candidates are located towards the upper left side of the map while less able candidates are located towards the lower left side of the map. Similarly, more difficult items are located towards the upper right side of the map while easier items are located towards the lower right side of the map.

Figure 2: LTE item bank person / item map



The item mean was set at 100; the candidate mean emerged at 131, the bottom of C1. The candidate mean was quite bell-shaped; the item mean showed a similar distribution, if slightly irregular in places.

Table 3 presents a distribution of item values by CEFR level.

Table 3: Item values at the CEFR levels

Item Type	Pre-A1	A1	A2	B1	B2	C1	C2
N	70	94	154	151	140	112	66
Mean	37.18	60.85	80.59	100.26	120.27	139.04	158.31
SD	7.93	5.49	5.63	5.98	5.87	5.85	5.53
Minimum	21.76	50.02	70.09	90.43	110.06	130.06	150.18
25th p'tile	31.61	56.67	76.51	95.23	115.31	133.97	153.56
50th p'tile	37.89	61.20	81.16	100.66	119.77	137.82	158.12
75th p'tile	43.91	64.85	85.51	105.90	125.59	143.68	162.68
Maximum	49.92	69.88	89.81	109.77	129.62	149.82	169.63

Minimum and maximum LID scale values for levels A1 to C2 emerged very close to the LID scale range values laid out in Table 1 above based on expert-judgement-assigned values. Test means were also very close to the midpoint of each level on the LID scale. This suggests that items are assessing at the desired levels.

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 & Stahl, 1990). 1.5 to 2.0 is considered just about acceptable, with figures beyond 2.0 unacceptable.

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.

Table 4 presents the infit and outfit statistics for each CEFR level.

Table 4: CEFR level fit statistics

	A1		A2		B1		B2		C1		C2	
	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit	Infit	Outfit
Valid	94	94	154	154	151	151	140	140	112	112	66	66
Mean	0.98	1.00	0.99	1.00	0.99	0.99	1.00	1.01	1.01	1.02	1.03	1.03
SD	0.06	0.20	0.09	0.21	0.07	0.14	0.09	0.18	0.10	0.16	0.14	0.17
Minimum	0.87	0.69	0.87	0.66	0.84	0.61	0.84	0.61	0.79	0.68	0.77	0.73
Maximum	1.20	1.92	1.37	2.16	1.29	1.43	1.41	1.79	1.32	1.51	1.38	1.47

Fit statistics were good at level mean values. At the extreme ends of the scale, at the A level for example, there was a degree of misfit. This may possibly be a result of small sample response size for approximately 100 of the 827 items in the dataset.

Item Type Analysis

There are four different listening and four reading item types in the adaptive test. Tables 5 and 6 below detail the constructs assessed in each item type, the (expert judge) assumptions regarding the relative demands of each item type, and the number of items currently in the adaptive test.

Table 5 first presents a breakdown of the Listening item types.

Table 5: Listening item types: Background detail

Item type	Constructs assessed	Relative demands	No. of items
LDMCR	Understanding spoken utterances and identifying the most appropriate response. (A1/A2 & C2) and interactions (B1-C1); awareness of functional language	A1-C2	123
LDMCV	Understanding key information in short spoken utterances; a focus on numbers, dates, spellings, prices etc	A1-A2, some B1/B2	41
LT2MC	Understanding short conversations; identifying opinion (sometimes unstated at C levels), standpoint, course of action, agreement/disagreement etc	B1-C2	128
LTMCC	Understanding longer monologues/dialogues; identifying fact, detail and chronology of events etc at A2-B1, opinion, cause-effect, speaker intention (sometimes unstated at higher levels) at B2-C2	A2-C2	100

Listening item types assess a range of constructs: some at a basic, essentially factual level (identifying numbers and dates etc), while others assess at a higher cognitive level (identifying opinion, agreement/disagreement etc.) Of the 393 listening items, the majority are broadly multi-level; a comparatively small number (40) focus on lower-level constructs, targeting CEFR A1/A2 Levels.

Table 6 provides a breakdown of the Reading item types.

Table 6: Reading item types: Background detail

Item type	Constructs assessed	Relative demands	No. of items
IC_R_Discrete	Understanding of vocabulary, collocation, phrasal verbs, idioms etc	A1-C2	142
IC_R_Cloze	Lexico-grammatical knowledge; vocab, linkers, phrasal verbs, collocation etc	A1-C2	170
RDMCV	Understanding the main idea of very short texts	A1-B1	38
RTMCT	Understanding longer texts ranging from detail and fact at lower levels (A2-B1) to complex argumentation, writer intention, summarising statements, unstated opinion etc at (B1) B2-C2	A2-C2	85

Reading item types also assess a range of constructs: some at a factual level (understanding of vocabulary), while others, as with Listening, assess at a higher cognitive level (understanding writer intention, unstated opinion etc.). As with the Listening items, the majority of the 434 Reading items are multi-level; only a small number are aimed at CEFR A1/A2 Levels, focusing on lower-level reading constructs.

Tables 7 and 8 below present item type difficulty from an analysis of the responses of the 48,000 candidates to whom the items have been administered. It should be recalled that, for purposes of analysis, the test midpoint is set at 100 (B1), with an SD of 20 (refer back to Table 1).

Table 7 first provides the analysis of the four Listening item types. The final row contains the expert-assigned target level. For the sake of readability, LID values have been rounded up to whole numbers.

Table 7: Listening item type values

	N	Mean	SD	Target level
LDMCR	123	122	33	A1-C2
LDMCV	41	65	12	A1-A2, some B1
LT2MC	128	129	27	B1-C2
LTMCC	100	137	29	A2-C2

Taking the mean as a reference point, the LTMCC items were seen to be the most demanding, with a mean of 137, or low-mid C1. While this item type assesses across levels, it also assesses certain higher level listening skills. LDMCV in contrast, being pitched at A1-A2, emerged with a mean of 65, or A1. As expected, the standard deviation for this task type is also by far the lowest as the range of levels tested is much smaller.

Table 8 presents the analysis of the Reading item types.

Table 8: Reading item type values

	N	Mean	SD	Target level
IC_R_Discrete	142	115	33	A1-C2
IC_R_Cloze	165	120	41	A1-C2
RDMCV	38	61	23	A1-B1
RTMCT	85	122	35	A2-C2

Regarding reading item types, RTMCT emerged as the most demanding item type, with a mean of 122, i.e., mid B2. This was closely followed by the IC_R_Cloze item type. The easiest type was RDMCV, at 61 (pre-A1). This task type (similar to LDMCV) had the lowest standard deviation as again the range of levels tested with this item type is narrower than the other item types.

Candidate Demographics Analysis

As a lead-in to the differential item functioning (DIF) analysis which is provided in the following section, an overview of the makeup of candidates is first presented. This overview, along with a summary of demographics, gives a picture of candidates who sat the LTE adaptive test over the three-year period mid 2020 to early 2023. The overview comprises four major categories: CEFR level obtained, country, gender and age. Second, crosstabulations by CEFR Level against country and gender are presented.

Overview of Major Categories

In terms of CEFR level candidature figures, there were few candidates at the CEFR A levels. This is to be expected as LTE is also available as a paper-based test, covering levels A1-B1, which is more appropriate for lower-level candidates. 65% of candidates were at B2 level and above. While there was some variation in the candidatures at the different CEFR levels over the past three years, the patterns of achievement at the different levels were broadly constant.

Regarding country of origin, while candidates from over 100 countries sat the LTE, many country candidatures were very small. Three countries – Poland, France and Greece – accounted for the majority of the candidature. With the exception of Greece, the largest candidatures were seen at B2 level.

In term of gender split, females accounted for 58% of the candidature. From A1-B2, there were more females than males. At C1, the genders were equal. It was only at C2 that more males were observed than females.

With regards age, the under 40s accounted for almost 70% of the candidature. For all age groups – apart from the 41-50 group – B2 was the level most commonly obtained.

Differential Item Functioning Analysis

This section extends the crosstabulation analysis presented with an investigation of Differential Item Functioning (DIF) into the three key variables. DIF analysis involves an exploration of whether any subgroup of candidates sitting a test is being unfairly disadvantaged. The exploration of potential bias among subgroup types typically involves investigating variables such as gender, first language, age etc. (Ferne & Rupp, 2007).

Rasch-based methods (Roznowski & Reith, 1999) have come to be the preferred statistical mode of analysis for DIF in terms of identifying latent traits. One extension of DIF is Differential Person Functioning (DPF), which involves the grouping of items into sets that share the same latent trait (e.g., Gierl et al., 2001). With over 800 items in the adaptive test, it was decided not to focus on the item level in this study. Rather, item groups are seen to be procedurally more informative and better indicators of both candidate performance and item precision than DIF (Linacre, 2012). DPF reports biases between candidates' actual responses against the estimated Rasch-calibrated item locations. Given the general acceptance of the term "DIF", however, it is "DIF" that is referred to in the current study.

The study follows the methodology described in Coniam & Lee (2021), where bias was investigated in LANGUAGECERT IESOL Listening and Reading tests. In the current study, analysis has been conducted using the computer program Winsteps (Linacre, 2010). Since 100 is the mid-point of the LANGUAGECERT Item Difficulty scale (see Table 1 above), Rasch logit values are rescaled to a mean of 100 and a standard deviation (SD) of 20 (Coniam et al., 2021b). As mentioned, DPF involves bundling items together; the analysis is therefore conducted on the basis of the four Reading and four Listening item types.

Three key statistics are reported in the analysis below. These are laid out in Table 9.

Table 9: Key statistics reported in DIF analyses

Statistic	**Gloss**	**Comment**
N	Number of responses analysed	
Item Facility (IF)	Percentage of correct responses	0.50 is taken as the IF threshold: an indicator that candidate correct responses were not successful merely by chance.
DIF Size	Difference between actual and Rasch calibrated locations	Positive values indicate that candidate responses were higher than calibrated values, and vice versa.

In analytic terms, DIF strengths may be graded into three categories: A, B and C (Zwick, 1999). 'A' indicates negligible DIF while 'C' is the most demanding category, indicating moderate-to-large DIF (greater than 0.64 logits). In the study, the threshold of 10 LID scale points, or half a logit, is taken as the limit for indicating possible biased responses.

DIF Analyses

A detailed summary of the DIF analyses is presented below. In the analysis of item type against gender, country and age, no DIF greater than 10 LID scale points (half a logit) on any of the three variables analysed was reported. No Category C, moderate-to-large DIF (Zwick, 1999), was observed.

In the tables below, DIF size biases above (or close to) 10 LID scale points are highlighted in red.

Table 10: DIF by gender

Analysis					Commentary
Gender	**Item type**	**N**	**IF**	**DIF size**	All Item Facilities (IF) are above 0.5, so it may be taken that candidate correct responses were not merely chance guesses. All DIF sizes are small, indicating that there would appear to be no bias regarding gender in the LTE data.
F	LDMCR	361548	0.74	-0.82	
F	LDMCV	16324	0.78	-0.4	
F	LT2MC	263231	0.62	-0.7	
F	LTMCC	136009	0.59	0	
F	IC_R_Discrete	362464	0.55	0.72	
F	IC_R_Cloze	277912	0.62	0	
F	RDMCV	25304	0.57	0	
F	RTMCT	138801	0.65	0	
M	LDMCR	252237	0.75	1.11	
M	LDMCV	11601	0.78	0	
M	LT2MC	183416	0.64	1	
M	LTMCC	95039	0.61	0.42	
M	IC_R_Discrete	248794	0.59	-1.17	
M	IC_R_Cloze	193919	0.65	-0.68	
M	RDMCV	16454	0.58	-0.76	
M	RTMCT	96882	0.66	0	
n/a	LDMCR	10218	0.71	0.51	
n/a	LDMCV	680	0.79	0.77	
n/a	LT2MC	7227	0.6	0	
n/a	LTMCC	3766	0.57	-1.02	
n/a	IC_R_Discrete	10465	0.55	0.57	
n/a	IC_R_Cloze	7860	0.6	0	
n/a	RDMCV	968	0.58	-0.6	
n/a	RTMCT	3921	0.64	-2.04	(n/a = not available)

Table 11: DIF by Country

Analysis					Commentary
Country	**Item type**	**N**	**IF**	**DIF size**	All Item Facilities (IF) (with one 0.49 in the Poland data) are again above 0.5, so it may be taken that candidate correct responses were not by chance. There does not seem to be a country bias in the LTE adaptive test data.
Germany	LDMCR	5842	0.75	0	
Germany	LDMCV	102	0.84	-8.99	
Germany	LT2MC	4420	0.64	-1.31	
Germany	LTMCC	2167	0.61	-2.17	
Germany	IC_R_Discrete	6167	0.54	1.88	
Germany	IC_R_Cloze	4500	0.62	0.63	
Germany	RDMCV	237	0.64	-2.68	
Germany	RTMCT	2239	0.67	-2.65	
Italy	LDMCR	9641	0.70	0.00	
Italy	LDMCV	716	0.78	1.86	
Italy	LT2MC	6751	0.59	0.00	
Italy	LTMCC	3592	0.58	-2.47	
Italy	IC_R_Discrete	9856	0.54	1.19	
Italy	IC_R_Cloze	7416	0.59	0.00	
Italy	RDMCV	1005	0.57	0.00	
Italy	RTMCT	3706	0.64	-3.18	
Poland	LDMCR	70220	0.73	-2.36	
Poland	LDMCV	3125	0.78	-1.54	
Poland	LT2MC	51226	0.64	-4.21	
Poland	LTMCC	26628	0.61	-4.62	
Poland	IC_R_Discrete	73241	0.49	5.05	
Poland	IC_R_Cloze	54009	0.58	2.51	
Poland	RDMCV	5000	0.56	0.73	
Poland	RTMCT	27002	0.66	-4.31	
France	LDMCR	178973	0.68	0	
France	LDMCV	12115	0.76	0	
France	LT2MC	126595	0.59	-0.89	
France	LTMCC	66882	0.55	-3.06	
France	IC_R_Discrete	182742	0.54	0.58	
France	IC_R_Cloze	137463	0.58	1.19	
France	RDMCV	18466	0.56	0	
France	RTMCT	68621	0.64	-0.51	
Greece	LDMCR	341757	0.77	0.00	

Greece	LDMCV	11831	0.79	0.00	
Greece	LT2MC	252034	0.65	1.39	
Greece	LTMCC	128948	0.62	2.94	
Greece	IC_R_Cloze	262831	0.67	-1.21	
Greece	RDMCV	16976	0.59	-0.60	
Greece	RTMCT	131298	0.66	1.33	
Greece	IC_R_Discrete	331518	0.60	-1.65	
Other	LDMCR	17570	0.74	1.37	
Other	LDMCV	716	0.80	-1.16	
Other	LT2MC	12848	0.63	0.00	
Other	LTMCC	6597	0.59	-1.13	
Other	IC_R_Discrete	18199	0.55	0.71	
Other	IC_R_Cloze	13472	0.64	-1.98	
Other	RDMCV	1042	0.58	0.99	
Other	RTMCT	6738	0.67	-0.76	

Table 12: DIF by Age

Analysis					Commentary
Age	**Item type**	**N**	**IF**	**DIF size**	There would not appear to be any bias as regarding age.
under 31	LDMCR	203828	0.70	0.00	
under 31	LDMCV	11829	0.77	-0.75	
under 31	LT2MC	146001	0.61	-1.15	
under 31	LTMCC	76783	0.58	-2.97	
under 31	IC_R_Discrete	207956	0.53	1.38	
under 31	IC_R_Cloze	156660	0.59	1.08	
under 31	RDMCV	18124	0.56	0.00	
under 31	RTMCT	78260	0.64	-0.76	
31-40	LDMCR	216352	0.75	0.00	
31-40	LDMCV	9729	0.78	0.70	
31-40	LT2MC	157526	0.64	0.00	
31-40	LTMCC	81735	0.61	0.63	
31-40	IC_R_Discrete	212635	0.57	0.00	
31-40	IC_R_Cloze	166339	0.64	0.00	
31-40	RDMCV	13959	0.57	0.00	
31-40	RTMCT	83092	0.66	0.00	
41-50	LDMCR	117139	0.77	0.00	
41-50	LDMCV	3864	0.80	-1.46	
41-50	LT2MC	86621	0.65	0.49	
41-50	LTMCC	44124	0.62	1.74	
41-50	IC_R_Discrete	114675	0.58	-0.92	
41-50	IC_R_Cloze	90066	0.66	-0.43	
41-50	RDMCV	5791	0.59	-0.59	
41-50	RTMCT	44984	0.67	0.00	
51-60	LDMCR	62154	0.76	0.00	
51-60	LDMCV	1993	0.79	0.00	
51-60	LT2MC	45968	0.64	0.75	
51-60	LTMCC	23204	0.60	2.82	
51-60	IC_R_Discrete	62247	0.58	-0.97	
51-60	IC_R_Cloze	47790	0.66	-1.02	
51-60	RDMCV	3076	0.59	0.00	
51-60	RTMCT	23857	0.66	0.60	
over 60	LDMCR	23701	0.75	1.49	

over 60	LDMCV	1098	0.78	2.47		(n/a = not available)
over 60	LT2MC	17203	0.62	2.87		
over 60	LTMCC	8686	0.59	4.94		
over 60	IC_R_Discrete	23444	0.62	-2.87		
over 60	IC_R_Cloze	18198	0.68	-3.16		
over 60	RDMCV	1644	0.61	-1.07		
over 60	RTMCT	9099	0.65	1.42		
n/a	LDMCR	829	0.67	2.22		
n/a	LDMCV	92	0.75	**9.89**		
n/a	LT2MC	555	0.62	-2.99		
n/a	LTMCC	282	0.52	0.48		
n/a	IC_R_Discrete	766	0.58	-2.19		
n/a	IC_R_Cloze	638	0.59	1.12		
n/a	RDMCV	132	0.60	-3.21		
n/a	RTMCT	312	0.63	3.10		

Finally, to explore how well current results might hold in the future, a Bayesian equivalence t-test was run against the adaptive test and DIF scores. The results are provided in Table 13.

Table 13: Bayesian equivalence t-test run on LTE adaptive test scores and DIF values

					95% Credible Interval	
	N	**Mean**	**SD**	**SE**	**Lower**	**Upper**
LTE scores	319486	127.28	35.33	0.06	127.16	127.40
DIF values	319486	127.96	40.62	0.07	127.82	128.10

In Table 13 above, the means of both sets of values together with credible interval values in the LID scale range of 127 (see Table 1) are located in the middle of the B2 range. The LTE scores and DIF values scores may therefore be taken as equivalent within their respective credible intervals.

The conclusion that may be drawn from the DIF study is that LANGUAGECERT tests are as bias free as one would wish against a backdrop of tests that are carefully and professionally developed. There was no predominance of DIF on either Reading or Listening item types against country, gender or age. Results generated from the LTE adaptive test may be therefore considered fair in the context of candidate background and language skill.

Conclusion

This chapter has outlined background studies which have contributed from different perspectives to the calibrating of the LANGUAGECERT LTE via the LID scale; and to how the LTE functions operationally in terms of candidate demographics and possible item bias. The LID scale is a comprehensive scale, linked to an item bank which provides both anchoring from individual tests with different frames of reference (Humphry, 2006) and individual item-based adaptive tests. Against this backdrop, the LANGUAGECERT scale should be viewed as a hybrid scale – in that it provides the foundation for the development and creation of both standalone and adaptive tests.

The engine facilitating the construction of LANGUAGECERT tests involves a complex item banking system containing large amounts of test material. This test material covers a wide range of content and construct characteristics which has been 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 (at CEFR level B1, for example) but also when developing tests which measure across multiple levels from A1 to C2.

The current chapter has outlined a number of related background studies, with two simulation studies conducted to ascertain item bank robustness. The first simulation study explored potential future item bank stability via imputing and analysing a larger dataset; the second simulation study involved a real-world test in terms of constructing tests from the item bank, administering those then-live tests to target sets of candidates and analysing the outcomes, i.e., candidate and item performance. Both studies contributed to a picture of a robust item bank.

In addition to the perspective of robustness as judged by item and test quality, two studies have reported on candidates and their backgrounds. The first provided a picture of the composition and background demographics of candidates who have taken the LTE over the three-year period 2020-2023. The second study explored potential bias among candidates in terms of whether any of the eight item types was unfairly disadvantaging any subgroup of candidates. The Differential Item Functioning investigation into the three key variables of country, gender and age reported no major item bias.

References

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology, 10*, 2714.

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

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, T. (2021). *Potential bias in LanguageCert IESOL items: A Differential Item Functioning analysis*. London, UK: LanguageCert.

Coniam, D., Lee, T., & Milanovic, M. (2022). *Exploring item bank stability in the creation of multiple test forms*. London, UK: LanguageCert.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*(2), 113-148.

Humphry, S. (2006). *The impact of differential discrimination on vertical equating*. ARC report. Western Australia: Department of Education & Training.

Lee, T., Coniam, D., & Milanovic, M. (2022). Exploring item bank stability through live and simulated datasets. *Journal of Language Testing & Assessment, 5*, 13-21.

Linacre, J. M. (2010). *WINSTEPS (Version 3.69)* [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2012). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.

Lunz, M., & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Profession, 13*, 425-444.

Pike, N., & Coniam, D. (2021). Adaptive testing and the LanguageCert Test of English adaptive test. *ELTNEWS*, 367.

Prieto, G., & Nieto, E. (2014). Influence of DIF on differences in performance of Italian and Asian individuals on a reading comprehension test of Spanish as a foreign language. *Journal of Applied Measurement, 15*(2), 176-188.

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 4: Similarity Detection in Writing Test Scripts at LANGUAGECERT

David Coniam and Vlasis Megaritis

Abstract

This chapter explores the issues surrounding plagiarism, a form of malpractice defined as cheating by collusion, by copying, by memorisation or by using previous candidates' work in LANGUAGECERT Writing Tests. The chapter first provides an overview of the area with a discussion of how and why plagiarism is becoming more of a problem in this digital age, and a categorisation of the different types of plagiarism that are prevalent. An overview of statistical and computational methods used to detect similarity in texts follows, together with a brief description of some of the most common tools to detect similarity in texts.

