CERTIFYING QUALITY IN ASSESSMENT AND LEARNING

Research and Validation at LANGUAGECERT

Volume 4

Edited by Leda Lampropoulou and David Coniam

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING
Research and Validation at LANGUAGECERT Volume 4

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

Publisher details: LANGUAGECERT, London, UK
Date of Publication: June 2026
ISBN: 978-9925-39-382-4

Foreword

Byron Nicolaides, CEO, PeopleCert Group

This fourth volume of Certifying Quality in Assessment and Learning is bound in red, representing the brand colours of City & Guilds, which we welcomed into the PeopleCert family of brands in 2025. City & Guilds, founded in London in 1878, delivers a comprehensive range of 2,000 vocational and technical qualifications across 25 industries and is recognised worldwide as the benchmark of quality in training and assessment.

The City & Guilds red joins the colour palette of brands that represent the PeopleCert trusted and complementary portfolio of ITIL, PRINCE2, DEVOPS INSTITUTE and LANGUAGECERT. Together, the colours mix and the brands combine to create a global powerhouse for lifelong learning. Committed to covering the spectrum of skilling, upskilling and reskilling to fulfil the PeopleCert vision of helping all our stakeholders build the brightest possible futures.

From the 7-year-old learning their first words of English to project managers, from private language schools to multinationals and from the vocational sector to government ministries, our mission is to help individuals turn their dreams and ambitions into reality, and organisations reach their targets and achieve their goals.

PeopleCert is the world's lifelong learning partner. We deliver our brand portfolio across 200+ countries and territories to millions of people, as well as to 50,000 corporations (82% of the Fortune 500) and 800 government organisations in 50 countries. Delivery is through an extensive network of 2,500 partners, more than 3,000 recognising institutions, and our award-winning, proprietary online invigilation platform. This global network speaks the PeopleCert common language of certification to connect and

create pathways between stakeholders, and provide proof of achievement, skills and abilities.

The Certifying Quality in Assessment and Learning series are snapshots of the LANGUAGECERT ongoing programme of research and validation that ensures the English language testing pathways are built on solid foundations and evidence of fitness for purpose. This research and validation programme exemplifies PeopleCert core values. The quality of the LANGUAGECERT research and tests, the innovation that defines the LANGUAGECERT portfolio, and the passion of the research and assessment teams to produce best-in-class language assessment. Proof of the integrity that is essential to the language testing ecosystem because all stakeholders must trust the validity, reliability and fairness of test scores. As is clarity: research needs to be shared and understood with and by stakeholders.

The global language testing ecosystem is complex and dynamic. LANGUAGECERT must respond with velocity to this environment, which is influenced by the interaction of many factors, such as world events, government regulation, student choice, and the challenges and opportunities of new technologies. Velocity means acting with agility, focus and precision to meet changing circumstances and stakeholder needs. The papers collected in this volume demonstrate LANGUAGECERT's ability to innovate rapidly, without sacrificing the validity, reliability and fairness that make our tests fit for purpose.

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

LANGUAGECERT fulfils the PeopleCert vision and mission by offering English tests for every learner, assessing and developing communicative ability for all ages and purposes.

We provide tests for young and school-age learners that lay out a pathway to proficiency, tests for admissions purposes to help students survive, thrive and flourish at university, as well as tests for the workplace, migration and adult learners. Our portfolio is built on the principles of communicative language learning, teaching and assessment, underpinned by research and validation, ensuring that all our tests are valid, reliable and fair.

Communicative language testing evolved from the communicative approach to language learning and teaching, which moved away from approaches such as the Grammar-Translation and Audio-lingual methods. The focus of learning, teaching and assessment shifted from an emphasis on textbook knowledge of how a language works to using language in the real world, from memorisation to spontaneity and from artificiality to authenticity.

Until recently, there was a broad consensus and application within the language assessment profession and community about what constitutes communicative language testing. The emergence of AI-first test design and delivery is at odds with the axioms of communicative language assessment. AI-first assessment models are led by the limitations of the respective algorithms rather than the communicative needs of test takers and test users. AI is a good servant, but a bad master.