The chapter then describes LANGUAGECERT's similarity detection tool SiD, which has been developed by PeopleCert for focused in-house scrutiny of all incoming scripts. To illustrate how SiD operates, and to provide a snapshot of the metric for determining similarity, exemplars of similarity at different levels of severity are then provided.

In 2023, a corpus was created of all computer-delivered LANGUAGECERT examination Writing Test scripts dating back to 2020. All computer-delivered Writing Test scripts are now passed through SiD, which examines them for similarity against the background corpus, as well as continually expanding the corpus in real time. All scripts, above a predetermined threshold of similarity, are scrutinised in order to determine whether malpractice has taken place.

The Writing Test similarity detector is just one of the tools in LANGUAGECERT's toolbox by which it ensures fairness and integrity in its examinations.

Keywords: similarity detection, cheating, writing tests, cosine similarity algorithm, Myers O(ND) algorithm

Background to Cheating with Particular Reference to Plagiarism

Cheating in examinations, including English language examinations, is a significant issue not only in academia but in classrooms around the world. With the English language becoming increasingly important for global communication, qualifications and visas for work and study purposes, the issue of cheating in English language examinations has come very much to the attention of assessment bodies and regulators. A brief overview of the literature on cheating and in particular plagiarism in English language examinations follows.

Many studies have investigated the issue of cheating in examinations. Whitley (1998) in his review of over 100 studies reported a number of reasons why students cheat on exams ranging from the importance of success, to the need for approval, to expected performance.

In a more recent large-scale survey, McCabe et al. (2012) reported that approximately 64% of students admitted to cheating on tests, while 58% admitted to some form of plagiarism. So'ud (2016) in a study of college of education students in the Sudan, reported that 100% of their interviewees admitted – for various reasons – to cheating in English language examinations. Wan and Li (2006) reported more than 60% of college students cheated at times, and about 10% cheated in examinations.

Using a variety of evidential sources, Huang & Garner (2009) reported a comparatively high level of cheating on the College English Test.

Digital content has seen a massive growth in recent years, and the internet has undoubtedly contributed to the prevalence of cheating in English language examinations. The ease of access to information and the ability to copy and paste from the internet has made it easier for students to be able to cheat (Noorbehbahani et al.,2022). While it has been argued that many students may not understand the concept of plagiarism and may not be aware of the consequences (Park, 2003), the fact is that cheating on examinations, on English language examinations, and on high-stakes English language examinations in particular, is at an all-time high, and on the increase (Iqbal et al., 2021).

To guard against cheating and malpractice, LANGUAGECERT has a rigorous set of test security principles related to online-delivered assessments (see: https://passport.peoplecert.org/docs/OLP_Exams_Candidate_Guidelines_Windows.pdf). Many of the security features echo those presented in Foster's (2013). To exemplify, upon first log-on, candidates need to follow a thorough 'onboarding' process; this includes an ID check, locking down their computer, checking there are no second monitors, and a room check through their webcam to show that the room is secure and that no other person or aids are present (see Coniam et al., 2021).

Types of Plagiarism

An array of different types of plagiarism are reported, both intentional and unintentional; see e.g., Bin-Habtoor & Zaher, 2012; Chowdhury & Bhattacharyya, 2018; Maurer, 2006. Different types of plagiarism are summarised below.

1. *Copy-and-paste plagiarism.* This is when a writer copies text from a source and pastes it into their own work without giving credit to the original author.
2. *Verbatim plagiarism.* This is when a writer copies text from a source word-for-word without giving credit to the original author.
3. *Paraphrasing plagiarism.* This occurs when a writer rephrases ideas or words without giving credit to the original author.
4. *Self-plagiarism.* This happens when a writer submits work that they have previously published without indicating that it has been published before.
5. *Mosaic plagiarism*. This is when a writer uses a combination of copied and original material in their work without properly citing the copied material.
6. *Accidental plagiarism*. This occurs when a writer inadvertently uses someone else's work or ideas without realising it, often due to a lack of understanding of proper citation practices.

7. *Structural Plagiarism*. This involves taking another person's ideas, sequence of arguments, selection of quotations from other sources, or even the footnotes that may have been used without giving due credit. Such plagiarism is not always easy to identify, as both texts have to be carefully scrutinised to identify similarities.

In the context of English language exams, types 1-3 are likely to be most prevalent and memorization is likely to play a role.

Statistical and Computational Methods Used to Detect Similarity

Over the past two decades, a considerable number of methods – which have also resulted in the development of an array of different plagiarism-checking tools – have been developed to identify similarity, or plagiarism; see e.g., Bin-Habtoor & Zaher, 2012; Chowdhury & Bhattacharyya, 2018; Maurer, 2006, who report on, review and evaluate such methods and tools.

A broad summary of the key methods is listed below.

1. *Linguistic analysis*. In this method, the language used in a text is analysed to identify patterns or characteristics – involving possible inconsistencies in writing style, vocabulary, and grammar – that may be suggestive of plagiarism (Pecorari, 2008).
2. *Database comparison*. In this method, texts are compared to a database of existing documents to identify matches or similarities. The database may be populated with previously published works, student papers, or any other relevant text that may be used for comparison (Si, et al., 1997).
3. *Citation analysis*. In this more academically-grounded method, citations in a text are analysed to determine if they are properly formatted and if they refer to valid sources. Citation analysis can also detect cases of self-plagiarism, where a writer submits work that they have previously published without proper citation (Mazov et al., 2016).
4. *Stylometric analysis.* In this method, the writing style of a text is analysed with a view to identifying patterns or characteristics – through changes in writing style or vocabulary – that may indicate plagiarism (Stein, et al., 2011).

5. *Machine learning*. In this method, machine learning algorithms are trained to detect plagiarism by analysing patterns and similarities in text. These algorithms use statistical models to identify similarities between documents and can be trained to recognise specific patterns or characteristics of plagiarism (Hunt et al., 2019).

6. *Text similarity analysis*. In this method, the text in two or more documents are compared to determine their level of similarity via algorithms which do string comparisons invoking mathematical functions. Among such algorithms are the Rabin-Karp and Jaro-Winkler distance algorithms (Leonardo and Hansun, 2017); the Levenshtein distance algorithm (Su et al., 2008); and the Smith-Waterman algorithm (Irving, 2004). Analysis is grounded on the basis that plagiarised text is likely to be similar or identical to the original source, with the algorithms producing output which reports the degree of similarity (see Vijaymeena & Kavitha (2016) for a summary of common algorithms).

As will be apparent from the detail presented below on the LANGUAGECERT similarity detection tool, the approach adopted by LANGUAGECERT, may be seen to be placed under method 6 above: text similarity analysis.

Similarity Detection Software Tools

As with other methods of detection, a number of software tools using different statistical and computational methods have been developed in an attempt to identify similarity, or plagiarism, in texts.

Bin-Habtoor & Zaher (2012) list 15 plagiarism detection tools. Naik et al. (2015) list over 30 tools. Heres & Hage (2017) compare nine tools. Chowdhury & Bhattacharyya (2018) present a survey of 31 tools, although they do not evaluate them. Mansoor & Al-Tamimi (2022) report on over 12 tools.

Summarising some of the various sources mentioned above, some of the key current software tools for detecting plagiarism are:

Turnitin is one of the most widely used pieces of similarity detection software (e.g., Meo & Talha, 2019). It uses a database of published works, student papers, and other sources to compare submitted documents for similarity. Turnitin provides a similarity score and highlights potential instances of plagiarism.

Plagiarism Detector X is a desktop application that can scan text documents for plagiarism. It uses a variety of algorithms, including text similarity analysis and database comparison, to detect plagiarism.

Grammarly is a popular writing assistant tool that can detect potential instances of plagiarism. It uses machine learning algorithms to analyse text and identify similarities to other documents.

Copyscape is an online tool that can scan web pages for plagiarism. It compares submitted text to a database of indexed web pages to identify potential instances of plagiarism.

Ephorus is a plagiarism detection tool used by educational institutions. It uses text similarity analysis to compare submitted documents to a database of published works and student papers.

Urkund is a plagiarism detection tool that can scan text documents for plagiarism. It uses a combination of text similarity analysis, database comparison, and citation analysis to identify potential instances of plagiarism.

SafeAssign is a plagiarism detection tool integrated into the Blackboard learning management system. It compares submitted documents to a database of published works, student papers, and other sources to identify potential instances of plagiarism.

The conclusion as to which software program or online tool is the best for detecting plagiarism depends on several factors, including the type of document being analysed, the type of plagiarism being investigated, and the resources available for analysis. Different tools use different algorithms and techniques to detect plagiarism, and their effectiveness can vary depending on the specific situation.

LANGUAGECERT has developed tools to automatically check the written text responses produced by candidates taking its English language exams. As an international English language exam board, operating all over the world and in different time zones the scope for cheating is significant. It is worth reiterating that checking written text is only one of the checks that need to take place to guard against cheating.

Plagiarism in English language exams may take a number of forms, as mentioned above. A serious form of plagiarism, or cheating, that LANGUAGECERT needs to detect involves essays which are significantly similar if not identical being submitted by different candidates.

The LANGUAGECERT focus rests initially on an in-house solution, relevant to scripts produced for LANGUAGECERT tests, in response to set prompts. Against this backdrop, the in-house similarity detector, SiD, has been developed which rates all input scripts for similarity against an existing corpus of past candidate scripts. The section below briefly outlines the LANGUAGECERT tool.

Background to Exploring Similarity in Texts

While the thrust of the current chapter involves a broad picture of the development and operation of the LANGUAGECERT similarity detector, some background technical detail is necessary. This section outlines, in lay terms as far as possible, some of the programming detail which underpins the operation of the tool.

The majority of the coding conducted in-house has been done in Python. This is an open-source computer programming language which consists of open-source libraries. Various of these libraries have been drawn upon in the three procedures outlined below.

In analysing candidate scripts with a view to detecting similarity, the LANGUAGECERT *Similarity Detector – SiD* – involves three core procedures. These are:

1. Vectorisation, i.e., converting the words in a text into numbers.
2. Measuring the similarity between vectors.
3. Qualitatively examining the output by highlighting similarities and differences between pairs of texts.

Procedures (1) and (2) form the core of the analysis. Procedure (3) can be viewed as the front end, where visualisations of similarities between scripts are presented to the end user. The sections below outline these procedures in the context of current implementations. Following this, procedures directly relevant to the construction and operation of the LANGUAGECERT tool SiD are provided.

Text Vectorisation

Before any comparison of texts may be conducted, the words in all texts need to be vectorised; that is, the words need to be converted into numerical representations which a software program can then meaningfully analyse. Egger (2022) presents a summary of different word (or "term-based") vectorisation techniques, with some of the most well-known described below.

The *Term Frequency - Inverse Dense Frequency* (TF-IDF) technique computes the importance a word in a document or corpus by comparing the frequency of the word in the document to its frequency across the entire corpus. It does not directly capture the meaning of the word, as it only takes into account its occurrence in the document or corpus (Ramos, 2003; Wang et al., 2020).

The *Hashing Vectorizer* is a vectorisation technique that is commonly used in natural language processing. It works by generating a fixed-length numerical representation of text data using a hashing function. Unlike other vectorisation techniques such as TF-IDF, it does not require the building of a dictionary or vocabulary (Idouglid and Tkatek, 2023).

Word2Vec is a predictive neural-based word embedding model that learns to represent words in a continuous vector space based on their contextual usage in a large corpus of text (Mikolov et al., 2013).

One of the most frequently used vectorisation techniques is the TF-IDF technique referred to above (Ramos, 2003); it is this procedure that is used in the LANGUAGECERT tool. TF-IDF was chosen because of its simplicity, its interpretability and its scalability.

Measuring Similarity Between Vectors

Once words have been vectorised, an algorithm is then required to measure the similarity between vectors. Some of the most common algorithms are outlined below.

The *Cosine Similarity* method measures the level of similarity between two vectors. It does this by calculating the cosine value of the angle between the two vectors, where the vectors are numerical representations of words in a document or a corpus (Connor, 2016).

The Manhattan Distance method computes the sum of the absolute differences or the absolute values of the differences between the corresponding dimensions or coordinates of the two points (Eugene, 1987).

The Jaccard similarity coefficient computes the relationship between words in two strings in terms of which words are shared and which are distinct (Diana and Ulfa, 2019).

The Dice coefficient defines the relationship between words in two strings as two times the number of terms which are common in the compared strings, divided by the total number of terms present in both strings (Küppers and Conrad, 2012).

Different researchers advocate different algorithms but the method adopted by LANGUAGECERT in its similarity detector is the Cosine Similarity method. The method was selected because it has been referred to as "standard" in similarity detection (Connor, 2016), and has been proven to be robust by a number of researchers (Saptono et al., 2018; Indriyanto and Sumitra, 2019; Davoodifard, 2022).

Identifying and Highlighting Differences in Scripts

Having measured the similarity between two vectors, the final step finding the differences or similarities between two pieces of text and highlighting the changes. This is the front end which is presented to users. Some of the most common difference (or 'diff') algorithms which do this are outlined below.

The Myers O(ND) difference algorithm works with textual strings. It calculates the best "difference" between two strings, which means finding the most concise sequence of 'edits' or changes to each text, such that string 1 is converted into string 2 (Myers, 1986).

Patience and Histogram algorithms enhance the Myers algorithm in certain ways to improve efficiency or performance (see Nugroho et al., 2020).

The Bentley-McIlroy algorithm operates using blocks of characters rather than single characters, as in Myers' algorithm (see Chang et al., 2008).

Although developed 40 years ago and enhanced over time (see e.g., Sjölund, 2021), Myers' algorithm is still widely used as a general-purpose difference detection tool, and is used to highlight the differences between two scripts. It is this algorithm, surrounded by a layer of pre-diff speedups and post-diff cleanups, that the LANGUAGECERT similarity detector currently uses.

The LANGUAGECERT Similarity Detector *SiD*

As outlined above, the LANGUAGECERT similarity detector (SiD) has been built following, to a large extent, well-researched best practice. Scripts input to the system are first converted into numbers using the *TF-IDF* technique. The *Cosine Similarity* algorithm is then invoked, which measures the level of similarity by calculating the cosine value between the two vectors. Myers' algorithm is then used to calculate and highlight the differences between two scripts.

The principal difference between the Cosine Similarity method and the Myers algorithm is that the Cosine Similarity is a measure of similarity between two texts which are represented as vectors without considering the relative position of words in these texts. Myers' algorithm is the front end which identifies the smallest set of insertions and deletions needed to 'transform' one sequence into the other. Appendix 1 outlines how texts which have a very high similarity score may be seen to appear qualitatively different in appearance.

The three-step operation outlined above represents the current operational state of the similarity detector. Following implementation and feedback from end users, it may be the case that the final procedure – highlighting textual similarities – may be performed by an algorithm other than the Myers', which is the procedure currently being implemented. The core operations performed by the *TF-IDF* technique and the *Cosine Similarity* algorithm which define the similarity score however, will not change.

The following section outlines the operation of the LANGUAGECERT similarity detector, SiD.

SiD in Practice

The system described below was implemented in 2022, and operationally affects all scripts coming in to the system on an ongoing, daily, basis.

A subcorpus exists for each prompt at each CEFR level. As new Writing Test prompts are created – which happens on a frequent and regular basis – new subcorpora are created to accompany the new prompts.

As scripts are input to the system, they are sorted and allocated to a specific subcorpus on the basis of CEFR level and question, i.e., prompt. All candidates have a unique identifier, so multiple takes of an examination, even responses to the same prompt, can be identified and traced.

A script which enters the appropriate subcorpus is then compared against every script that exists in the subcorpus. This means that any given script will be compared against thousands of other scripts, with a similarity score (derived from the Cosine Similarity algorithm) calculated for every script analysed.

Interpreting SiD's Output

This section presents examples of scripts from different candidates outlining degrees of similarity at various percentiles. Appendix 1 provides examples of how similar, yet apparently different, texts appear as pairs of scripts. Actual output will contain two scripts (the left-hand script being "Script 1" and the right-hand script "Script 2") presented horizontally, side by side. Any given pair of scripts need to be viewed in the context of Script 2 being 'derived' from Script 1, with coloured highlighting as outlined below.

- Green text in Script 2 indicates text which appears in Script 2 but not in Script 1.
- Red text in Script 1 indicates text which when 'deleted' from Script 1 may be seen to result in the text observed in Script 2. This may involve words and phrases in Script 1 being 'left out', 'substituted' or 're-arranged' in Script 2.
- White text indicates text which is the same in both scripts. The more white text there is, the more similar the two scripts will tend to be.

Consider Figure 1 below which is extracted from Figure 4 further down this section. The figure contains the first line from two high-similarity scripts. Apart from some minor differences, the two lines will be seen to be almost exactly the same.

For readability's sake, the lines have been enlarged, with Script 2 appearing beneath Script 1.

Figure 1: One comparable line from two similar texts

Script 1	Hey Johnsons, I hope this letter finds you well. I was thrilled to hear that you have
Script 2	Dear Johnson, I hope this letter finds you well. I was thrilled to here that you have

The first word in Script 1 is “Hey”; in Script 2, it is “Dear”.

The second word in Script 1 is “Johnsons”. In Script 2, the second word is “Johnson”, the “s” on the word “Johnson” in Script 2 having been ‘removed’.

Further along the line, “hear” in Script 1 appears as “here” in Script 2.

The preponderance of white text in the two extracts in Figure 1 underscores the high similarity between the two texts.

Appendix 1, as mentioned, provides examples of what similar, yet apparently different, texts might look like in terms of ‘deletions’ in one text (Script 1) and ‘insertions’ in another (Script 2).

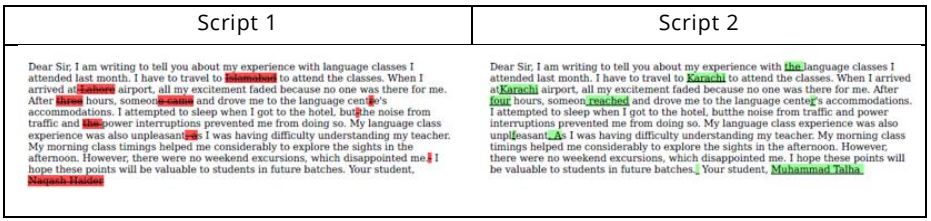
As Appendix 1 illustrates, two scripts can visually contain a comparatively large amount of red and green highlights (indicating potential differences) alongside a very high similarity score. Therefore, because of the large number of differences, generally, in the context of an examiner scrutinising two scripts which have an extremely high similarity score (above 0.9, say), the more red and green text there is, and the less white text, the lower will be the degree of similarity between the two scripts, and the less likelihood of cheating having occurred. For an examiner looking at two scripts with a preponderance of white text, a warning sign of potential malpractice is flagged, and the scripts concerned are then forwarded for more detailed investigation.

To give a flavour of the procedure in practice, and the type of output provided to a scrutinising examiner, some exemplar pairs of scripts exhibiting different levels of similarity are presented below.

One issue regarding the occurrence of similarity rests on the extent to which candidates reuse, or incorporate, detail from the prompt. Such recycling of given text is very much the case at lower CEFR levels (A1 to B1). Although recycling is less prevalent at B2 and above, a certain amount of reuse of given words and phrases still exists.

Figure 2 presents two scripts with a 0.90 similarity.

Figure 2: 0.90 similarity

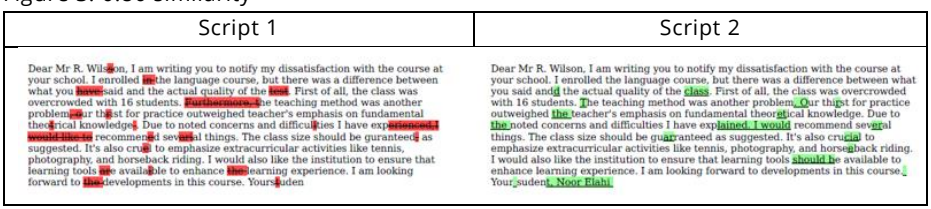


Script 1	Script 2
Dear Sir, I am writing to tell you about my experience with language classes I attended last month. I have to travel to Islomabad to attend the classes. When I arrived at Lahore airport, all my excitement faded because no one was there for me. After three hours, someon e came and drove me to the language cente r's accommodations. I attempted to sleep when I got to the hotel, but the noise from traffic and the power interruptions prevented me from doing so. My language class experience was also unpleasant as I was having difficulty understanding my teacher. My morning class timings helped me considerably to explore the sights in the afternoon. However, there were no weekend excursions, which disappointed me. I hope these points will be valuable to students in future batches. Your student, Naqash Haider	Dear Sir, I am writing to tell you about my experience with the language classes I attended last month. I have to travel to Karachi to attend the classes. When I arrived atKarachi airport, all my excitement faded because no one was there for me. After four hours, someon reached and drove me to the language cente r's accommodations. I attempted to sleep when I got to the hotel, butthe noise from traffic and power interruptions prevented me from doing so. My language class experience was also unpleasant. As I was having difficulty understanding my teacher. My morning class timings helped me considerably to explore the sights in the afternoon. However, there were no weekend excursions, which disappointed me. I hope these points will be valuable to students in future batches. Your student, Muhammad Talha

Apart from minimal changes such as place and person names and a couple of other minor differences, Script 2 is essentially the same as Script 1.

Figure 3 presents two scripts with 0.80 similarity.

Figure 3: 0.80 similarity

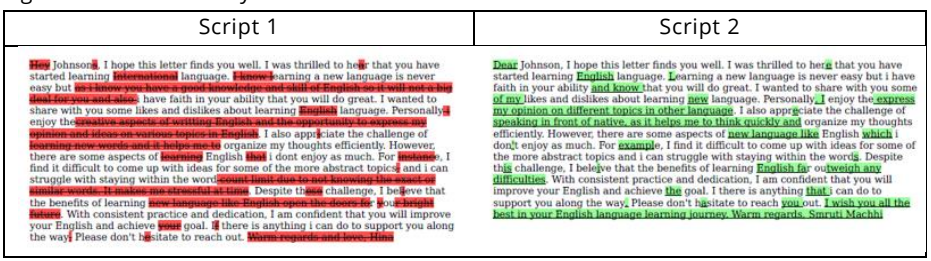


Script 1	Script 2
Dear Mr R. Wilson, I am writing you to notify my dissatisfaction with the course at your school. I enrolled in the language course, but there was a difference between what you said and the actual quality of the class. First of all, the class was overcrowded with 16 students. Furthermore, the teaching method was another problem. Our thirst for practice outweighed teacher's emphasis on fundamental theoretical knowledge. Due to noted concerns and difficulties I have experienced, I would like to recommend several things. The class size should be guranteed as suggested. It's also crucial to emphasize extracurricular activities like tennis, photography, and horseback riding. I would also like the institution to ensure that learning tools are available to enhance the learning experience. I am looking forward to the developments in this course. Yours truly, Aiden	Dear Mr R. Wilson, I am writing you to notify my dissatisfaction with the course at your school. I enrolled the language course, but there was a difference between what you said and the actual quality of the class. First of all, the class was overcrowded with 16 students. The teaching method was another problem. Our thirst for practice outweighed the teacher's emphasis on fundamental theoretical knowledge. Due to the noted concerns and difficulties I have explained. I would recommend several things. The class size should be guaranteed as suggested. It's also crucial to emphasize extracurricular activities like tennis, photography, and horseback riding. I would also like the institution to ensure that learning tools should be available to enhance learning experience. I am looking forward to developments in this course. Your sudent, Noor Elahi

While there is a high degree of similarity between the two scripts, there are notable differences.

Figure 4 presents two scripts with 0.70 similarity.

Figure 4: 0.70 similarity



Script 1	Script 2
Hey Johnson, I hope this letter finds you well. I was thrilled to hear that you have started learning International language. I know learning a new language is never easy but no i know you have a good knowledge and skill of English so it will not a big deal for you and also. have faith in your ability that you will do great. I wanted to share with you some likes and dislikes about learning English language. Personally I enjoy the creative aspects of writing English and the opportunity to express my opinion and ideas on various topics in English. I also appriciate the challenge of learning new words and it helps me to organize my thoughts efficiently. However, there are some aspects of learning English that i dont enjoy as much. For instance, I find it difficult to come up with ideas for some of the more abstract topics and i can struggle with staying within the words count limit due to not knowing the exact or similar words. It makes me stressful at time. Despite these challenge, I believe that the benefits of learning new language like English open the doors for your bright future. With consistent practice and dedication, I am confident that you will improve your English and achieve your goal. If there is anything i can do to support you along the way. Please don't hesitate to reach out. Warm regards and love. Hina	Dear Johnson, I hope this letter finds you well. I was thrilled to hear that you have started learning English language. Learning a new language is never easy but i have faith in your ability and know that you will do great. I wanted to share with you some of my likes and dislikes about learning new language. Personally, I enjoy the express my opinion on different topics in other language. I also appreciate the challenge of speaking in front of native, as it helps me to think quickly and organize my thoughts efficiently. However, there are some aspects of new language like English which i don't enjoy as much. For example, I find it difficult to come up with ideas for some of the more abstract topics and i can struggle with staying within the words. Despite this challenge, I beleive that the benefits of learning English far outweigh any difficulties. With consistent practice and dedication, I am confident that you will improve your English and achieve the goal. I there is anything that i can do to support you along the way. Please don't hesitate to reach you out. I wish you all the best in your English language learning journey. Warm regards. Smruti Machhi

The broad structure of the two scripts is comparable and hence the comparatively high degree of similarity at 0.70. There is more originality in Script 2, however, compared with the two 0.80 similarity scripts above.

Figure 5 presents two scripts with 0.60 similarity.

Figure 5: 0.60 similarity

Script 1	Script 2
To The Director, I am writing this letter regarding the issues I experienced while completing a short English language course at the Highview School. Firstly, every class accommodated almost 20 students, violating the school admission policy. Not only did it create several sitting issues, but teachers also could not deliver their lectures appropriately. In addition, due to overstrength, some chapters were left unfinished. Secondly, I was excited to learn some horse riding skills, but unfortunately, the absence of coaching staff made it almost impossible to practice the skill correctly. Moreover, safety equipments was also missing, leading participants to inflict injuries during the ride. I would request adhering to the admission policy or expanding the class size for a healthy learning environment. Furthermore, hire professional coaches and purchase some new safety gears for horse riding to avoid potential incidents. Looking forward to hearing from you. Yours sincerely, Ayesha	To The Director, I am writing this letter regarding the issues I encountered while completing a summer course at the Highview School. Every class had almost 20 students, which was against the course admission policy. Not only did it create several sitting issues, but it also distracted teachers from focusing on students individually. Therefore, we could not complete the syllabus timely. I was excited to gain some horse riding experience, but unfortunately, there was no training staff available at the stable, which was a huge disappointment. Moreover, some participants got severely injured during the ride due to the lack of safety equipment. I would request to expand the class size or ensure an adequate seating plan for a healthy learning environment. in addition, hiring seasoned training staff and safety tools for horse riding will help the students avoid future incidents. Looking forward to hearing from you. Yours sincerely, Faisal

At 0.60 similarity, the greater degree of originality in Script 2 is becoming apparent. There is much less white text.

Figure 6 presents two scripts with 0.50 similarity.

Figure 6: 0.50 similarity

Script 1	Script 2
Hello. My name is Khalid and I am here to complain about The Language Centre. I have attended the institutionbut it was a very bad educational experience. Starting from the airport arrival, noone came to recieve me at the airport. There was not any opportunity to practice the languages as all the people were strangers. I had to wait for three hours. After that time some people came and they took me to the city centre. Accomodations were too expensive. The management should have told the students to make accomodations arrangements prior to their departure so that any inshan could be avoided but nothing was told to us. Teachers were cooperative but	Hi, My name is Joshua and I am here to share my experience of attending The Language centre. It was a bad experience as no one came to recieve me at the airport. All the people were strangers and there was not any opportunity to practice the language. I was really worried at that time. Some people came and they took me to the city centre. Inquired about te accomodation rates and I found that rooms were too expensive. The admistration should have informed about the accommodation rates prior to the departure of the student so that he could make the arrangements accordingly. The lessons were also not impressive. The teachers

At 0.50 similarity, the degree of difference between the two scripts extends, although there is still some similarity – as in the two scripts above due in part to the recycling of words from the prompt.

Conclusion

This chapter has presented a picture of how LANGUAGECERT approaches and engages with the issue of similarity – potential cheating – in LANGUAGECERT Writing Tests. The chapter has outlined the working of the LANGUAGECERT similarity detector SiD in terms of how the system processes scripts within the system, and the type of output that is provided.

The identification of textual similarity and differences have been presented from two complementary perspectives – the Cosine Similarity method and Myers' O(ND) difference algorithm respectively. These metrics generate output which provides a baseline quality check in terms of potential malpractice.

As the current chapter illustrates, LANGUAGECERT takes the issues of cheating or malpractice extremely seriously. Ways in which LANGUAGECERT does this have been illustrated above. It is clear from the illustrations that the issue of similarity / cheating / plagiarism must be tackled strenuously and continuously. The LANGUAGECERT similarity detector outlined in this chapter represents but one element in LANGUAGECERT's striving to maintain honesty, integrity and fairness in LANGUAGECERT's English language examinations.

References

Bin-Habtoor, A. S., & Zaher, M. A. (2012). A survey on plagiarism detection systems. *International Journal of Computer Theory and Engineering, 4*(2), 185.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS), 26*(2), 1-26.

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.

Connor, R. (2016). A tale of four metrics. In *Similarity Search and Applications: 9th International Conference, SISAP 2016, Tokyo, Japan, October 24-26, 2016, Proceedings 9* (pp. 210-217). Cham: Springer International Publishing.

Davoodifard, M. (2022). Automatic detection of plagiarism in writing. *Studies in Applied Linguistics & TESOL at Teachers College, Columbia University, 21*(2), 54-60.

Diana, N. E., & Ulfa, I. H. (2019, March). Measuring performance of n-gram and Jaccard-similarity metrics in document plagiarism application. In *Journal of Physics: Conference Series* (Vol. 1196, No. 1, p. 012069). IOP Publishing.

Egger, R. (2022). Text representations and word embeddings: Vectorizing textual data. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications* (pp. 335-361). Cham: Springer International Publishing.

Foster, D., & Layman, H. (2013). Online proctoring systems compared. *Webinar*. Retrieved from http://www.slideshare.net/caveonweb/caveon-webinar-series-online-proctoring-best-practicesoct-2013-slideshare-final

Heres, D., & Hage, J. (2017, November). A quantitative comparison of program plagiarism detection tools. In *Proceedings of the 6th Computer Science Education Research Conference* (pp. 73-82).

Huang, D., & Garner, M. (2009). A case of test impact: Cheating on the College English Test in China. In *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment* (pp. 59-76).

Hunt, E., Janamsetty, R., Kinares, C., Koh, C., Sanchez, A., Zhan, F., ... & Oh, P. (2019, November). Machine learning models for paraphrase identification and its applications on plagiarism detection. In *2019 IEEE International Conference on Big Knowledge (ICBK)* (pp. 97-104). IEEE.

Idouglid, L., & Tkatek, S. (2023). Word embedding methods of text processing in big data: A comparative study. In *Artificial Intelligence and Smart Environment: ICAISE'2022* (pp. 831-836). Cham: Springer International Publishing.

Indriyanto, I., & Sumitra, I. D. (2019, November). Measuring the level of plagiarism of thesis using vector space model and cosine similarity methods. In *IOP Conference Series: Materials Science and Engineering* (Vol. 662, No. 2, p. 022111). IOP Publishing.

Iqbal, Z., Anees, M., Khan, R., Hussain, I. A., Begum, S., Rashid, A., ... & Hussain, F. (2021). Cheating during examinations: Prevalence, consequences, contributing factors and prevention. *International Journal of Innovation, Creativity, and Change, 15*(6), 601-609.

Irving, R. W. (2004). Plagiarism and collusion detection using the Smith-Waterman algorithm. *University of Glasgow*, 9.

Küppers, R., & Conrad, S. (2012, September). A set-based approach to plagiarism detection. In *CLEF (Online Working Notes/Labs/Workshop)*.

Leonardo, B., & Hansun, S. (2017). Text documents plagiarism detection using Rabin-Karp and Jaro-Winkler distance algorithms. *Indonesian Journal of Electrical Engineering and Computer Science, 5*(2), 462-471.

Mansoor, M. N., & Al-Tamimi, M. S. (2022). Computer-based plagiarism detection techniques: A comparative study. *International Journal of Nonlinear Analysis and Applications, 13*(1), 3599-3611.

Maurer, H. A., Kappe, F., & Zaka, B. (2006). Plagiarism—A survey. *Journal of Universal Computer Science, 12*(8), 1050-1084.

Mazov, N. A., Gureev, V. N., & Kosyakov, D. V. (2016). On the development of a plagiarism detection model based on citation analysis using a bibliographic database. *Scientific and Technical Information Processing, 43*(4), 236-240.

McCabe, D., Butterfield, K., & Trevino, L. (2012). *Cheating in college: Why students do it and what educators can do about it*. Baltimore, MD: The Johns Hopkins University Press.

Meo, S. A., & Talha, M. (2019). Turnitin: Is it a text matching or plagiarism detection tool? *Saudi Journal of Anaesthesia, 13*(1), 48-51.

Myers, E. W. (1986). An O(ND) difference algorithm and its variations. *Algorithmica, 1*(1-4), 251-266.

Naik, R. R., Landge, M. B., & Mahender, C. N. (2015). A review on plagiarism detection tools. *International Journal of Computer Applications, 125*(11), 16-22.

Noorbehbahani, F., Mohammadi, A., & Aminazadeh, M. (2022). A systematic review of research on cheating in online exams from 2010 to 2021. *Education and Information Technologies, 27*(6), 8413-8460.

Nugroho, Y. S., Hata, H., & Matsumoto, K. (2020). How different are different diff algorithms in Git? Use--histogram for code changes. *Empirical Software Engineering, 25*, 790-823.

Park, C. (2003). In other (people's) words: Plagiarism by university students—Literature and lessons. *Assessment & Evaluation in Higher Education, 28*(5), 471-488.

Pecorari, D. (2008). *Academic writing and plagiarism: A linguistic analysis*. London, UK: Bloomsbury Publishing.

Qualifications and Curriculum Authority. (2007). *Regulatory principles for e-assessment*. Retrieved from https://publications.parliament.uk/pa/cm200607/cmselect/cmeduski/memo/test&ass/ucm3102paper4.pdf

Ramos, J. (2003, December). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning* (Vol. 242, No. 1, pp. 29-48).

Saptono, R., Prasetyo, H., & Irawan, A. (2018). Combination of cosine similarity method and conditional probability for plagiarism detection in the thesis documents vector space model. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10*(2-4), 139-143.

Sjölund, M. (2021, September). Evaluating a tree diff algorithm for use in Modelica tools. In *Modelica Conferences* (pp. 529-537).

So'ud, M. D. N. (2016). The effect of cheating in English examinations on the process of the pedagogical evaluation. *International Journal of Evaluation and Research in Education, 17*(4), 145-155.

Stein, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. *Language Resources and Evaluation, 45*, 63-82.

Su, Z., Ahn, B. R., Eom, K. Y., Kang, M. K., Kim, J. P., & Kim, M. K. (2008, June). Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm. In *2008 3rd International Conference on Innovative Computing Information and Control* (pp. 569-569). IEEE.

Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal, 3*(2), 19-28.

Wan, Y. K., & Li, H. T. (2006). A study on college students cheating in examinations. *Journal of Yuxi Teachers College, 22*(3), 86-90.

Appendix 1: Categorising Similarity yet Difference in Texts

Example 1: Both texts exactly the same. *Similarity score 1.0*

Candidate Name 1	Candidate Name 2
I love summer. My favorite month is August. I eat watermelons in August. I hate winter.	I love summer. My favorite month is August. I eat watermelons in August. I hate winter.

Example 2: Changing the order of one sentence. *Similarity score 1.0*

Candidate Name 1	Candidate Name 2
~~I love summer.~~ My favorite month is August. I eat watermelons in August. I hate winter.	My favorite month is August. I eat watermelons in August. I hate winter. I love summer.

Example 3: Changing the order of two sentences. *Similarity score 1.0*

Candidate Name 1	Candidate Name 2
~~I love summer.~~ My favorite month is August.~~I hate winter.~~ I eat watermelons in August.	My favorite month is August. I eat watermelons in August. I hate winter. I love summer.

Example 4: Changing the order of all sentences. *Similarity score 1.0*

Candidate Name 1	Candidate Name 2
I ~~love summer. I eat watermelons in August~~. My favorite month is August.I ~~hate~~ winter.	I hate winter. My favorite month is August. I love summer. I eat watermelons in August.

s

Example 5: Making typos: “favoritemonth” instead of “favorite month”. *Similarity score 0.81*

Candidate Name 1	Candidate Name 2
I ~~love summer. I eat watermelons in~~ August ~~.~~My favorite~~ month is August.I hate winter.~~	I hate winter. August is my favoritemonth. I love summer. I eat watermelons in August.

“Favoritemonth“ and “ favorite month “ are considered two different words, by the scoring algorithm that is why the score changes

Chapter 5: SELT IESOL Writing Test Quality

David Coniam, Irene Stoukou, Tony Lee and Michael Milanovic

Abstract

This chapter 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 encompassed over 11,000 candidates, 60 examiners and 18 different tasks. Using principally Many-Facet Rasch Analysis (MFRA), the study explores the consistency of marking in terms of examiner, task, and rating scale fit and severity.

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: writing tests, Many-Facet Rasch Analysis, fit, severity

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 & 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 & McNamara (1995). A considerable amount of research exists on examiner reliability (Saito, 2008; Webb et al., 1990); consistency (Elder et al., 2007; Lumley & McNamara, 1995); severity (Engelhard & Myford, 2003). Other investigations of factors affecting examiners' rating have focused on: mother tongue, expertise, educational qualifications, professional background (Barkaoui, 2007; Cumming, 1990; Johnson & 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 & McNamara, 1995; Kang et al., 2019; Webb et al., 1990; Weigle, 1998). For details of the training of LANGUAGECERT Writing Test examiners, see Papargyris & 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 & 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 the 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 detail on the dataset constructed for the analysis. This comprises the four facets used in the Many-Facet Rasch Analysis (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 & 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. 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 Many-Facet Rasch Analysis (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 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 in line with the relevant test level?

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, with a lower score indicating that the task is easier. For rating scales, a higher score indicates a more demanding scale.

In the analysis below, three 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

As mentioned, the facet map is an initial visual guide, permitting 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, and reflecting the requirement 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 detail on 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

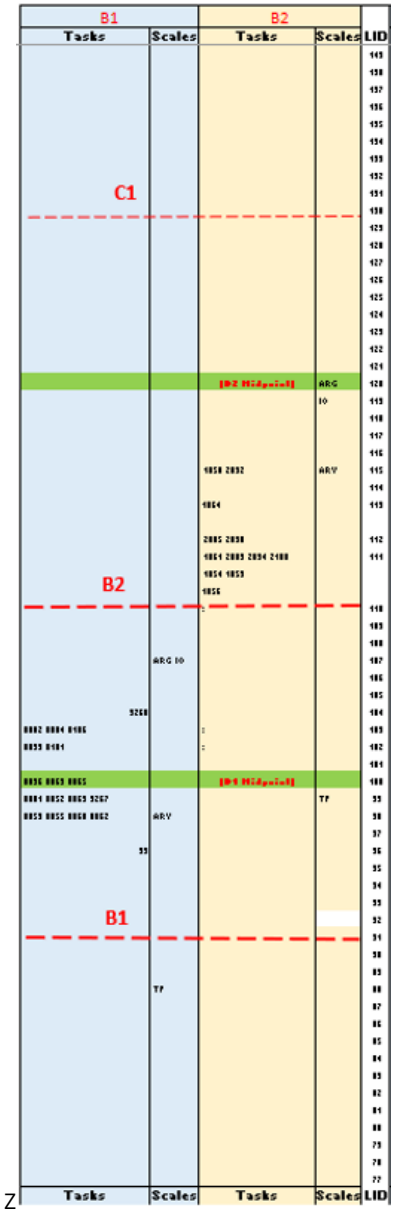
Rating scales

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 below.

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, infit, as mentioned, 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. There were three examiners exhibiting misfit at B1 and 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 TestDaF Writing test, reports an examiner severity spread of 4.26 logits. Park (2004) reports an examiner severity range of 5.24 logits.

The issue of examiner 'severity/leniency', it should be noted, 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: (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 Rafoth, 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 extends 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 & 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 sum then, in light of the analysis reported, the SELT B1 and B2 English Language Writing Tests may be seen as being robust and fit for purpose.

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* (pp. 86-107). 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 Professions, 13*, 425-444.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54*, 913-925.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). 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. https://doi.org/10.46451/ijts.2022.01.13

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet 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

Examiner	*LID*	*Infit*	*S.E.*
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

continued from previous column

Examiner	*LID*	*Infit*	*S.E.*
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

953535	126.71	1.02	1.53		1664778	66.02	0.86	1.35
1681139	77.6	1.00	1.07		1655247	111.32	0.79	3.05
1664747	53.26	1.00	1.28		2187924	75.80	0.79	1.4
28729	124.95	1.00	1.09		2433349	80.42	0.76	14.93
1664753	84.79	0.99	1.02		1858871	114.98	0.76	2.93
2116474	84.71	0.98	1.05		2248452	102.98	0.75	0.88
1676916	78.98	0.98	1.08		2228716	144.41	0.69	9.73
1681140	56.35	0.96	1.99		18078	74.83	0.68	17.05
17955	108.45	0.95	10.6		1668577	104.00	0.68	1.23
1366256	111.29	0.95	3.92		2085519	109.77	0.63	4.41
1652245	94.44	0.94	1.64		1211463	124.27	0.5	11.89
1643603	112.1	0.94	1.07		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

Examiner	*LID*	*Infit*	*S.E.*
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

Examiner	*LID*	*Infit*	*S.E.*
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)

Task ID	LID	Infit	S.E.
3268	104.07	1.05	0.91
0084	103.32	0.99	0.69
0106	103.00	1.04	0.98
0082	102.51	0.99	0.72
0093	102.25	1.00	0.69
0101	101.82	1.02	0.73
0096	100.37	0.91	0.67
0065	99.84	1.06	0.73
0063	99.57	1.01	0.65
0052	99.12	0.94	0.73
0081	98.88	0.90	0.74
3267	98.79	1.01	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

B2

(Mean: 120; SD: 20)

Task ID	LID	Infit	S.E.
1058	126.21	0.90	1.32
2092	124.4	1.05	0.93
2090	122.12	0.96	1.32
1064	121.73	0.93	1.27
2085	119.3	0.98	0.91
2100	119.08	0.97	1.27
1061	118.54	0.93	0.89
1059	118.43	0.95	0.94
2083	118.08	1.05	0.90
2094	118.01	1.02	0.90
1054	117.76	1.01	0.90
1056	116.34	0.92	0.92

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)

Scale	LID	Infit	S.E.
IO	106.93	1.02	0.35
ARG	106.81	0.81	0.33
ARV	98.37	0.79	0.34
TF	87.89	1.38	0.37

B2

(Mean: 120; SD: 20)

Scale	LID	Infit	S.E.
ARG	129.38	0.73	0.55
IO	128.02	1.21	0.59
ARV	123.67	0.81	0.56
TF	98.94	1.24	0.61

Chapter 6: Aligning LANGUAGECERT SELT Tests to the LANGUAGECERT Item Difficulty (LID) 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 a previous study which established that the LANGUAGECERT SELT B1–C1 tests are robust though the use of externally-referenced anchoring.

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.

Keywords: scale alignment, listening tests, reading tests, externally-referenced anchoring

Introduction

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 and work to the UK, since 2020.

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). The previous study (Milanovic et al., 2022) 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 constitute 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 [Note 1]. 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 represent 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 line with the key test qualities of validity and reliability (Bachman & Palmer, 2010), the LST tests assess the communicative skills that test takers will be expected to control 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 judgment 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 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 methodology surrounding externally-referenced anchoring relates to the use of Rasch measurement. 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.

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

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
C2	3	111
C1	6	581
B2	6	2,732
B1	9	10,808

Via externally-referenced, or vertical, anchoring (see 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. 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
C2	151-170	160	90-100	95
C1	131-150	140	75-89	82
B2	111-130	120	60-74	67
B1	91-110	100	40-59	50
A2	71-90	80	20-39	30
A1	51-70	60	10-19	15

Externally-Referenced Anchoring

The methodology used in the current study is based on, as mentioned, 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 Questions

The research question being pursued in the current study may be summarised thus:

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

Background Statistical Analysis

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 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 proposed 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.

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. As mentioned, 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.

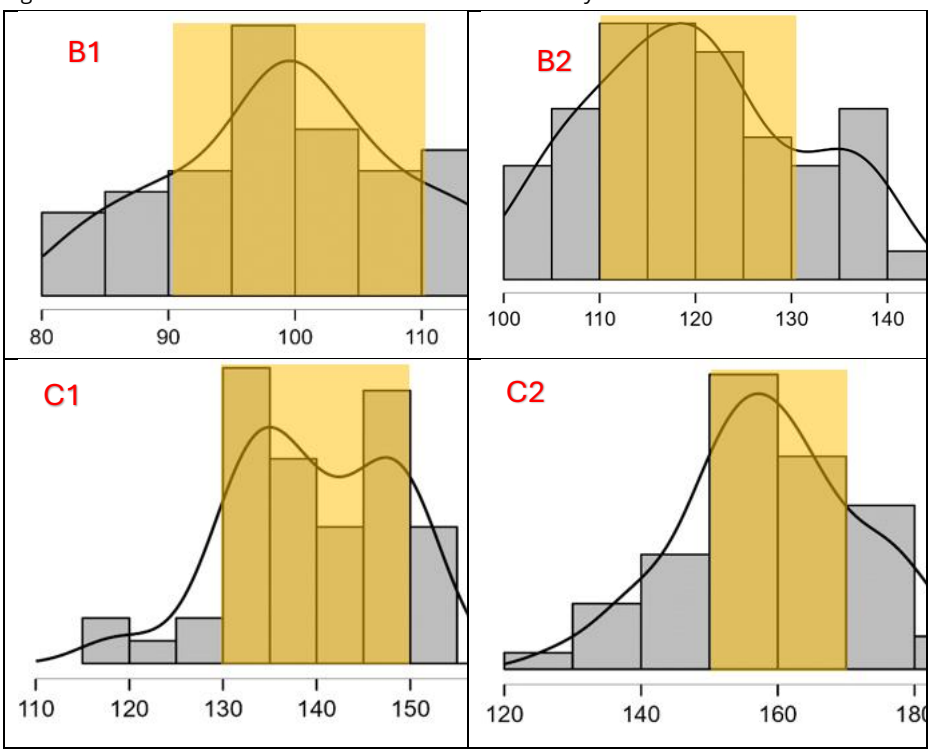
Table 3: Percentile distributions in composite LANGUAGECERT SELT Test tests

	B1		B2		C1		C2	
No. of items		52		52		52		56
Mean		100		120.00		140.00		160
SD		9.59		10.83		9.28		14.09
Maximum		119.55		141.02		165.98		198.53
75th percentile	110	**105.64**	130	**126.43**	150	**147.69**	170	**167.96**
50th percentile		**99.45**		**119.29**		**139.50**		**159.15**
25th percentile	91	**94.04**	111	**112.78**	131	**133.45**	151	**150.72**
Minimum		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. There is a degree of divergence, although this 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 targetted at their intended level.

To provide an accessible visual impression, test difficulty distributions are now presented in graph form in Figures 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.

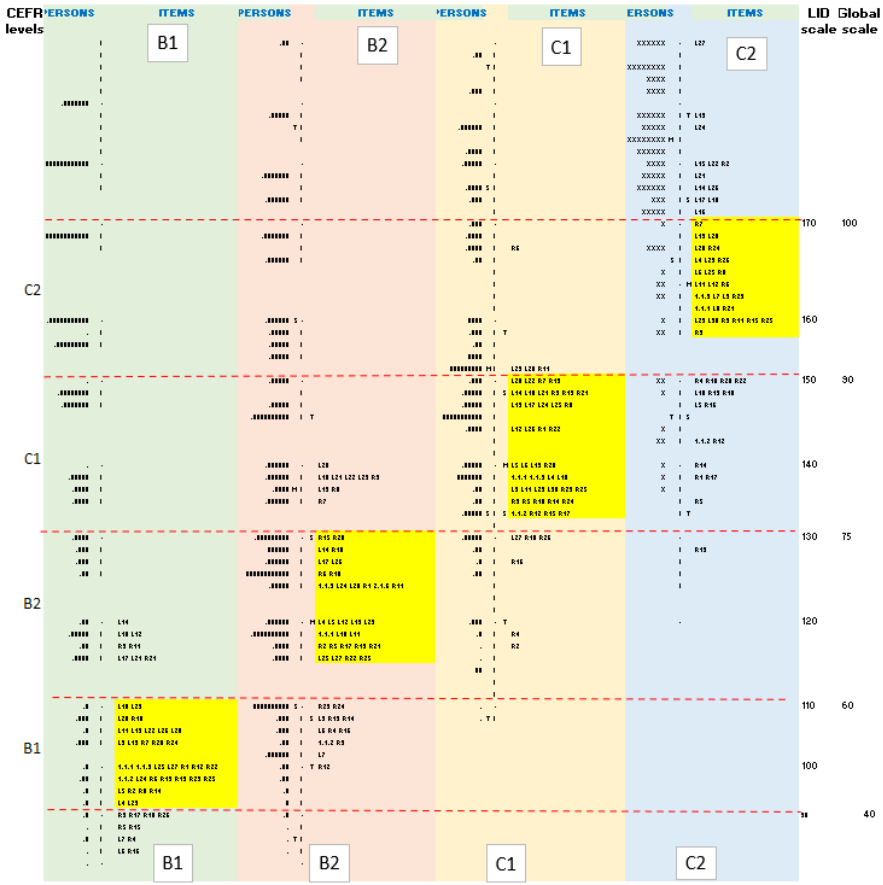
Placing LANGUAGECERT SELT on the LID Scale

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.

The maps should be read such 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 2: LANGUAGECERT SELT Test Common Scale



As can be seen from Figure 2, 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 targetted at the appropriate level.

It is also clear from Figure 2 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 B2 test, the majority of items assess at B2 level, but substantial numbers 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.

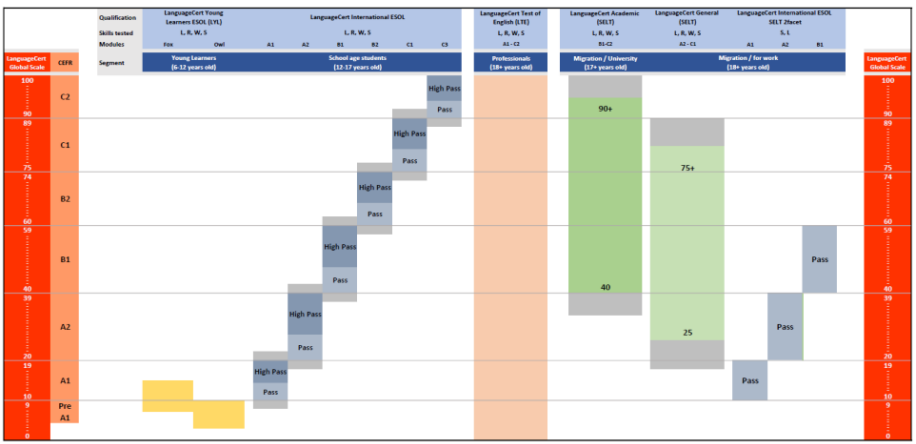
The research question pursued in the study was that LANGUAGECERT SELT tests could be accurately placed on the LID scale and hence the CEFR, accuracy being defined as good Rasch infit and outfit statistics being obtained in the analysis at each of the four test levels. Rasch levels were indeed within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

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge. https://doi.org/10.4324/9780429030499.

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. doi.org/10.46451/ijts.2022.01.12.

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

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

Zwick, Rebecca, Dorothy T. Thayer & Charles Lewis. 1999. An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36(1). 1-28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x.

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

Test level	Rasch statistics summary
B1	SELT B1 All | 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|
B2	SELT B2 All | 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|
C1	SELT C1 | 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|
C2	SELT C1 | 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|

SECTION 2:
ORIGINAL RESEARCH

Chapter 7: 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.

An analysis is presented of a sample of the Listening and Reading test forms which comprise the LANGUAGECERT SELT tests which 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; and as a consequence; 3) 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.

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

Introduction

This report 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) establishing Secure English Language Tests (SELT) tests for movement and 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, mapped to the Common European Framework of Reference (CEFR) for UK Visas & Immigration (UKVI) worldwide, covering all visa type requirements to live, work or study 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 needs to 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, and which were produced over the period 2020-2021. All test forms comprise 52 items.

Against the key test qualities of validity and reliability (Bachman & 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 & 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 matches 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 & 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 in place – to be robust (Coniam et al., 2022).

The LST B1-C1 test forms analysed 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 forms are separate from one another, there are no linking items or test takers by which direct cross-calibrating may be conducted. Nonetheless, 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 previously-established item values by expert judgement. It is therefore this methodology – externally-referenced anchoring – which is used in the current study to explore 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 the 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 & 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; Gable & Wolf,1993; Bachman et al, 1995. Relatively 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 LANGUAGECERT SELT Tests

The LST suite comprises tests at CEFR levels B1 to C2. Examination specifications reflect the requirements of the CEFR; with test materials writers having 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 judgment 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 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, detail on which 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.

Externally-Referenced Anchoring, CEFR levels and Test Forms

The methodology used in the current study is based on, as mentioned, 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.

Table 2: LID scale

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:

3) The extent to which a test's midpoint corresponds to the LID scale level.

4) 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 may be summarised thus:

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

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 externally-referenced anchoring are now examined, in particular means recorded at the 25th and 75th percentiles. 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
(LID scale range: 91-110; midpoint: 100)

	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.

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

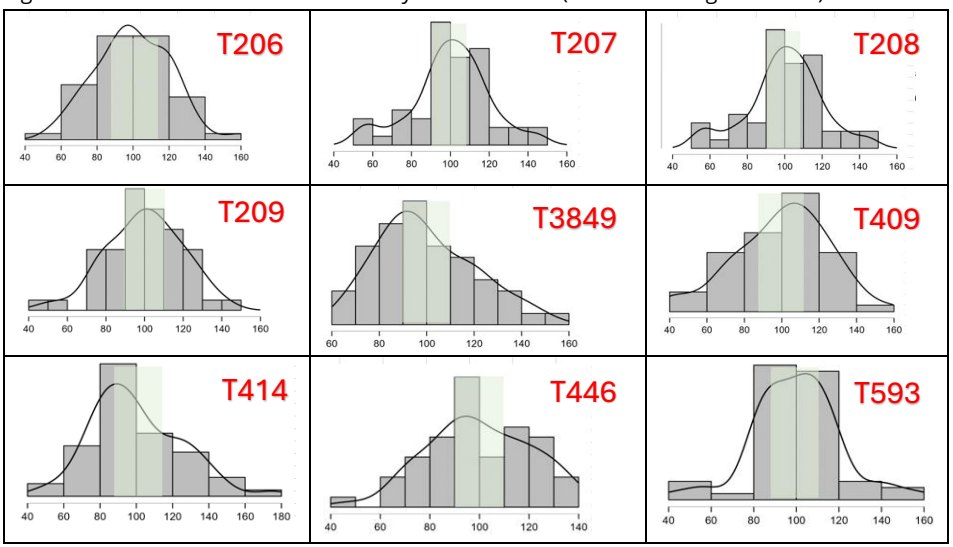
The C1 test forms show a close match with their LID scale ranges. At both 25th and 75th percentiles, all six test forms are close to the upper and lower scale ranges of 150 and 131. This means that all six tests have been well targeted at the C1 level.

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.

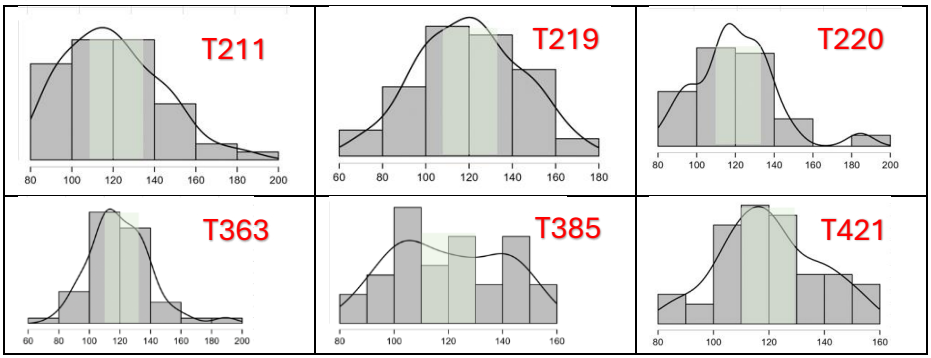
Figure 1: IESOL SELT B1: Item difficulty distributions (LID scale range: 91-110)



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.

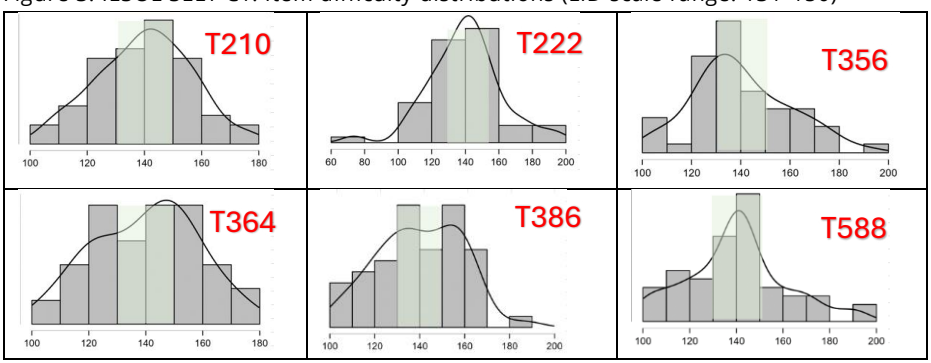
Figure 2: IESOL SELT B2: Item difficulty distributions (LID scale range: 111-130)



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 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. This research will be reported upon in a subsequent paper.

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.

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge. https://doi.org/10.4324/9780429030499.

Coniam, D., Lee, T., Milanovic, M., & Pike, N. (2021a). Validati*ng 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.

Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings* (2nd ed.). New York, NY: Springer Science & Business Media.

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing*, 28(1), 77–104.

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. doi.org/10.46451/ijts.2022.01.12.

Linacre, J. M. (2018). *Winsteps: Rasch measurement computer program user's guide*. 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. doi.org/10.46451/ijts.2022.01.13.

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, Rebecca, Dorothy T. Thayer & Charles Lewis. 1999. An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement,* 36(1). 1-28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x.

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 .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 11.08 ITEM RELIABILITY .99|
T209	SELT B1 T209 | PERSON 10810 INPUT 1411 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 42.3 51.8 146.37 12.11 1.00 .2 .93 .1| | P.SD 9.7 1.5 32.77 7.78 .10 .6 .46 .7| | REAL RMSE 14.40 TRUE SD 29.44 SEPARATION 2.05 PERSON RELIABILITY .81| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 1146.5 1404.9 100.00 1.82 .99 .1 .93 -.4| | P.SD 138.9 6.3 19.40 .43 .13 3.1 .33 3.1| | REAL RMSE 1.87 TRUE SD 19.31 SEPARATION 10.34 ITEM RELIABILITY .99|

T384

```
SELT B1 T384
-------------------------------------------------------------------------------
| PERSON   10810 INPUT    1365 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      41.5      51.8      143.43   11.20       .99     .1    .95     .1|
| P.SD       9.8       1.6       31.91    6.48       .14     .7    .57     .9|
| REAL RMSE  12.94 TRUE SD   29.17  SEPARATION  2.25  PERSON RELIABILITY  .84|
|-----------------------------------------------------------------------------|
| ITEM      52 INPUT      52 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN    1090.0    1358.8      100.00    1.79       .99     .1    .95    -.3|
| P.SD     162.8       6.5       20.37     .33       .13    3.2    .36    3.3|
| REAL RMSE   1.82 TRUE SD   20.28  SEPARATION 11.12  ITEM   RELIABILITY  .99|
-------------------------------------------------------------------------------
```

T409

```
SELT B1 T409
-------------------------------------------------------------------------------
| PERSON   10810 INPUT    1344 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      43.3      51.7      151.86   12.38      1.00     .2    .92     .1|
| P.SD       8.8       2.5       31.74    7.22       .13     .6    .70     .7|
| REAL RMSE  14.33 TRUE SD   28.32  SEPARATION  1.98  PERSON RELIABILITY  .80|
|-----------------------------------------------------------------------------|
| ITEM      52 INPUT      52 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN    1120.1    1335.4      100.00    2.14       .99     .1    .92    -.6|
| P.SD     146.9       6.8       25.02     .87       .14    3.1    .35    3.1|
| REAL RMSE   2.31 TRUE SD   24.91  SEPARATION 10.77  ITEM   RELIABILITY  .99|
-------------------------------------------------------------------------------
```

T414

```
SELT B1 T414
-------------------------------------------------------------------------------
| PERSON   10810 INPUT    1401 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      41.8      51.8      145.66   11.19       .97     .1   1.00     .2|
| P.SD       9.3       1.2       31.86    5.60       .19     .7    .85    1.1|
| REAL RMSE  12.52 TRUE SD   29.30  SEPARATION  2.34  PERSON RELIABILITY  .85|
|-----------------------------------------------------------------------------|
| ITEM      52 INPUT      52 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN    1126.1    1396.3      100.00    1.86       .99    -.1   1.00    -.5|
| P.SD     196.0       5.5       24.41     .47       .14    3.1    .62    3.2|
| REAL RMSE   1.92 TRUE SD   24.33  SEPARATION 12.67  ITEM   RELIABILITY  .99|
-------------------------------------------------------------------------------
```

T446

```
SELT B1 T446
-------------------------------------------------------------------------------
| PERSON   10810 INPUT     655 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      41.0      51.6      141.82   11.18      1.00     .1    .94     .1|
| P.SD       9.8       2.7       32.46    7.13       .12     .7    .49     .8|
| REAL RMSE  13.26 TRUE SD   29.63  SEPARATION  2.23  PERSON RELIABILITY  .83|
|-----------------------------------------------------------------------------|
| ITEM      52 INPUT      52 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN     516.1     650.3      100.00    2.56       .99     .1    .94    -.2|
| P.SD      76.6       4.2       20.68     .62       .13    2.4    .34    2.5|
| REAL RMSE   2.63 TRUE SD   20.51  SEPARATION  7.80  ITEM   RELIABILITY  .98|
-------------------------------------------------------------------------------
```

T593

```
SELT B1 T593
-------------------------------------------------------------------------------
| PERSON   10810 INPUT     641 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN      41.7      51.7      145.88   12.24       .99     .1    .96     .1|
| P.SD      10.0       2.2       34.51    7.93       .14     .7    .63     .8|
| REAL RMSE  14.58 TRUE SD   31.27  SEPARATION  2.14  PERSON RELIABILITY  .82|
|-----------------------------------------------------------------------------|
| ITEM      52 INPUT      52 MEASURED                INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   ZSTD|
| MEAN     514.6     637.2      100.00    2.69       .99     .0    .96    -.2|
| P.SD      71.6       3.6       20.83     .71       .15    2.5    .43    2.6|
| REAL RMSE   2.78 TRUE SD   20.64  SEPARATION  7.43  ITEM   RELIABILITY  .98|
-------------------------------------------------------------------------------
```

Appendix 2: LST B2: Fit Statistics and Person Reliabilities

Test no.	Rasch statistics summary
T211	SELT B2 211 | PERSON 2732 INPUT 528 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ 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 IMNSQ ZSTD OMNSQ 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 | PERSON 2732 INPUT 569 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ 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 8.85 TRUE SD 28.45 SEPARATION 3.21 PERSON RELIABILITY .91| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ 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 | PERSON 2732 INPUT 547 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 33.5 51.8 138.19 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 IMNSQ ZSTD OMNSQ 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 | PERSON 2732 INPUT 573 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ 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 10.85 TRUE SD 28.78 SEPARATION 2.65 PERSON RELIABILITY .88| | ITEM 52 INPUT 52 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ 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 8.09 ITEM RELIABILITY .98|

T385

```
SELT B2 385
-------------------------------------------------------------------------------
| PERSON    2732 INPUT     280 MEASURED              INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   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     IMNSQ   ZSTD  OMNSQ   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
-------------------------------------------------------------------------------
| PERSON    2732 INPUT     235 MEASURED              INFIT        OUTFIT    |
|          TOTAL     COUNT     MEASURE  REALSE     IMNSQ   ZSTD  OMNSQ   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     IMNSQ   ZSTD  OMNSQ   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 | 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 | 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 | 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 5.03 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 | PERSON 581 INPUT 120 MEASURED INFIT OUTFIT | | TOTAL COUNT MEASURE REALSE IMNSQ ZSTD OMNSQ ZSTD| | MEAN 32.2 52.0 155.07 7.54 1.00 .0 1.08 .1| | 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 1.08 .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

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

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 8: Exploring Test Gender Bias in the LANGUAGECERT SELT IESOL Speaking and Listening Tests

Michael Milanovic, Tony Lee, Leda Lampropoulou and David Coniam

Abstract

This chapter examines LANGUAGECERT's two-skills Secure English Language Testing (SELT) International ESOL Speaking and Listening (IESOL) tests. These tests are offered at CEFR levels A1, A2 and B1 and aimed at candidates applying for a visa to migrate or work in the UK, providing evidence of ability to operate in English. The purpose of the current study explored the unbiased nature of the two-skills test, affirming that test results may be seen to be robust and reliable.

An overview is first provided of where two-skills tests are positioned in the broader picture of language skills assessment. An analysis of the A1, A2 and B1 tests is then presented over the period when the tests were administered, i.e., from 2020 to 2023. With the three tests graded in line with CEFR difficulty levels, a study of test bias from the perspective of gender which was explored via differential item functioning (DIF) reported negligible-to-no bias.

Within the constraints of high pass rates, the chapter concludes that the three SELT IESOL Speaking and Listening tests, perform within operational expectations. The SELT IESOL Speaking and Listening tests are robust tests, are functioning as intended and returning reliable results.

Keywords: differential item functioning, Speaking and Listening (IESOL) tests, gender bias,

Introduction

In an era of communicative language teaching and assessment, there is a general recognition that assessment should cover all four language skills (see e.g., Guerrero, 2000; O'Sullivan et al, 2022; Powers, 2010). In the majority of assessment situations, evidence of ability in all four skills is the norm – in school situations and in applying for entrance to university etc – in part to encourage washback and for integrated instruction to be provided in all four skills. The conventional four-skills testing approach, which has been widely used in language assessment for decades, aims to comprehensively evaluate learners' language abilities across all four modalities, providing a comprehensive picture of their overall language proficiency. As language teaching methodologies have evolved and our understanding of language acquisition has deepened, some educators have, however, begun to question the efficacy and practicality of assessing all four skills in a single test.

There is a case for two-skills tests, specifically speaking and listening, where such as authenticity, efficiency, and alignment with communicative language teaching approaches, and the ability to use language for real-life communication is seen as a key competence.

It has come to be accepted that different language learners will exhibit differing levels of ability in the different language skills. Bachman (1985) argued that a divisible model of language ability with a general factor plus distinct traits is a plausible model for how language ability may be compartmentalised. Bachman (1990) extended the earlier research, examining various aspects of language proficiency, including the ability to use language skills separately and in combination.

It has been argued that listening and speaking are theoretically and practically not easily separable (see Douglas, 1997) and that the two skills should be integrated in assessment. Children learn their first language almost exclusively through listening and responding to spoken input, with some estimations that at least half the time spent in communicative interaction involves listening (see Wagner, 2018).

The two-skills speaking and listening test format has the potential to address several key concerns associated with four-skills tests. In contexts where reading and writing skills are not seen as relevant, a focus on the testing of speaking and listening skills, can create more authentic and communicatively meaningful assessment experiences. In this context, testing these skills also aligns with the principles of communicative language teaching.

Frost et al. (2011) state that while language assessment has traditionally focused on measuring the four skills independently, such a focus may be problematic since many 'real world' communicative acts involve the integration of two or more skills, as well as other non-linguistic cognitive abilities.

Gender is considered a key variable in terms of gauging fairness and lack of bias in high stakes tests (see e.g., Ozdemir and Alshamrani, 2020; Song et al., 2015). Against this backdrop, in the current study, gender is explored via DIF in the context of the LANGUAGECERT SELT IESOL Speaking and Listening tests.

Two-skills Tests

A number of two-skills tests have been developed; their main features and focuses are summarised below.

Tavil (2010) reports the successful implementation of an integrated two-skills listening and speaking test which assessed candidates' oral/aural skills through information-gap tasks at a Turkish university.

Frost et al. (2011) investigated how candidates integrate stimulus materials into their speaking performances on an integrated listening-then-speaking summary task. They conclude that the use of an integrated listening and speaking task together with its associated rating scale functions well as a measure of speaking proficiency.

Lion et al. (2013) describe the use of the ALTA Clinician Cultural and Linguistic Assessment, an oral/aural Spanish Speaking and Listening Test administered to physicians in the USA by the ALTA language testing service. The situation required solely an oral/aural test since the study wished explicitly to evaluate American physicians' ability to communicate directly with Spanish-speaking patients.

Cao (2019) outlines the Computerized English Listening and Speaking Test (CELST) which was developed in 2011 and assesses English pronunciation, listening proficiency, interactional competence. The author claims that the CELST meets the requirements of a good oral test, by focusing on information exchange, creating contextualised situations and authenticity to incorporate interaction into language communication.

Rukthong and Brunfaut (2020) investigated listening in the context of integrated tasks such as listening-to-speak. They conclude that the listening/speaking summarisation test task which they developed illustrates that test-takers use a range of cognitive processes strategies in processing listening input.

Four providers offer two-skills Speaking and Listening tests at CEFR levels A1-B1 for visa applicants to meet UK Home Office English language requirements. These are LANGUAGECERT, the IELTS SELT Consortium, Trinity College London and Pearson (see https://www.gov.uk/guidance/prove-your-english-language-abilities-with-a-secure-english-language-test-selt).

An overview of the LANGUAGECERT SELT IESOL Speaking and Listening tests follows.

The LANGUAGECERT IESOL Speaking & Listening Tests

The LANGUAGECERT IESOL Speaking and Listening Test (IESOL S&L) series of graded examinations provide 'steps up the ladder' of proficiency and are suitable for non-native speakers of English who in particular need to demonstrate that they have met the required level of English as specified by the UK Home Office.

The qualifications demonstrate a candidate's ability to communicate using English in real life situations, as may be seen to be appropriate at the respective CEFR levels (A1 to B1 in this case). For details see https://www.LANGUAGECERT.org/en/language-exams/english/LANGUAGECERT-esol-selt.

LANGUAGECERT IESOL Speaking and Listening Test Makeup

The LANGUAGECERT International IESOL Speaking and Listening (IESOL S&L) tests are structured such that candidates respond to speaking and listening tasks which elicit a range of skills. Table 1 elaborates.

Table 1: Speaking and Listening Test tasks

Test Parts	Skill and Focus	Task
Part 1: Respond to questions on familiar matters and communicate personal information	A1 and A2: Give personal information. B1: Express opinions and ideas in addition to the above.	Give and spell name Give country/place of origin Answer three to four questions
Part 2: Initiate and respond appropriately in social situations	A1, A2, B1: communicate in real-life situations using a range of functional language to elicit or respond as appropriate. The sophistication and length of the expected candidate output increases through A1 to B1.	Two situations are presented by the interlocutor at each level and candidates are required to respond to and initiate interactions.
Part 3: Exchange information and opinions	A1 and A2: Exchange information to complete a simple task . B1: Co-operate to reach agreement/decision. The sophistication and length of the expected candidate output increases through A1 to B1.	Exchange information to identify similarities and differences in pictures of familiar situations at A1 and A2 levels. Hold a short discussion to make a plan, arrange or decide on something using visual prompts at B1.
Part 4: (subparts a & b): Understand a short monologue delivered by the marking interlocutor; deliver a short, uninterrupted talk on a relevant topic	A1 and A2: Demonstrate the ability to understand and use sentences and produce a piece of connected spoken English B1: Narrate, describe or communicate ideas and express opinion(s). The sophistication and length of the expected candidate output increases through A1 to B1.	Listen to the monologue and answer the questions. After 30 seconds of preparation time, talk about a topic provided by the interlocutor. Preliminary – half a minute Access – 1 minute B1 – 1 and a half minutes Answer follow-up questions .

The format of the tests and the nature of the assessment criteria reflect the broad multi-faceted construct underlying the Speaking and Listening tests. Communicative ability is the primary focus, while accuracy and range become increasingly important as the CEFR level of the test increases. For example, a level '3' for A1 grammar is defined as:

control of a restricted range of A1 grammar
several errors occur with some A1 grammar

Test Data

The data in the current dataset was compiled from tests administered over the period mid 2020 to early 2023. Table 3 provides details of sample sizes over the period.

Table 3: Sample detail

CEFR level	Candidates
A1	12,868
A2	5,758
B1	22,968

The largest candidature is at B1 level, reflecting the popularity of the respective visa type.

Purpose of the study and its Research Question

As mentioned earlier, the purpose of the study was to investigate whether acceptable quality levels were maintained in terms of the two-skills tests in relation to gender bias or more accurately, lack of it.

Test data and the Global Scale

At LANGUAGECERT, tests, items, and candidate test results are linked to the CEFR via the LANGUAGECERT Global Scale (Milanovic et al., 2023). Global Scale ranges for the three CEFR levels explored in the current study are provided in Table 4.

Table 4: Global Scale (GS) ranges

CEFR level	*GS level cut point*
A1	10
A2	20
B1	40
B2	60
C1	75
C2	90

Examiner, task and candidate facets were explored using Rasch measurement. This involved investigating where the different facets are located on the Global Scale, and where they are located relative to each other. The results of these analyses are not reported here given that the main focus of this chapter is gender bias.

Table 5 first presents details of sample sizes for the different test levels and pass rates.

Table 5: Sample sizes and pass rates

CEFR level	*Candidates*	*Pass rate (%)*	*Mean (max. 30)*	*SD*	*SEM*
A1	12,868	11,043 (85.82%)	23.75	6.33	0.06
A2	5,758	5,136 (89.20%)	24.99	5.95	0.08
B1	22,968	21,976 (95.68%)	27.20	4.45	0.03

KEY: SD=Standard Deviation; SEM=Standard Error of the Mean

As may be seen, pass rates are high for all test levels. The pass mark, as mentioned above, is 18/30. All tests have a mean score considerably above this. Measurement error is nonetheless small. Part of the reason for such high pass rates may be attributed to 'candidate readiness'. With the IESOL S&L tests, the situation is somewhat different from how 'candidate readiness' may be perceived in a school situation. In the latter situation, a student generally takes a test when they are ready for it, often as recommended by their teacher. In contrast, on the IESOL S&L tests, the candidate profile is different by virtue of the fact that the majority of candidates need proof of ability in order to be eligible for the issuing or renewing of a visa. In this context, many candidates sit an IESOL S&L test that is considerably below their actual proficiency level. Many IESOL S&L candidates, for instance, have lived in the UK for many years and are virtual native speakers, i.e., at CEFR C2 level. Such candidates nonetheless need to pass a B1, or even an A1, level test as proof of ability. This is the main reason that such high pass rates emerge.

For many candidates, then, whether they take an A1 or a B1 test makes little difference: many are still going to be C1 or above. The issue is further complicated by the high-stakes nature of the test where a pass is required in order to obtain a visa. School students taking a test which is suggested to be at their level generally accept and live with the results – even a fail grade. In contrast, many IESOL S&L candidates who are marginal and who failed a test the first time around will often retake the test until they achieve a pass. Such a situation exacerbates the high pass rates. In the current study, regarding candidates who have taken a test multiple times, only the candidate's best result has been included in the dataset.

On a methodological point, high pass rates, it should be noted, complicate analyses. Statistical analyses generally need 'space' – i.e., a range of test scores – to be able to conduct sufficient, yet accurate, computations. The lack of such space – as with the current tests with pass rates above 85% – somewhat constrains statistical analysis.

Differential Item Functioning Analysis

This section presents an investigation of differential item functioning (DIF) into the key variable, gender. DIF analysis involves an exploration of whether any subgroup of candidates in a test is being unfairly disadvantaged. In the exploration of potential bias among subgroup types, gender is a key variable that is seen to be worthy of investigation (Ferne & Rupp, 2007).

Rasch-based methods (Roznowski & Reith, 1999) have come to be the preferred statistical mode of analysis for DIF in terms of identifying latent traits. One extension of DIF which has been used in previous studies is Differential Group Functioning (DGF). DGF involves grouping items into sets that share the same latent trait (e.g., Gierl et al., 2001). DGF, which is used in the current analysis, reports biases between candidates' actual responses against the estimated Rasch-calibrated item locations. For ease of reference, however, given the general acceptance of the term "DIF", it is "DIF" that is used in the current study.

In analytic terms, the most demanding category – indicating moderate-to-large DIF strength – is stated as being greater than 0.64 logits (Zwick, 1999). In LANGUAGECERT terms, 0.64 logits equate to approximately 10 Global Scale points. It is this threshold which is taken as the limit for indicating possible bias in the current study.

IESOL S&L candidates are not required to provide demographic detail when registering for the test. Consequently, certain detail is incomplete. Table 6 provides details of test sample sizes and the number of candidates who supplied details of their gender.

Table 6: Sample size and gender detail

CEFR level	*Candidates*	*Stating gender*	*Male*	*Female*
A1	12,868	7,167 (55.70%)	1,751 (24.43%)	5,416 (75.57%)
A2	5,758	1,207 (20.96%)	388 (32.15%)	819 (67.85%)
B1	22,968	6,457 (28.11%)	3,130 (48.47%)	3,327 (51.53%)

Among the three tests, more females than males provided their demographic details, with A1 candidates being the most responsive test group of the three. The available sample size is nonetheless sufficiently large to be able to conduct DIF analyses.

Following the analysis of rating scales above, a DIF analysis was conducted on gender against rating scale. DIF size differences between DIF and actual Global Scale values are provided in Table 7 below.

Table 7: DIF by gender

Gender	Rating scale	**DIF size**
F	TF	0.32
F	ARG	0.86
F	ARV	0.77
F	PIF	0.54
F	LR	-2.07
M	TF	-0.47
M	ARG	0.45
M	ARV	-0.17
M	PIF	0.00
M	LR	0.00

As can be seen from the table above, the largest DIF value was 2.07, considerably below the proposed threshold of 10 scale points. From this, it can be concluded that neither gender can be seen to be unfairly disadvantaged with ratings awarded on the IESOL S&L tests.

Conclusion

This chapter has presented an examination of LANGUAGECERT's two-skills Secure English Language Testing (SELT) International ESOL Speaking and Listening tests, administered in the period 2020 to 2023. The purpose of the study has been to explore the quality of the test and the robustness of results with particular reference to gender bias.

The two-skills tests are offered at CEFR levels A1 to B1, being aimed at candidates who are applying for a visa to migrate, work or study in the UK. The ability focus is on oral/aural skills as evidence of spoken English proficiency.

Pass rates were high, with all tests reporting pass rates of 85% or higher – a reflection of the generally high ability of the candidature and the requirement that candidates possess a pass on a particular test if they are to meet certain UK visa or study requirements. Within these constraints, the three LANGUAGECERT IESOL Speaking and Listening tests have been shown to function reliably, with examiners, tasks and rating being seen to be within operational limits.

In closing, we would therefore state that the SELT IESOL Speaking and Listening tests may be considered robust, that they function as intended, and provide unbiased results.

References

ALTA Language Testing Services. (2023). *Clinician cultural and linguistic assessment (CCLA)*. Retrieved from http://www.altalang.com/language-testing/ccla.aspx.

Bachman, L. F. (1985). An examination of some language proficiency tests from a communicative viewpoint. ERIC.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations*. TOEFL monograph series, 8. Princeton, NJ: ETS.

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345-369.

Guerrero, M. D. (2000). The unified validity of the four skills exam: Applying Messick's framework. *Language Testing, 17*(4), 397-421.

Hidri, S. (2018). Assessing spoken language ability: A many-Facet Rasch analysis. In *Revisiting the assessment of second language abilities: From theory to practice* (pp. 23-48).

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.

Lee, T., Papargyris, Y., Milanovic, M., Pike, N., & Coniam, D. (2023). *Aligning LanguageCert SELT tests to the LanguageCert Item Difficulty (LID) scale*. London, UK: LanguageCert.

Linlin, C. (2020). Comparison of Automatic and Expert Teachers' Rating of Computerized English Listening-Speaking Test. *English Language Teaching, 13*(1), 18-30.

Lion, K. C., Thompson, D. A., Cowden, J. D., Michel, E., Rafton, S. A., Hamdy, R. F., ... & Ebel, B. E. (2013). Clinical Spanish use and language proficiency testing among pediatric residents. *Academic Medicine, 88*(10), 1478-1484.

Milanovic, M., Pike, N., Papargyris, Y., Lee, T., & Coniam, D. (2023). *The LanguageCert Global Scale*. London, UK: LanguageCert.

O'Sullivan, B., Motteram, J., Skipsey, R., & Dunlea, J. (2022). The importance of the four skills in the Japanese context.

Ozdemir, B., & Alshamrani, A. H. (2020). Examining the Fairness of Language Test Across Gender with IRT-based Differential Item and Test Functioning Methods. *International Journal of Learning, Teaching and Educational Research, 19*(6), 27-45.

Powers, D. E. (2010). The case for a comprehensive, four-skills assessment of English-language proficiency. *R & D Connections, 14*, 1-12.

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31-53.

Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. *Papers in Language Testing and Assessment, 4*(1), 97-124.

Tavil, Z. M. (2010). Integrating listening and speaking skills to facilitate English language learners' communicative competence. *Procedia-Social and Behavioral Sciences, 9*, 765-770.

Wagner, E. (2018). Assessing listening. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 29-44). John Benjamins Publishing Company.

Chapter 9: The Use and Impact of Pre-task Planning Time in the Monologic Task of LANGUAGECERT Speaking Tests

Leda Lampropoulou

Abstract

Extensive oral tasks or monologues of different types (e.g., presentations, storytelling) are often used as second language acquisition tasks in the fields of language learning and language testing. Pre-task planning time is a common provision to test-takers who may use different strategies to prepare their response. High-stakes tests, such as the LANGUAGECERT IESOL suite of tests, include planning time prior to monologic tasks and offer test-takers the opportunity for note-making. While the language assessment literature supports planning time for reasons of face validity and fairness, research studies do not consistently support correlations between planning and performance.

The current study examined the differences between the scores of test-takers who used note-making as a strategy and those who did not. The research questions investigated: (i) whether test-takers who make notes during planning time in the monologue task of an L2-English B2 speaking task are awarded higher scores on their spoken performances than test-takers who do not and (ii) test-takers' perceptions of their use of planning time.

The findings suggest that making notes did not improve test-takers' performance against any of the rating criteria used in the assessment. It also revealed that most test-takers use their planning time to generate their main propositions.

Keywords: speaking exams, oral tests, planning time, pre-task planning, LANGUAGECERT exams

Background and Study Scope

The study has two related focuses. Firstly, it examines whether using note-making during pre-task planning time creates gains for test-takers which are transferable into scoring. Secondly, it aims to provide insights into test-taker perceptions of their use and focus during planning time. In this regard, the background and research literature, and in particular, oral assessment task types and task planning, are first explored; following this, the study is outlined and pedagogical implications are discussed.

Second language (L2) assessments use various task types, frequently in combination, to operationalise the construct of speaking to measure, as accurately as possible, test-takers' L2 spoken ability. Brown (2004) placed these on a continuum of a five-level taxonomy. According to that, the simplest tasks in terms of cognition and ease of completion involve mere repetition of oral input and are classified as imitative. In between task types range from intensive and responsive to interactive, with the length of the expected response, turns, and exchanges adding to the tasks' complexity. Extensive language production tasks such as story narration and both formal and informal presentations were classified as the most complex and demanding. Brown's (2004) taxonomy included planning in the design of the extensive production tasks, a practice also reflected in L2 speaking assessments, such as the speaking tests in the LANGUAGECERT International English for Speakers of Other Languages (IESOL) suite of exams.

Ellis (2009) and Foster and Skehan (1996) explored the influence of task planning on L2 oral production, focusing on syntactical *complexity*, *grammatical accuracy,* and *fluency* (CAF). Linguists have also researched the connection between planning and the length of planning time available (e.g., Li et al. 2014). Pang and Skehan (2014) examined task planning in a task-based language teaching (TBLT) setting, where pre-task planning is contrasted with task repetition and rehearsal.

In a meta-analysis of task planning and oral L2 production, Johnson and Abdi Tabari (2022) observed that two theoretical models have been used predominantly – Shekan's (1998) trade-off hypothesis and Robinson's (2011) cognition hypothesis. Briefly, the trade-off hypothesis suggests that attentional capacity to the three CAF elements is limited and that attending to increase performance in one area may take attention away from the other two and result in a weaker performance in those areas. The cognition hypothesis suggests that task complexity will raise both language complexity and accuracy, at the expense of fluency. Johnson and Abdi Tabari (2022) stated, nonetheless, that no research findings have consistently and unambiguously confirmed a positive relationship between planning and oral L2 production. Considering the absence of a firm conclusion of a seemingly fair assessment method and intuitively sensible good practice, the current study set out to examine the relationship between pre-task planning time and L2 oral production in the IESOL Speaking Test. Specifically, the study investigates the impact of making notes during pre-task planning time on the test-takers' performance in an extensive speaking task – a monologue – as part of a LANGUAGECERT IESOL speaking test at B2 level. In particular, it looks at the effect that can be observed on test-taker scores – depending on whether test-takers use a note-making strategy to prepare their monologue on the assigned topic. Conducted in a formal assessment setting, scores awarded reflect test-takers' performance under the criteria of task fulfilment (TF), grammatical range and accuracy (GRA), lexical range and accuracy (VOC), and pronunciation intonation and fluency (PIF). The study also then briefly explores the key areas on which test-takers chose to direct their focus during that designated time.

Planning is "a problem-solving activity" (Ellis, 2005) through which learners select and determine a strategy to express a speech act, either in an automatic or a controlled manner depending on the learner's proficiency level, with a higher automatisation ability available at higher L2 proficiency levels (De Bot, 1992). De Bot (1992) extended Levelt's (1989) system of first language (L1) production to also apply it to L2 speech. In both systems, oral production comprises three gradual psycholinguistic processes: conceptualisation; formulation; articulation. Investigations into planning have mainly focused on its impact on the speech produced. Ellis (2005) classified planning into pre-task and within-task planning, with four issues generally identified:

- planning time length and availability
- learners' language proficiency level
- task type and task complexity
- the lack or presence of structured guidance

In L2 teaching, task-based language teaching (TBLT) has been used to operationalise strategic planning and to explore its impact on the complexity, accuracy, and fluency (CAF) of learners' performance. Crookes (1989) and Yuan and Ellis, (2003) found significant improvements in all three CAF measures when pre-task planning time was allowed. Most studies have, however, only been able to report similar findings for aspects of fluency and complexity (e.g., Foster & Skehan 1996; Ortega, 1999). Geng and Ferguson (2013) found that planning time before a complex task positively affected the syntactic complexity of test-taker responses although at the expense of fluency.

An overview of research on planning time and its impact on the CAF of the spoken language produced has revealed a relative consensus on the value of providing L2 learners with an opportunity to plan their response in a teaching and learning context even if the benefits are more apparent for the complexity and fluency indicators and not always observable in the errors that earners make.

Table 1 presents an overview of research on planning time and its impact on the complexity, accuracy, and fluency of the speech produced. Out of the 19 studies reviewed, all but five observed a positive effect on complexity. Similarly, in fourteen of the studies a positive effect was observed on fluency, while only one saw a negative effect. The most inconsistent findings were observed for accuracy. Six studies found a positive effect, while another six concluded in mixed results.

Table 1: Overview of research on pre-task planning time in a TBLT setting and its effect on performance

Researcher(s)	Planning time	Positive effect on		
		Complexity	Accuracy	Fluency
Crookes (1989)	10 min.	✓	✓	✓
Foster & Skehan (1996)	10 min.	✓	mixed	✓
Skehan & Foster (1997)	10 min.	mixed, with a trade-off effect		
Menhart (1998)	1 min., 5 min., 10 min.	✓	✓	✓
Ortega (1999)	10 min.	✓	mixed	✓
Foster & Skehan (1999)	10 min.	✓	✓	✓
Yuan & Ellis (2003)	10 min. and within task	✓	mixed	×
Kawauchi (2005)	10 min.	✓		✓
Sangarun (2005)	15 min.	✓		✓
Skehan & Foster (2005)	10 min.	✓		✓
Gilabert (2007)	10 min.	✓	✓	✓
Mochizuki & Ortega (2008)	5 min.	✓	✓	✓
Guara-Tavares (2009)	10 min.	✓	✓	✓
Ahangari & Abdi (2011)	10 min.	✓	✓	
Sasayama & Izumi (2012)	5 min.	✓		✓*
Genc (2012)	10 min.		✓	
Geng & Ferguson (2013)	10 min.	✓	✓	✓
Nielson (2013)	10 min.	✓	mixed	✓
Khorami & Khorasani (2018)	10 min.		mixed	

Note: A significant positive effect in the specific area is symbolised with a tick (✓) in the respective column, whereas the absence of a statistically significant effect is symbolised with a cross (×). Inconsistent findings are described as mixed. Blank cells are in place when researchers did not research that area or did not report a finding for that skill. * This research found a negative effect on fluency.

A literature review of the research conducted to investigate the presence of a similar positive impact on the scores achieved by planners in an assessment setting, nonetheless, suggests a somewhat dissimilar picture. Overall, the findings of these studies do not consistently align with the conclusions reached by researchers investigating planning time within a learning context. Even within the assessment context, findings vary. An overview of the studies on the use of pre-task planning time in oral exams is presented in Table 2. Out of the thirteen studies reviewed, a positive effect on scoring was observed on only five of them, but that was significant only in three of them.

Table 2: Overview of research on pre-task planning time in a testing setting and its effect on scores

Researcher(s)	Planning time	Effect on scoring
Wigglesworth (1997)	1 min.	no
Iwashita et al. (2001)	3 min.	no
Elder et al. (2002)	3 min.	no
Tavakoli & Skehan (2005)	5 min.	no
Xi (2010)	1 min.	yes, minimal
Elder & Iwashita (2005)	3 min.	no
Weir et al. (2006)	1 min.	yes
Elder & Wigglesworth (2006)	1 min., 2 min.	no
Nitta & Nakatsuhara (2014)	3 min.	yes
Li et al. (2014)	30 sec., 1 min.	no
Li et al. (2014)	2 min., 3 min., 5 min.	yes
O'Grady (2019)	30 sec., 1 min., 5 min., 10 min.	yes, minimal
Innue & Lam (2021)	1 min. and 30 sec.	no

The stark contrast between the two contexts, i.e. learning and assessment, is not as surprising as it may initially seem, as there are certain foundational differences between the two settings.

The amount of time provided to examination test-takers to plan their responses was, with very few exceptions, considerably shorter than what was offered to classroom learners. In TBLT, ten minutes of planning time has been used in most studies. In contrast, assessments that developed research-informed specifications and included planning time needed to also adhere to the principle of practicality (Bachman & Palmer 1996) and limited the time offered to test-takers to one minute only in most of the test-based studies. Consequently, the time allowance for planning in a speaking test may be insufficient for an improvement in scoring to be observed.

Several test-based studies have employed CAF to evaluate test-taker performances (Li et al., 2014; Nitta & Nakatsuhara, 2014). The most common practice remains, nonetheless, that of trained markers assessing test-takers' responses using the respective test's rating scales or an adaptation of these. The resulting scores are then used for the research project's calculations. The absence of similar positive effects on the test-takers' spoken performance may therefore be attributed to the nature of the scale not being sufficiently sensitive for the improvements to be measured or that the CAF gains were too minor to be apparent (Wigglesworth, 1997; O'Grady, 2019).

Despite the relatively conflicted findings regarding the significance of the impact, a conclusion unanimously reached by test-based researchers was that pre-task planning time is justified and should be granted to test-takers. Swain's (1985) principle that tests ought to be biased towards eliciting a test-taker's best possible performance is widely regarded as good practice (Elder & Wigglesworth, 2006; O'Grady, 2019). Additional reasons in favour of providing test-takers with planning time in speaking tests include arguments on construct validity, authenticity, and fairness (O'Grady, 2019; Wigglesworth, 1997) [Note 1].

What appears to have been less researched in this area is the test-takers' their perspectives and insights into what occurs during planning time. Since inconclusive findings reported may suggest that test-takers use ineffective strategies to plan their responses, further investigation is needed into the impact of planning time as well as into test-takers' perceptions, to better comprehend the value of making notes as a planning strategy.

Research Questions

The research questions addressed in the study were:

RQ 1: To what extent are test-takers who make notes during planning time in the monologue task of an L2-English B2 Speaking Test awarded higher scores on the different rating scales on their spoken performances than test-takers who do not make notes?

RQ 2: What are test-takers' perceptions of the strategic planning time offered prior to the English B2 Speaking Test monologue task?

Methodology, Participants, Data

Data collection was completed in two phases. Test-takers' scores in the live exams provided the data for the quantitative analysis, whereas their responses to the questionnaire formed the data to explore their perspectives.

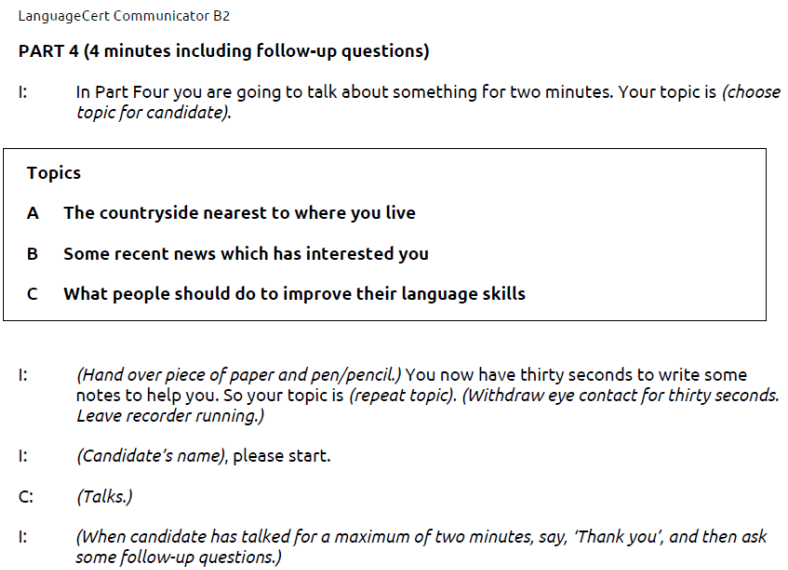
Participants

Participants who agreed to participate in the study were test-takers registered to take the LANGUAGECERT IESOL Speaking test at B2 level of the Common European Framework of Reference for Languages (CEFR). Of the 50 participants who consented to take part in the study, 31 were students at an English language school (a LANGUAGECERT test centre) while the remaining 19 had registered for the exams that LANGUAGECERT conducts in Greece in its premises in Athens. There were 31 female and 19 male participants. The majority of the participants were from Greece (n=45), while four were Albanian and one was Indian. On average, the participants had been studying English as a foreign language for six years.

Live Speaking Tests

Each test-taker in the study sat a face-to-face IESOL B2 Speaking exam. The exams were conducted as per the normal process i.e., a live interview with one interlocutor and one test-taker per session. A recording device was used to record audio only, based on which the test-taker's performance is assessed, at a later stage. A LANGUAGECERT IESOL B2 Speaking test exam paper was used. The study focused on the final task: the monologue. The framework and a sample task are shown in Figure 1 below, with (I) referring to the interlocutor and (C) referring to the test-taker.

Figure 1: Speaking Test Part 4 Rubric - Interlocutor Framework



LanguageCert Communicator B2

PART 4 (4 minutes including follow-up questions)

I: In Part Four you are going to talk about something for two minutes. Your topic is *(choose topic for candidate).*

Topics

A The countryside nearest to where you live

B Some recent news which has interested you

C What people should do to improve their language skills

I: *(Hand over piece of paper and pen/pencil.)* You now have thirty seconds to write some notes to help you. So your topic is *(repeat topic). (Withdraw eye contact for thirty seconds. Leave recorder running.)*

I: *(Candidate's name)*, please start.

C: *(Talks.)*

I: *(When candidate has talked for a maximum of two minutes, say, 'Thank you', and then ask some follow-up questions.)*

The format for the B2 Speaking test involves the interlocutor choosing a topic from a selection of three equivalent topics. For equivalence and reliability purposes, interlocutors were instructed to use the same topic for all test-takers and to ignore the other two options. The selected topic was: "A time when your family helped you". Interlocutors announced the topic to the test-takers orally only and informed them that they were given 30 seconds as planning time to "write some notes to help [them]". Test-takers were given a pencil and a piece of paper, and the interlocutors then repeated the task topic. Planning time began at that point and lasted 30 seconds. If the test-taker insisted on starting their monologue early, however, this was permitted.

Test-takers kept their notes and could consult them during their talk. The interlocutor retrieved the notes at the end of the test, with test-takers aware that these would not be assessed.

Of the 55 sheets that were returned, five were unnamed and excluded from the study. For the remaining, 44 had at least one word noted while the remaining six were blank.

Post-test Questionnaire

A questionnaire was used for two reasons. Firstly, previous researchers had identified the need for further insight into the test-takers' perceptions of pre-task planning. Secondly, since the creation of very brief notes does not allow significant mapping via analytic discourse analysis, a post-test questionnaire was administered that enquired into the use of planning time for the monologue part of the Speaking Test.

The questionnaire items comprised short questions, no negative constructions, and were written in language below the CEFR B2 level at which test-takers were being assessed.

A first draft was piloted on two volunteers from within LANGUAGECERT's research team, following which, a second draft was then piloted with ten mock test-takers. The final version of the questionnaire was then developed. It included only four items and did not enquire into test-taker demographics. The first item asked test-takers whether they had made notes. Based on their initial response, they were asked what they focussed on during planning time, regardless of note-making. The last item asked test-takers whether the planning time allowed had been sufficient.

Out of the fifty test-takers who had consented to participate in the study only one did not wish to complete the questionnaire. The remaining forty-nine agreed and answered all items, with questionnaire completion taking place as close as possible to the Speaking Test itself, while test-takers exited the examination room. The questionnaire items can be found in Appendix 1.

Scoring the Speaking Performances

Five trained interlocutors followed a scripted framework, interacting with the test-taker throughout the test. To avoid contamination, the interlocutors were not briefed on the details of the study. Interlocutors in face-to-face exams do not assess the test-takers but record the audio of the session to be assessed by a different marker.

An experienced marker rated all the Speaking test performances using the standard IESOL markscheme. For the study, the marker was asked to first listen to and rate the monologue task individually, and to rate the test-taker's performance on the rest of the test (i.e., Parts 1-3) at a later stage. Marking was done over five days to minimise any halo effect and to avoid marker fatigue. A recording or parts of each test were available to ensure confident rating. The marker was not given access to the test-takers' notes. The test-takers' performances were then all second marked and inter-rater reliability was good (α=0.81). For a better understanding of the marks available to be awarded, the rating scale at Communicator – CEFR B2 level with the analytic markscheme and descriptors per mark can be found in Appendix 2.

Quantitative Data Analysis

For each test-taker, five different raw marks were generated as per table 3 below.

Table 3: Marks available

Rating scale	Marks available
Task Fulfilment	0 – 3
Grammatical Resources	0 – 3
Lexical Resources	0 – 3
Pronunciation, Intonation, Fluency	0 – 3
Total	0 - 12

Test-takers were divided into two groups. Group A (n=44) comprised test-takers who used their planning time to make notes writing down anything apart from the topic. Group B (n=6) consisted of those who only noted down the title of the topic or produced no notes. Samples of notes produced by the test-takers can be found in Appendix 3.

Since the marker used a rating scale to award specific marks, there was already some indication that the data, being ordinal, should be analysed as not normally distributed. To confirm this, a descriptive statistics analysis of the test-takers' scores per criterion was performed using the program IBM SPSS Statistics (Version 27) to run a test of normality. A Shaphiro-Wilk test is suggested for a sample size of up to 50. The null hypothesis is that data are normally distributed. The results of the tests are shown in Table 4. For all four criteria, significance P <.001, which means the data should be handled as non-parametric.

Table 4: Shaphiro-Wilk tests for data normality

Rating scale	Statistic
Task Fulfilment	(W = .747, p <.001)
Grammatical Resources	(W = .657, p <.001)
Lexical Resources	(W = .704, p <.001)
Pronunciation, Intonation, Fluency	(W = .519, p <.001)

As can be seen, on all rating scales, the results of Shaphiro-Wilk tests showed that the data were not normally distributed, so non-parametric tests were adopted to analyse the data of this study.

The effect of note-making during planning time on the test-takers' performance was then investigated by comparing the scores of Group A (note makers) and Group B (non-note makers) in the monologue. The nonparametric Mann-Whitney U test was used to compare the performance of the two groups.

Results

Monologue Rating Scale Scores

The main research question examined whether test-takers who make notes during planning time in the monologue task are awarded higher scores on the four criteria than test-takers who do not make notes. Mann-Whitney U test results are reported in turn below for the rating scales. Table 5 reports the results for Task Fulfilment.

Table 5: Mann-Whitney U test on TF scores

Mann-Whitney U	U = 125, p = .803
Group A	Md = 2, n = 44, mean rank = 25.34
Group B	Md = 2, n = 6 mean rank = 26.67

The Mann-Whitney U Test results did not reach significance (p = .85), indicating that note makers do not score higher than non-note makers for topic development.

Table 6 reports the results for Grammatical Range and Accuracy.

Table 6: Mann-Whitney U test on GRA scores

Mann-Whitney U	U = 81, p = .067
Group A	Md = 1 n = 44, mean rank = 26.66
Group B	Md = 1 n = 6, mean rank = 17.00

The Mann-Whitney U Test revealed no significant difference in the GRA score of the two groups. The mean rank analysis of the GRA scores shows a much larger difference than the one reported for the TF criterion. However, the difference in test-takers' scores failed to reach statistical significance, suggesting that note makers do not use grammatical structures which are awarded higher scores than non-note makers.

Table 7 reports the results for vocabulary range and accuracy between note-makers and non-note-makers.

Table 7: Mann-Whitney U test on VRA scores

Mann-Whitney U	U = 114, p = .538
Group A	Md = 1 n = 44, mean rank = 25.91
Group B	Md = 1 n = 6, mean rank = 22.50

The Mann-Whitney U Test results did not reach significance indicating no substantial difference in the VRA scores of test-takers who made notes versus those who did not.

Table 8 presents the results of the Mann-Whitney U test for pronunciation intonation and fluency.

Table 8: Mann-Whitney U test on PIF scores

Mann-Whitney U	U = 111, p = .310
Group A	Md = 2, n = 44, mean rank = 25.03
Group B	Md = 2, n = 6, mean rank = 28.92

The Mann-Whitney U Test revealed no significant difference in the PIF score of the two groups – mirroring the null findings in the other three criteria and suggesting that there is no significant difference in the performances between note-making test-takers and those who did not make notes in pronunciation intonation and fluency aspects.

In summary, none of the tests conducted to examine the impact of making notes during planning time on the scores awarded for the four criteria of the rating scale produced statistically significant results, indicating that note-making test-takers were not awarded significantly different scores from non-note-makers.

Post-test Questionnaire

The second research question explored test-takers' perceptions of the strategic planning time offered prior to the Speaking Test monologue task. Some items on the questionnaire (see Appendix 1) allowed multiple responses: for example, what test-takers spent their planning time on, regardless of whether they had made notes.

Table 9 provides detail on questionnaire item 1, whether test-takers made notes in the planning time prior to the monologue.

Table 9: Questionnaire Item 1 - Making notes during planning time

Did you make notes during planning time?	Responses
Yes, I made a lot of notes.	7 (14%)
Yes, but just some words.	36 (74%)
No, I didn't make any notes.	6 (12%)

Most of the test-takers (74%) stated they had made notes, but just a few words. The ratio of note sheets only containing just a few words is also an accurate representation of the collected sheets. There were a few test-takers (12%) who stated they had made no notes at all, and this number coincides with the number of sheets that were returned blank.

For the next set of questions, Question 2 (Q2) and Question 3 (Q3), test-takers were asked to specify how they had chosen to spend the thirty seconds of time they had at their disposal, or what their notes' purpose was, if they had made any.

Table 10 presents the available options and the responses for each one.

Table 10: Questionnaire Item 2 and Item 3 responses (combined groups) – Planning time usage

During planning time, I focussed on...	Responses
Generating ideas	38 (78%)
Structuring my monologue	10 (20%)
Planning my grammatical structures	4 (8%)
Selecting useful vocabulary items	8 (16%)
Calming down	5 (10%)
Nothing in particular	2 (4%)

As can be seen from table 10, the most prevalent response was noting down or thinking of ideas to talk about, with 78% of the respondents reporting this as their main focus. 20% reported planning the monologue's structure. 16% reported focusing on vocabulary. Other options accounted for 10% of respondents or less.

To distinguish between planning strategies test-takers used and examine the relationship within note makers and non-note makers a crosstabulation of the focus areas within the two groups was undertaken. Table 11 shows the results of the crosstabulation of the questionnaire responses.

Table 11: Planning time usage by test-taker group

Planning Time Usage	Yes Notes	No Notes	Total
Ideas	34 (79.1%)	4 (66.7%)	38
Structure	10 (23.3%)	0 (0.0%)	10
Grammar	4 (9.3%)	0 (0.0%)	4
Vocabulary	8 (18.6%)	0 (0.0%)	8
Calm down	2 (4.7%)	3 (50.0%)	5
Nothing	0 (0.0%)	2 (33.3%)	2
Total	43	6	49

As can be seen from Table 11, 79% of note-makers focussed on thinking of and writing down the ideas to talk about.

The final question on the questionnaire enquired into the adequacy of the provided planning time. Table 12 depicts test-takers' responses to that regardless of whether they had made notes during that time or not.

Table 12: Questionnaire Q4 responses: 30" to plan your monologue. Was this time enough?

Respondents	Planning time was enough	Responses	
Note makers		Yes	27 (63 %)
	No	16 (37 %)	
Non note makers	Yes	4 (67 %)	
	No	2 (33 %)	

The majority of test-takers responded that the time they were provided with sufficed to plan their response. Respondents from both groups agreed by approximately the same percentages – 63% and 67% respectively – that they did not need more time to plan better. A follow-up question was asked about what they would use the extra time for. Most test-takers responded that they would have used it to think of more ideas to talk about while a few others mentioned they would have used it to relax. The frequencies analysis for all items is presented in Table 13.

Table 13: Questionnaire frequency statistics (n=49)

	Made Notes	Ideas	Structure	Grammar	Vocabulary	Calm down	Nothing	Enough time
Mean	.88	.78	.20	.08	.16	.10	.04	.61
SD	.33	.42	.41	.28	.37	.31	.2	.49

Discussion

This study investigated the potential effects of planning one's monologue through making notes on test-taker performance in a B2 speaking test setting using LANGUAGECERT IESOL speaking test-takers. Assessed performances of note makers were compared on four rating scales with the performances of test-takers who did not use a note-making strategy.

The research found that note-makers were not awarded statistically significant higher marks on any of the criteria. This suggests that notes did not help test-takers fulfil the given task more fully or more coherently, nor did they demonstrate a consistently higher level of GRA, a superior VOC, or a more natural and effective PIF. In other words, the analyses outcomes seem to suggest that there is no difference on the test-taker's performance, regardless of whether they prepare a response through making notes or not.

These results may appear counter-intuitive and in conflict with the TBLT literature which suggests that pre-task planning time offered in a classroom setting can substantially improve test-takers' task performance when speech production is measured against CAF indices (Foster & Skehan 1996; Geng & Ferguson 2013). Fluency and complexity, the two areas where gains from planning are most frequently observed in TBLT studies, did not appear to behave any differently from the rest of the criteria examined in the current study. A relatively uncomplicated explanation is that TBLT studies in most cases offered learners much more planning time (ten minutes in TBLT compared to one to two minutes for most test tasks). However, studies that allowed test-takers five to ten minutes to prepare results were still not able to confirm a meaningful effect on test-takers' performances (Tavakoli & Skehan, 2005).

Nevertheless, and perhaps more importantly, the current study's null results are consistent with the body of research conducted under exam conditions. These suggest that strategic planning did not have a meaningful or substantial effect on test-takers' spoken production (Innue & Lam 2021; Wigglesworth & Elder 2010). Thus, the conclusion is that using note-making as a strategy for optimal performance in the monologue section of the LANGUAGECERT IESOL speaking test does not produce an observable improvement on spoken performance and may be of limited effectiveness, as currently used by test-takers.

By way of reinforcing the fairness argument and adding to it an element of face validity, responses to the appropriacy of the offered length of planning time revealed that test-takers were relatively split between those who were happy with the time provided (approximately 60%) and those who would have wished for more (approximately 40%). Consequently, although extending planning time may be against test practicality and unsupported by the study's findings, reducing or removing it altogether might jeopardise score acceptance by test-takers and stakeholders (Innue & Lam, 2021; Wigglesworth & Elder, 2010).

Limitations and Implications

In the process of acknowledging the limitations of this research some additional points should be considered. Firstly, the rating scale's (in)capacity to measure the kind of gains produced by planning and note-making could perhaps account for the absence of a demonstrable effect on scores. The issue of the appropriacy and fitness of the rating scale was encountered with different types of scales and measures. O'Grady (2019) used an empirically derived binary-choice, boundary-definition (EBB) scale, and an analytic rating scale but did not find substantial differences between the generated test scores. It may therefore be useful to explore whether a different, perhaps longer than four bands, scale would guide markers to different assessment decisions that may allow gains from strategical planning to manifest in test-takers' scores.

Secondly, the quality of test-takers' notes during planning should be considered because only a very limited amount of the notes would qualify as good plans for an effective oral response. Very few test-takers used arrows or a bullet list and none sketched notes in the form of a mind map. These poorly written plans demonstrate the test-takers' overall lack of good note-making skills which otherwise could be conducive to an improved response to the oral task. The possibility for test takers' note-making skills to be ineffective and unable to assist them in producing a better oral response than they would produce without any notes entails certain pedagogical implications. Teachers preparing students for LANGUAGECERT oral exams - but also for any speaking examination that contains similar tasks where the opportunity for making notes is offered to the test takers - might want to teach planning skills in a structured and explicit manner, to help their learners develop and sharpen them. Despite the fact that note making is a life skill which will also be useful to the learner beyond the test, it is generally undervalued. The interlocutors conducting the speaking exams are a case in point. In discussing their views on the task after their had conducted the exams, they reflected that, as language teachers, they had never explicitly taught their students how to make notes in preparation for that speaking task. This cannot be generalised, and a study on how teachers teach planning strategies could shed light on the matter. In the meantime, teachers should perhaps consider honing such skills through practicing different planning strategies, in an attempt to explore what works for each learner, and that will be a welcome positive washback of the speaking assessments.

A significant limitation in this study was that the participants were not equally split into note makers and non-note makers as the majority of the test-takers opted for some sort of notes. In this study the aim was to investigate the note-making strategy keeping the live exam conditions untouched. A future study could also examine how test-takers who normally make notes would perform if deprived of the note-making option.

Conclusion

The present study sought to examine the use of note-making as a pre-task planning strategy and its effect on L2 oral performance at a CEFR B2 level (Council of Europe 2001) speaking test task aiming at eliciting a monologue. The view that there is a positive correlation between planning and performance has been widely endorsed in language teaching and supported in the relevant research (Nielson, 2013; Yuan and Ellis, 2003). Nevertheless, a similar effect is not consistently present in a testing setting (Inoue & Lam, 2021; Nitta & Nakatsuhara, 2014; Wigglesworth & Elder, 2010).

The current study, conducted in a face-to-face exam context, reflects the findings of other relevant studies that using note-making as a pre-task planning strategy does not have a significant effect on test-takers' performance in terms of their scores. This finding was consistent across all four criteria comprising the test's markscheme. In this sense, this study can be used to complement the body of research on pre-task planning time usage by adding to the range of task types (a 2-minute monologue), the specific strategy (note-making), and the specific planning time provided (30 seconds). It also contributes to the less researched area of test-takers' perceptions of their own use of planning time and note-making.

These findings, however, should not be taken to imply that pre-task planning time is redundant in an oral test or to suggest that assessment developers should eliminate them from test task specifications. The current study argues that planning time of as little as thirty seconds should be included in the design of all extensive monologic tasks; the provision of a longer planning period should be considered for more demanding tasks which may require more than idea generation.

Apart from researchers and test developers interested in designing research-informed test specifications, SLA practitioners such as teachers of English as a foreign language (EFL) and materials developers may also benefit from being aware of the results and the pedagogical implications of this type of studies. The poor quality of the test takers' note making practices as observed in their returned indicates there may be a learning gap in planning strategies and note-making skills.

Note

1. Construct validity refers to the capacity of the generated test scores to be generalised and interpreted meaningfully and legitimately into the intended real-world use or target language use (TLU) domain (Bachman & Palmer, 1996).

References

Bachman, L.F., & Palmer, A.S. (1996). *Language Testing in Practice*. Oxford University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Longman.

Bui, G., & Huang, Z. (2018). L2 fluency as influenced by content familiarity and planning: Performance, measurement, and pedagogy. *Language Teaching Research, 22*(1), 94–114. https://doi.org/10.1177/1362168816656650

De Bot, K. (1992). A bilingual production model: Levelt's "speaking model" adapted. *Applied Linguistics*, *13*, 1–24.

Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, *11*(4), 367–383. https://doi.org/10.1017/S0272263100008391

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing, 19*(4), 347–368. https://doi.org/10.1191/0265532202lt235oa

Elder, C., & Iwashita, N. (2005). Planning for test performance: Does it make a difference? In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 219–239). John Benjamins.

Elder, C., & Wigglesworth, G. (2006). An investigation of the effectiveness and validity of planning time in part 2 of the IELTS speaking test. *IELTS Research Reports* (*Vol. 6*, pp. 1–28). IELTS Australia and British Council.

Ellis, R. (2005). Planning and task-based performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 3–34). John Benjamins.

Foster, P., & Skehan, P. (1996). The influence of planning time on performance in task-based learning. *Studies in Second Language Acquisition, 18*, 299–234. https://doi.org/10.1017/S0272263100015047

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and (+/− here and now): Effects on oral production. In M. Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 44–68). Multilingual Matters.

Genc, Z. (2012). Effects of strategic planning on the accuracy of oral and written tasks in the performance of Turkish EFL learners. In A. Shehadeh & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts research and implementation* (pp. 67–89). John Benjamins.

Geng, X., & Ferguson, G. (2013). Strategic planning in task-based language teaching: The effects of participatory structure and task type. *System*, *41*, 982–993.

Guara-Tavares, M. (2009). The relationship among pre-task planning, working memory capacity, and L2 speech performance: A pilot study. *Linguagem & Ensino*, *12*(1), 165–194. https://doi.org/10.1590/S0103-18132013000100002

Inoue, C., & Lam, D.M.K. (2021). The Effects of Extended Planning Time on Test takers' Performance, Processes, and Strategy Use in the Lecture Listening-Into-Speaking Tasks of the *TOEFL iBT*® Test. ETS Research Report Series, 2021: 1-32. https://doi.org/10.1002/ets2.12322

Iwashita, N., McNamara, T. & Elder, C. 2001: Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning* 51, 401–36. https://doi.org/10.1111/0023-8333.00160

Johnson, M. D., & Abdi Tabari M. (2022). Task Planning and Oral L2 Production: A Research Synthesis and Meta-analysis, *Applied Linguistics*, amac026. https://doi.org/10.1093/applin/amac026

Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 143–165). John Benjamins.

Khorami, A., & Khorasani, R. (2018). The effects of planning time and proficiency level on accuracy of oral task performance. *Global Journal of Foreign Language Teaching*, *7*(4), 155–168. https://doi.org/10.18844/gjflt.v7i4.3004

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge University Press.

Li, L., Chen, J., & Sun, L. (2014). The effects of different lengths of pre-task planning time on L2 learners' oral test performance. *TESOL Quarterly*, *49*(1), 38–66. https://doi.org/10.1002/tesq.159

Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativization. Language Teaching Research, 12(1), 11–37. https://doi.org/10.1177/1362168807084492

Nielson, K. (2013). Can planning time compensate for individual differences in working memory capacity? *Language Teaching Research*, *18*(3), 272–293. https://doi.org/10.1177/1362168813510377

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on oral task performance. *Language Testing*, *31*(2), 147–175. https://doi.org/10.1177/0265532213514401

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. Language Testing, 36(4), 505–526. https://doi.org/10.1177/0265532219826604

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*(1), 109 - 148. https://doi.org/10.1017/S0272263199001047

Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and second language reporting in narrative retelling. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp.
95–128). John Benjamins.

Robinson, P. (ed.). (2011). *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*. John Benjamins. https://doi.org/10.1075/tblt.2

Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 111–143). John Benjamins.

Sasayama, S., & Izumi, S. (2012). Effects of task complexity and pre-task planning on Japanese EFL learners' oral production. In A. Shehadeh & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts research and implementation* (pp. 23–43). John Benjamins.

Skehan, P., & Foster, P. (1997). Task type and processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 185–211. https://doi.org/10.1111/1467-9922.00071

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193–219). John Benjamins.

Swain, M. (1985). Large scale communicative testing: A case study. In Y. Lee, C. Fok, R. Lord & G. Low (Eds.), *New directions in language testing* (pp. 35–46). Pergamon Press.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–277). Amsterdam: John Benjamins.

Weir, C., O'Sullivan, B., & Horai, T. (2006). Exploring difficulty in speaking tasks: An intra-task perspective. *IELTS Research Reports* (Vol. 6, pp. 1–42). IELTS Australia and British Council.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85–106. https://doi.org/10.1177/026553229701400105

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*(1), 1–24. https://doi.org/10.1080/15434300903031779

Xi, X. (2010). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing*, *27*(1), 73–100. https://doi.org/10.1177/0265532209346454

Yuan, F., & Ellis, R. (2003). The effects on pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *21*(1), 1–27.

Appendix 1: Questionnaire Items

1) During your speaking exam, you were given some preparation time to plan your monologue.
Did you make any notes during your preparation time?

a) Yes, I made a lot of notes.
b) Yes, but just some words.
c) No, I didn't make any notes.

2) If you made notes, what was the purpose behind them? (You may choose more than one response, if necessary.)

a) To note down the ideas to speak about.
b) To structure my monologue.
c) To plan what grammar I will use.
d) To note down useful vocabulary.
e) Other:

3) If you didn't make notes, what did you use your preparation time for? (You may choose more than one response, if necessary.)

a) To think of the ideas to speak about.
b) To think of how to structure my talk.
c) To think about the grammar I will use.
d) To think about useful vocabulary.
e) To calm myself down before I start talking.
f) I wasn't thinking of anything.
g) Other:

4) You were given 30" to prepare. Was the time enough?

a) Yes.
b) No.

5) Level of the exam you took (circle): A1 / A2 / B1 / B2 / C1 / C2

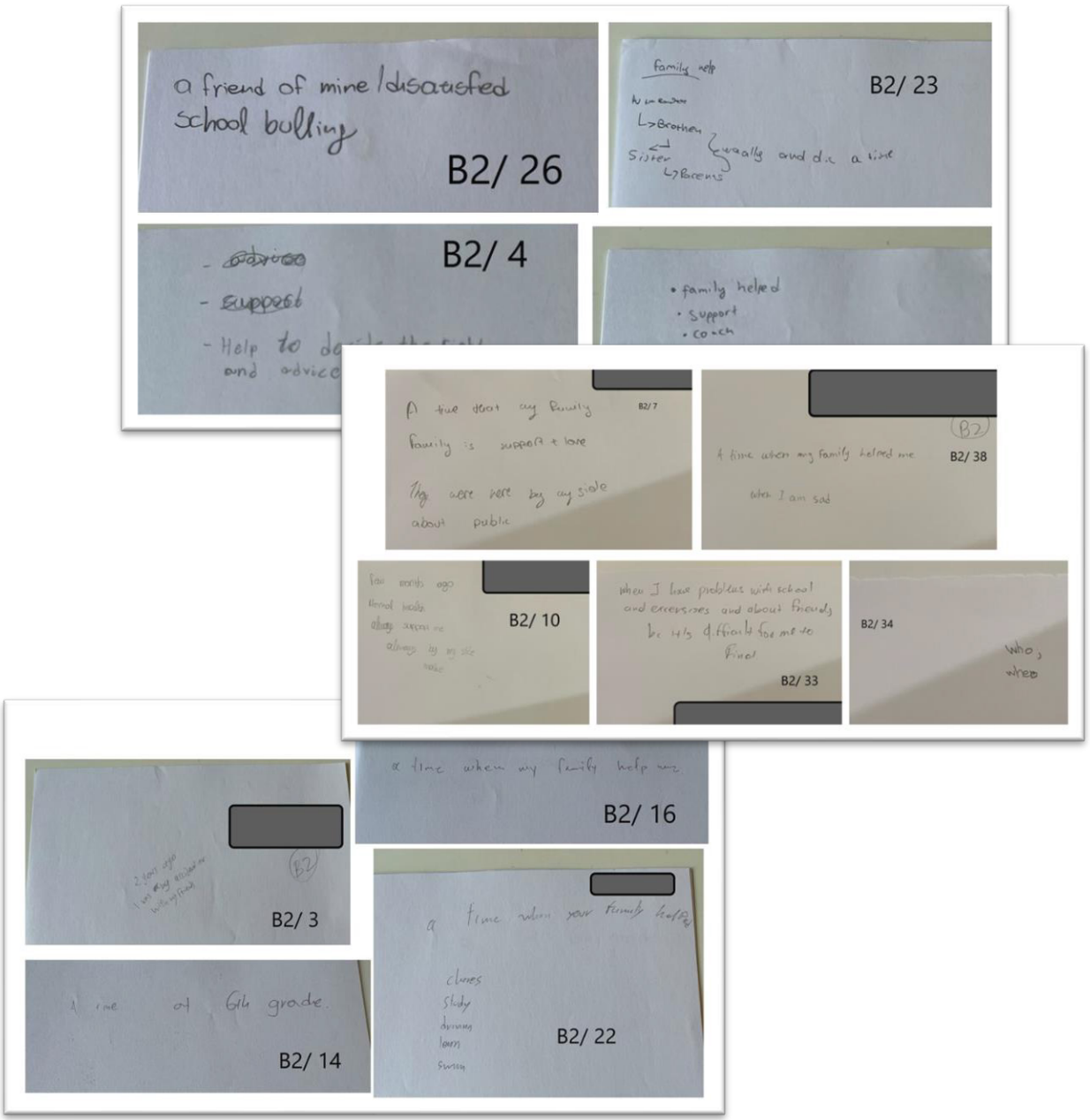
Appendix 2: LANGUAGECERT IESOL Speaking B2 Markscheme and Descriptors

	Task Fulfilment and Coherence	**Accuracy and Range of Grammar**	**Accuracy and Range of Vocabulary**	**Pronunciation, Intonation and Fluency**
3	- Tasks are fulfilled with ease and confidence - Turn taking is spontaneous and natural - Contributions are fully relevant and detailed - Significant points are appropriately highlighted with supporting detail - Discourse is clear and coherent and produced in an appropriate style with a wide range of B2 level cohesive devices	- A wide range of B2 level grammar is used - There is a consistently high level of accuracy and control - Occasional errors may occur, but are often corrected	- A wide range of B2 level vocabulary is used to deal with the tasks - Choice of vocabulary is generally appropriate and effective	- Pronunciation is clear and natural - Intonation is used to convey meaning effectively - The flow of language is maintained effectively - No evident hesitations

2	- Tasks are fulfilled with relative ease - Turn taking is naturally handled - Contributions are mostly relevant - Intended message is clearly communicated. Misunderstandings are rare - Discourse is mostly clear and coherent with use of B2 level cohesive devices	- A good range of B2 level grammar is used - There is a good level of accuracy and control - Some errors may occur, but the message is always communicated	- A sufficient range of vocabulary is used to deal with the B2 tasks - Choice of vocabulary is generally appropriate and effective - Some vocabulary errors occur, but do not impede communication	- Pronunciation is reasonably clear and easily understood - Stress and intonation patterns are appropriately used to help convey meaning - The flow of language is generally maintained despite some hesitation - No undue strain on the listener
1	- Tasks remain largely unfulfilled - Interaction is only maintained with the support of the interlocutor - Little natural turn taking takes place - Contributions lack relevance - Intended message is only communicated with difficulty - Ideas are linked together simply and	- Range of grammar is too limited to deal with the B2 level tasks - Frequent errors are noticeable, and may impede communication	- Range of vocabulary is too limited to deal with the B2 level tasks - Vocabulary errors may make the message difficult to follow	- Unclear pronunciation leads to undue strain on the listener - Inappropriate stress and intonation patterns impede communication - Frequent hesitations are evident, with repetition and attempts to repair language

	may be difficult to follow			
0	- The tasks are unfulfilled and intended message is not successfully communicated - Ideas are difficult to follow and not linked together into connected speech - OR insufficient sample of language to assess	- Inadequate range of grammar - Frequent errors impede communication - OR insufficient sample of language to assess	- Lacks the vocabulary to deal with the B2 level tasks - The message is obscured by vocabulary errors - OR insufficient sample of language to assess	- Unclear pronunciation and/or intonation prevents clear understanding - Frequent hesitation places strain on the listener - OR insufficient sample of language to assess

Appendix 3: Samples of Test-Takers' Note making Sheets



Chapter 10: A Comparability Study of Handwritten versus Typed Responses in High-Stakes English Language Writing Tests

Irene Stoukou, Yiannis Papargyris and David Coniam

Abstract

This chapter investigates fairness in writing test scores in terms of candidates who completed a writing test either by hand or typed, on a computer. The data for this large-scale comparability study comprise candidates taking English language writing tests at four CEFR levels – B1 to C2 in the period 2019–2022. The data were analysed via effect size differences and equivalence tests. Measured by effect size, a small amount of difference was apparent in scores obtained between the two production modes at B1, B2 and C1 levels. At C2 level, there was a medium effect size, indicative of a difference in favour of computer-produced scripts. Differences observed on equivalence tests – an adaptation of the standard t-test – were not found to be statistically significant. The contribution of the research to knowledge lies in the fact that (with the exception of C2 level) – whether writing tests are written by hand or on computer, while there is a slight skew towards higher scores with computer-processed texts, candidates generally receive similar scores in both modes. Practically, candidates may elect to write either on paper or on computer without fear of bias.

Keywords: handwritten scripts; computer-processed scripts; effect size; equivalence t-tests

Introduction

There is a substantial literature on score equivalence obtained from handwritten (HW) and computer-processed (CP) scripts. Indeed, research into score equivalence between handwritten and computer-processed scripts stretches back to the 1960s when the word-processing of scripts first began. To put these issues into perspective, the current section first presents an overview of the research – which presents provide contrasting results. The section concludes with the research gap being explored in the current research.

While some studies have revealed better performance by candidates writing by hand; others have reported the opposite, with higher CP scores; and, in contrast, no significance has been found for either mode of delivery in other studies. A review of the research from the different angles is presented below.

Some of the earliest research was by Marshall and Powers (1969), in whose study neat handwritten essays scored higher than typed ones. Mazzeo and Harvey's (1988) study of handwritten and computer-processed scripts indicated better performance in HW mode, which they attributed, understandably at the time, to lack of familiarity with the technology.

Arnold et al. (1990) reported computer-processed scripts receiving lower scores than handwritten scripts. Sweedler-Brown (1991) reported likewise, although only with lower ability scripts. In Powers et al.'s (1994) and Russell and Tao's (2004) studies, students' HW scripts scored higher than the same students' comparable CP scripts. Bridgeman and Cooper (1998) in a study involving Graduate Management Admissions Test scores reported higher scores with HW than with CP scripts. Klein and Taub (2005) reported a teacher bias for legible HW scripts. In Breland et al.'s (2005) study of TOEFL candidates, HW scores, related to general English language ability, were reported.

While numerous studies have reported handwriting-based scripts to have received higher scores, there have also been many studies reporting computer-processed-based scripts to have received higher scores. Some studies showing such advantage are outlined below.

An overall advantage for CP texts has been reported in certain studies (Sprouse and Webb, 1994; Peacock, 1988; Hughes and Akbar, 2010). On the issue of quality, Peacock (1988) reported an advantage for low-quality CP scripts.

Peacock (1988) also reported an advantage regarding text type for CP essays where the essays were not related to external sources.

In Canz et al.'s large-scale (2020) study, CP scripts received higher grades despite raters being highly trained raters.

Russell and Plati (2000) reported lower secondary school students performing better under CP conditions. In Goldberg et al.'s (2003) meta-analysis of 26 writing studies of K-12 students writing in CP or HW modes, results indicated higher text quality for the CP scripts.

Other confirmatory studies for students achieving higher grades in CP mode include Russell and Haney (1997) and Russell and Plati (2001).

In addition to studies citing an advantage for either mode, here have also been studies where neither mode has been reported as conferring an advantage, as outlined below.

While positive findings have been reported for both modes, a number of studies have reported no significant difference in terms of grade received in either CP or HW mode. Among these are: Wise and Plake, 1989; Wright and Linacre, 1994; Taylor et al., 1999; Russell, 1999; MacCann et al., 2002; Horkay et al., 2006; Boulet et al, 2007; King et al., 2008; Mogey et al., 2010; Chan et al., 2018.

As may be seen from the studies reported above, there is evidence for all positions: that under certain conditions CP scripts receive higher scores; under others that HW scripts score higher, with many studies also reporting no significant difference between modes.

Differences notwithstanding, it is nonetheless the case that with improvements in technology in terms of usability, speed and lower cost (see Lim and Wang, 2016), the use of a computer to produce essays in a variety of situations – classwork, homework and examinations – is increasing. Indeed, with the recent COVID-19 pandemic, greater acceptance has been observed of the use of computers and technology (Hodges et al., 2020).

In light of the above, it is worth considering the question of whether the ability or preference to use a computer in an examination is related to age. Older candidates do not necessarily opt for CB tests as such; it is simply the route they follow which leads them to an online-proctored environment (i.e., navigating the internet, selecting an exam provider online, registering, booking a slot and managing their time etc.). Against this backdrop, for more mature candidates, the CB component is simply part of the overall context.

Over the past three years, that is, during the period of COVID in 2019–2022, many examination bodies experienced exponential increases in online-administered examinations (see e.g., Ockey, 2021). LANGUAGECERT English language tests are available in either traditional centre-based or online proctored (OLP) delivery modes (Coniam et al., 2021). During the COVID pandemic, LANGUAGECERT saw a great increase in its OLP mode of delivery, and a concomitant increase in writing tests produced by computer as opposed to being handwritten. While the research outlined above has presented different perspectives on the two modes of delivery – computer-processed versus handwritten – and how the mode might confer an advantage on scores, little research has been conducted in the past few years – and certainly not in the context of the huge increase in computer-processed writing test scripts against the backdrop of COVID.

This is therefore the research gap that the current study fills in the context of computer-processed versus handwriting writing test scripts. Using comparatively large writing test datasets (a considerable number of which had been administered during the COVID pandemic period) at differing CEFR levels of ability, the study explores to what extent the mode of script production impacts on candidate score.

Method

This section outlines the study in terms of research design, the data and data analysis. Background to the LANGUAGECERT Writing Test is first presented to situate the study.

The IESOL Writing Test

The data in the study come from four examinations – at CEFR levels B1–C2, which form part of the International ESOL (IESOL) suite of English language tests. The Writing Tests comprise two different writing tasks tapping a range of writing skills. Table 1 elaborates.

Table 1: IESOL 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
C1	a neutral or formal text for a public audience	150-200	a text using informal language	250-300
C2	a neutral or formal text for a public audience	200-250	a text using informal language	250-300

Each task is scored on four levels (0-3) against four subscales which for the most part are double-marked before final scores are amended or confirmed and signed off by a more senior member of the assessment team, usually a chief examiner. (Refer to https://www.LANGUAGECERT.org/en/language-exams/).

Candidates may take the examination either at a physical centre or by online-proctored mode. If they take the examination at a centre, they generally handwrite. While it is possible to do a computer-based test at a physical centre, this option is not very popular; most candidates handwrite tests at centres. When tests are taken online, a locked-down computer is used. It should be noted that the term 'computer-processed' is used in the current chapter to indicate that candidates write on a 'bare-bones' computer; they do not have access to a word processor or any of the more advanced facilities such as grammar/spellchecking that a word processor offers.

All Writing Test markers hold professional accredited English language qualifications and experience as English language teachers. All prospective markers undergo a standardisation and training programme before being certified as qualified markers (for details, see Papargyris and Yan, 2022). The training programme involves marking sample scripts and prospective markers must demonstrate they can mark accurately and consistently before they are certificated as markers. Checking takes place by a group of chief examiners during live marking, and if markers are suspected of marking inaccurately and/or inconsistently, they may be removed from the marking session and/or retrained or even dismissed. Markers are monitored on an ongoing basis as well as attending standardisation sessions, again on a regular basis. LANGUAGECERT markers mark across CEFR levels (Papargyris and Yan, 2022). At any one time, there may well be in the region of 200 markers marking different numbers of scripts at the different CEFR levels. While the scope of the current study does not involve an examination of marker performance, the reader is referred to Coniam et al. (2022) where an exploration using Many-Facet Rasch Analysis into marker performance can be found.

The Current Study

This section presents details on candidates' scores against the two modes of script production. Table 2 below provides detail on the number of candidates at each CEFR level for each mode. The data collection period extended over the three-year period from mid 2019 to mid 2022. Although not germane to the current study, it should be noted that the current study involved 143 different markers.

Table 2: Candidate sample sizes

Level	Mode	N	Level sample
B1	CP	3108	22727
	HW	19619	
B2	CP	14878	27590
	HW	12712	
C1	CP	7674	10330
	HW	2656	
C2	CP	2869	4363
	HW	1494	

Legend: CP=computer-processed; HW=Handwritten

At B1 level, the candidature comprises many school students. It is therefore not perhaps surprising that the majority of scripts at this level were handwritten. As one moves up the level, and demands of certification for study, work, immigration purposes come more to the fore, candidates tend to be slightly older and more computer literate. More online-proctored (OLP) examinations take place at this level, a situation exacerbated by COVID, and support for why computer-processed (CP) scripts outnumber handwritten (HW) scripts at B2-C2.

Hypotheses

The hypothesis pursued in the study is that scores awarded to either of the two modes of script production – computer-processed or handwritten – will not be significantly different. Three sub-hypotheses are pursued:

1. The difference between the mean scores for the two written script modes will be less than 5% for any given CEFR level.
2. Only small effect size differences will be noted between the two modes.
3. On equivalence tests, significance will not emerge against specified upper and lower bounds for any CEFR level.

Findings and Discussion

Two sets of data for the Writing Test are presented. The first set of analyses contains descriptive statistics: means (maximum 25) for the two modes, standard deviations and effect size differences. The second set of analyses consists of equivalence independent samples t-tests ("equivalence tests"). Equivalence tests – as opposed to regular t-tests – permit for significance to be explored by specified upper and lower bounds (Lakens, 2017). The two bounds define the extent of variation of t values with respect to the populations being tested. If the t value falls within the estimated range, the two populations may be seen to be equivalent.

Descriptive Statistics

Descriptive statistic results are in provided in Table 3 for the two types of writing for the four CEFR levels. The final two right-hand columns contain detail on score and effect size differences between the two modes. Effect size differences are reported in terms of Cohen's d, for which a small effect is generally 0.2, a medium effect 0.5, and a large effect 0.8 (Glen, 2021).

Table 3. Writing mode descriptives

Level	Mode	N	Mean	SD	Raw score (%) difference	Effect size differences (Cohen's d)
B1	CP	3108	18.75	4.63	0.80 (3.20%)	0.17
	HW	19619	17.95	4.72		
B2	CP	14878	18.85	4.68	0.80 (3.21%)	0.17
	HW	12712	18.04	4.67		
C1	CP	7674	17.60	4.80	1.13 (4.53%)	0.23
	HW	2656	16.46	4.86		
C2	CP	2869	18.13	4.77	2.57 (10.28%)	0.55
	HW	1494	15.56	4.46		

Key: CP=computer-processed; HW=handwritten

Equivalence Tests

Table 4 below presents equivalence test results comparing handwritten (HW) and computer-processed (CP) script production modes. Upper and lower bounds have been set at +/- 0.05 of the raw score (see Lakens, 2017). As mentioned, critical decisions regarding equivalence revolve around whether estimated t values are between the upper and lower bound. In Table 4 below, p values indicate significance with respect to upper and lower bound t values going beyond specified bounds.

Table 4: Equivalence samples t-tests

Test Level	*Statistic*	*t*	*df*	*p*
B1	*upper bound*	*9.36*	*22725*	*< .001*
	t value	*8.81*	*22725*	*< .001*
	lower bound	*8.26*	*22725*	*1.00*
B2	*upper bound*	*15.12*	*27588*	*1.00*
	t value	*14.23*	*27588*	*< .001*
	lower bound	*13.34*	*27588*	*< .001*
C1	*upper bound*	*9.99*	*10328*	*1.00*
	t value	*10.45*	*10328*	*< .001*
	lower bound	*10.91*	*10328*	*< .001*
C2	*upper bound*	*16.92*	*4361*	*1.00*
	t value	*17.26*	*4361*	*< .001*
	lower bound	*17.59*	*4361*	*< .001*

Discussion

The results above provide a consistent picture: at all levels, candidates who produced computer-processed scripts scored higher than did candidates who produced handwritten scripts. This finding echoes the study by Goldberg et al. (2003) who analysed studies of students writing in CP or HW modes, with results indicating CP scripts being rated more highly. As Lim and Wang (2016) report, the use of a computer to produce essays in many school situations is increasing. It may simply be the case that such increasing use of the computer results in a vicious, or virtual, cycle (depending on one's point of view), whereby writing on computer becomes the norm and the mode to which people, examination candidates included, are simply becoming more accustomed.

The results for higher scores obtained on computers may be due to a number of factors. One consistent feature mentioned by LANGUAGECERT markers in post-marking reports is that of the legibility (or lack of it) encountered in many handwritten scripts. Be that as it may, the main issue is that at CEFR levels B1-C1, the difference between the two modes is less than 5%, a figure generally taken as being indicative of significance.

What then might be the possible reasons for candidates using a computer to produce their script – in particular at the higher CEFR levels – to obtain comparatively higher scores? One possible explanation may be found in the candidates' background. In a survey (in mid-2022) of over 40 LANGUAGECERT Writing Test markers, markers noted that, at the CEFR A and B levels, there were more younger candidates. These younger candidates were more used to writing on paper than using a computer. More proficient candidates – in particular those at C2 – were noted by some markers as being older and more computer literate. Markers perceived these two factors as helping to account for the skew towards higher scores achieved on computer-processed scripts.

Conclusion

This study has reported on comparability of scores awarded to candidates who completed Writing Tests either through handwriting or by using a computer at CEFR levels B1 to C2.

The key hypothesis in the study was that mean scores and performance on the Writing Test in either mode would not be significantly different from each other; i.e., that candidate scores would not be influenced by the writing mode. Specifically, three hypotheses were being investigated.

The first hypothesis was that differences between the mean scores for the two modes of test production would be less than 5% for any given CEFR level. This was the case for levels B1, B2 and C1. It was not the case for C2 where differences were greater than 5%. While the hypothesis was confirmed for B1, B2 and C1, it was rejected for C2.

The second hypothesis was that, at worst, only small effect sizes would be reported between the two writing modes. This was indeed the case with B1, B2 and C1. At the C2 level, however, a medium effect size was observed, causing the hypothesis to be rejected.

The third hypothesis was that, for any given CEFR level, significance between upper and lower bounds would not be observed on equivalence t-tests. Significance was not observed for either bound at any test level. Consequently, the two script writing modes can be taken as broadly equivalent, and the hypothesis can be accepted.

While differences at B1 and B2 were minimal, it could be seen that as one moved up the CEFR levels, the relative score gain conferred by using a computer increased. At B1 and B2 the difference was 3%. At C1, it was 5%, and at C2, 10%.

As mentioned above, use of a computer in an examination may be seen to be related to age in that older candidates simply follow an online path which leads to an online-proctored environment (i.e., navigating the internet, selecting an exam provider online, registering, booking a slot and managing their time etc.). For older candidates, the CP component in terms of how a test is taken may well be seen as simply a part of an online path they have followed.

The current study has been purely quantitative. A further study is currently exploring Writing Test markers' views on the effect of certain linguistic or textual features on candidates' scripts. Echoing markers' comments alluded to above, a more fine-grained examination lies in determining to what extent demographic factors such as age might have an effect on results obtained from writing tests by hand versus on computer.

Another aspect of the interaction between digital environment and textual production, worth exploring in the future, is that of task requirements vis-a-vis the support each environment allows. In a digital environment for instance, candidates have the option of employing a variety of content control features (provided these are made available by the test provider). Such features may significantly contribute to the authoring, editing, and proofreading of longer, complex and structurally challenging texts and thus account for the increasing discrepancy between scores, which culminates at C2.

The research literature revealed support for all modes: for handwritten scripts, for scripts written on computer, and for there being no difference. The current study, however, lends support to the view that, while differences remain, it is computer-processed scripts that certain candidates tend to score higher on.

A generally greater uptake of the use of computers is seen in the production of text – for all purposes, not just examinations. In the light of such uptake, one potential solution to the discrepancy score situation, as one looks to the future, is that all scripts be computer processed. Indeed, many professional examinations – law examinations, for example (Steel et al., 2019) – are now required to be done solely on computer as are the Association of Chartered Certified Accountants' (ACCA) financial and accounting examinations.

The COVID pandemic has accelerated the computer processing of scripts, with many more candidates taking exams online rather than on paper (Fuller et al., 2020; Abduh, 2021). For such a move to be accepted more widely, however, school students in particular need to have easy access to a computer and to be computer literate. This is contingent upon schools moving increasingly towards total computer-based work, with each child having their own laptop for continual school and home use, as with Uruguay's Plan Ceibal (see Segovia et al., 2022), for example. In the UK, the government Office of Qualifications and Examinations Regulation (Ofqual) has recently announced a three-year plan to explore the possibility of across-the-board online testing for students (Ofqual, 2022). Indeed, in the long run, what Mogey and Fluck (2015) describe as "post-paper assessment" is possibly what education and assessment authorities should be considering. Whether these changes will happen quickly will be observed and reported on in due course.

References

Abduh, M. Y. M. (2021). Full-time online assessment during COVID-19 lockdown: EFL teachers' perceptions. *Asian EFL Journal*, 28(1.1), 26-46.

Association of Chartered Certified Accountants Association. *Computer-based exams*. (n.d.). https://www.accaglobal.com/vn/en/student/exam-support-resources/fundamentals-exams-study-resources/f1/technical-articles/computer-based-exams.html.

Arnold, V. (1990). Do students get higher scores on their word-processed papers? A study of bias in scoring hand-written vs. word-processed papers. The Educational Resources Center. Whitter, CA Rio Hondo College.

Boulet, J. R., McKinley, D. W., Rebbecchi, T., & Whelan, G. P. (2007). Does composition medium affect the psychometric properties of scores on an exercise designed to assess written medical communication skills?. *Advances in Health Sciences Education*, 12(2), 157-167.

Breland, H., Lee, Y. W., & Muraki, E. (2005). Comparability of TOEFL CBT essay prompts: response-mode analyses. *Educational and Psychological Measurement*, 65(4), 577-595.

Bridgeman, B., & Cooper, P. (1998). Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test. http://eric.ed.gov/?id=ED421528.

Canz, T., Hoffmann, L., & Kania, R. (2020). Presentation-mode effects in large-scale writing assessments. *Assessing Writing*, 45, 100470.

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing*, 36, 32–48. https://doi.org/10.1016/j.asw.2018.03.008.

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. https://doi.org/10.5539/elt.v14n8p58.

Coniam, D., Stoukou, I., Lee,. T., & Milanovic, M. (2023). SELT IESOL Writing Test quality. London, UK: LanguageCert.

Fuller, R., Joynes, V., Cooper, J., Boursicot, K., & Roberts, T. (2020). Could COVID-19 be our 'There is no alternative'(TINA) opportunity to enhance assessment?. *Medical Teacher*, 42(7), 781-786.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment,* 2(1).

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. *EDUCAUSE Review*. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, 5(2). https://ejournals.bc.edu/index.php/jtla/article/view/1641.

Hughes, J., & Akbar, S. (2010). The influence of presentation upon examination marks. 11th Annual Conference of the Subject Centre for Information and Computer Sciences, 178–182.

King, F.J., F. Rohani, C. Sanfilippo, N. White. (2008). Effects of handwritten versus computer-written modes of communication on the quality of student essays. Center for Advancement of Learning and Assessment (CALA Report). http://www.cala.fsu.edu/files/writing_modes.pdf, 2008.

Klein, J., & Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing*, 10, 134–148. https://doi.org/10.1016/ j.asw.2005.05.002.

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.

MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology*, 33(2), 173-188.

Marshall, J. C., & Powers, J. C. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement*, 6, 97–101. https://doi.org/10.1111/j.1745-3984.1969.tb00665.x.

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature. New York: College Entrance Examination Board.

Mogey, N., & Fluck, A. (2015). Factors influencing student preference when comparing handwriting and typing for essay style examinations. *British Journal of Educational Technology*, 46(4), 793-802.

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: Letting the students choose. *ALT-J Research in Learning Technology*, 18, 29–47. https://doi.org/10.1080/09687761003657580.

Ofqual. (2022). Ofqual corporate plan 2022 to 2025. Coventry, UK: https://www.gov.uk/government/organisations/ofqual.

Ockey, G. J. (2021). An overview of COVID-19's impact on English language university admissions and placement tests. *Language Assessment Quarterly*, 18(1), 1-5.

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.

Peacock, M. (1988). Handwriting versus word processed print: An investigation into teachers' grading of English language and literature essay work at 16+. *Journal of Computer Assisted Learning*, 4, 162–172. https://doi.org/10.1111/j.1365-2729.1988.tb00173.x.

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31, 220–233. https://doi.org/10.1111/j.1745-3984.1994.tb00444.x.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Educational Policy Analysis Archives*, 7(20), 1–47.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, 5(3), 1-19.

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *Teachers College Record*. http://www.tcrecord.org/Content.asp?ContentID=10709.

Russell, M., & Tao, W. (2004). The influence of computer-print on rater scores. *Practical Assessment, Research & Evaluation*, 9(10), 1–14.

Segovia, G. D., Jang, E. H., Manuel, C., & Staal, E. (2022). Uruguay: Rethinking teacher training and global education through Plan Ceibal. In Reimers, F.M., Budler, T.A., Irele, I.F., Kenyon, C.R., Ovitt, S.L., & Pitcher, C.E. *Reimagining our Futures Together. A New Social Contract For Education*, pp. 449-477. Paris: UNESCO.

Sprouse, J. L., & Webb, J. E. (1994). The Pygmalion effect and its influence on the grading and gender assignment on spelling and essay assessments. *ERIC Document*, ED 374096.

Steel, A., Moses, L. B., Laurens, J., & Brady, C. (2019). Use of e-exams in high stakes law school examinations: Student and staff reactions. *Legal Education Review*, 29, 1.

Sweedler-Brown, C. O. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic scoring of essays. *Research and Teaching in Developmental Education,* 8, 5–14.

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

Trobia, A. (2011). Cronbach's Alpha. In Lavraka, P. (ed.) *Encyclopedia of survey research methods*, Vols 1 & 2, pp. 168-169. Thousand Oaks, Ca.: Sage Publications.

Wise, S., & Plake, B. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, 8(3), 5–10.

Wright, B., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370. http://www.rasch.org.

Chapter 11: The Delivery of Speaking Tests in Traditional or Online Proctored Mode: A Comparability Study

Michael Milanovic, Tony Lee and David Coniam

Abstract

This chapter investigates the comparability of 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 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 that mode of test delivery does not significantly affect test taker scores.

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 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, 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 & 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 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 & Ilyas, 2013, Berrada et al., 2021).

Khan & 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 & 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 does online assessment. Online teaching, which now has 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 & 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, working on their own. 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 around security, the "vulnerability" of online tests and academic dishonesty (see Corrigan-Gibbs et al., 2015; Coniam et al., 2021). Such issues are key, especially when examinations 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 & 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 assessing 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).

Sujana (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 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. However, the process is entirely machine mediated.

Against the backdrop of the covid pandemic, assessment of all forms moved, with differing degrees of success (Ali & Dmour, 2021), to various online modes. As might be expected – following the discussion above of the complexities of assessing speaking – it was indeed assessing students' oral proficiency that emerged as most challenging for many educators. Forrester (2020) elaborates upon 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 & 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: Castillo & Doe, 2017; Lee, 2020.

Examinations Conducted using Online Invigilation / 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 & 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 & 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 & 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 using 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 score for examinations taken onsite by examinees. While they attribute some of the differences in results to the makeup 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 grades awarded being: 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 of assessment that is benchmarked against the CEFR (see Papargyris & 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 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.

In terms of test reliability, since Speaking Test scores are obtained via the four rating scales, 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	Omega	95% CI lower and upper bounds
B1	0.64	0.64-0.65
B2	0.62	0.62-0.63
C1	0.65	0.64-0.66
C2	0.72	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 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. 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.

Hypotheses

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:

(1) That, at worst, only small effect size differences between the two modes will be observed.

(2) That on equivalence tests, significance will not 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 hypotheses were being investigated.

The first hypothesis 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, the fact that a small-to-medium effect size was observed for C2 meant that the hypothesis could not be accepted.

The second hypothesis 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, and the hypothesis was accepted. 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 actual makeup 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 perhaps 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 do 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 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 exacerbated 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.

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. doi.org/10.24059/olj.v21i1.885.

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. doi.org/10.18178/ijiet.2021.11.5.1515.

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 P. Falvey & D. Coniam (Eds.), *Radical Solutions for Education in a Crisis Context* (pp. 191-205). Singapore: Springer.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology, 54*(2), 199-231.

Castillo, M. S., & Doe, R. (2017). Mobile and nonmobile assessment in organizations: Does proctoring make a difference? *Psychology, 8*(06), 878. doi.org/10.4236/psych.2017.86057.

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. doi.org/10.1016/j.caeo.2021.100061.

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. doi.org/10.46451/ijts.2022.01.03.

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. doi.org/10.5539/elt.v14n8p58.

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. doi.org/10.1145/2810239.

Fall, T., Adair-Hauck, B., & Glisan, E. (2007). Assessing students' oral proficiency: A case for online testing. *Foreign Language Annals, 40*(3), 377-406. doi.org/10.1111/j.1944-9720.2007.tb02865.x.

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. Retrieved from 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 P. Falvey & D. Coniam (Eds.), *Radical Solutions for Education in a Crisis Context* (pp. 85-98). Singapore: Springer.

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. Retrieved from 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. doi.org/10.1080/08923647.2020.1796376.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures, 14*(1), 1-24.

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. doi.org/10.12669/pjms.36.COVID19-S4.2795.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. doi.org/10.1177/1948550617697177.

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. doi.org/10.13106/jafeb.2020.vol7.no8.653.

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.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Madsen, H. S. (1983). *Techniques in testing*. New York: Oxford University Press.

Mays, T. J. (2021). Teaching the teachers. In P. Falvey & D. Coniam (Eds.), *Radical Solutions for Education in a Crisis Context* (pp. 163-176). Singapore: Springer. doi.org/10.1007/978-981-15-7869-4_11.

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.

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.

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, 10*(1), 439.

Sujana, I. M. (2016). Assessing oral proficiency: Problems and suggestions for elicitation techniques. Retrieved from 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.

Todd, R. W. (2020). Teachers' perceptions of the shift from the classroom to online teaching. *International Journal of TESOL Studies, 2*(2), 4-16.

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. doi.org/10.1016/j.iree.2020.100196.

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 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 ρ 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* (2nd ed.). 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, K. (1999, October 8). *Interpreting correlation coefficients*. Retrieved from http://acad.cl.uh.edu/itc/educ6032/course/resources/unit2/index.htm

Glaser, D. N. (1999). The controversy of significance testing: Misconceptions and alternatives. *American Journal of Critical Care, 5*(5), 291-296.

Goodman, L. (1990). Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. *Sociological Methodology, 20*, 249-294.

Gronlund, N. E. (1985). *Measurement and evaluation in teaching*. New York: Macmillan.

Haney, W., Fowler, C., Wheelock, A., Bebell, D., & Malec, N. (1999). Less truth than error? An independent study of the Massachusetts Teacher Tests. *Education Policy Analysis Archives, 7*(4).

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, P. (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. Retrieved from https://testing.byu.edu/handbooks/bettertests.pdf.

ISBN: 978-9925-34-960-9