At LANGUAGECERT, our approach to test design and delivery is human-led and people-first. AI is a tool and not the craftsman. Humans are always in the loop, in charge and making the final decisions. Our hybrid approach combines the speed and efficiency of AI for item generation, marking and malpractice detection with human expertise, knowledge and judgment.

AI offers testing organisations, test takers and test users many benefits in terms of speed and efficiency. Faster question paper production, faster marking and faster results. But efficiency must not be at the expense of effectiveness. AI-generated items and marking must be as good and effective as human item writing and marking.

For a test to be valid, reliable and fair, it must be fit for the purpose it is intended for. A fit-for-purpose test is efficient and effective in eliciting and assessing the right range of real-world communicative ability at the right levels and authentically representing the targeted domain of language use.

A fit-for-purpose test must also be trusted, and to be trusted, a test must be transparent. Without transparency, a test's validity, reliability and fairness cannot be meaningfully evaluated by test users. It is important that all our stakeholders, from the parents of our youngest learners, our test takers, to recognising institutions, companies, and governments, can see for themselves the research that underpins and is incorporated into our tests.

Open research practices and the evidencing of validity, reliability and fairness are ethical responsibilities for test providers. They inform decision making by test users, contribute to and advance knowledge in the field of language assessment and strengthen public trust in testing.

This fourth volume of Certifying Quality in Assessment and Learning is part of our commitment to transparency, equability and accountability.

Contents

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 Chief Research Advisor at LANGUAGECERT. He has been working and conducting research in English language teaching, education and assessment for over 50 years. His main publication and research interests are in language assessment, language teaching methodology, and academic writing and publishing.

Peter Falvey is an Honorary Professor at The Education University of Hong Kong. He is a teacher educator and a former Head of Department in the Faculty of Education of the University of Hong Kong. His main publication and research

interests are in language assessment, second language writing methodology, and text linguistics.

Anthony Green is Professor in Language Assessment and Director of the Centre for Research in Language Learning and Assessment at the University of Bedfordshire. He has extensive experience in language teaching and assessment and has published several influential works on language assessment and L2 writing, including Exploring Language Assessment and Testing (Routledge). His research interests involve the relationships between assessment, teaching and learning.

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.

Johnathan Jones is a researcher and lecturer at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire. His work focuses on language assessment, second language listening and speaking, and factors affecting test performance, particularly comprehensibility and fluency. His work also addresses validity, validation, and quantitative approaches to language testing, supporting the development of fair, evidence-based assessments for educational and professional contexts.

Leda Lampropoulou is Head of Research at LANGUAGECERT, where she leads the organisation's research strategy and chairs the Research Committee. With over 15 years' experience in language testing and second language acquisition, she specialises in speaking assessment, test reliability, and benchmarking. She has led concordance studies and large-scale alignment to international

frameworks, supporting principled, data-informed assessment development and decision-making. She holds an MA in Language Testing from Lancaster University and serves on the UKALTA Executive Committee.

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. These include being Senior Advisor to the National Educational Examinations Authority (NEEA) in China as well as a Visiting Professor with CRELLA of the University of Bedfordshire.

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

Volume 4 of Certifying Quality in Assessment and Learning brings together a set of studies positioned at the intersection of innovation, measurement, and comparability in contemporary language assessment. Across its three thematic strands—Innovation & Automation, Measurement, Calibration & Validation, and Comparative Studies & Learning Contexts—the volume reflects a field undergoing rapid transformation, shaped by advances in artificial intelligence and digital delivery while remaining grounded in core principles of validity, reliability, and fairness.

Section 1: Innovation & Automation

The opening section, *Innovation & Automation*, foregrounds the opportunities and constraints associated with integrating emerging technologies into large-scale assessment systems. Chapter 1 sets the conceptual foundation by examining the role of artificial intelligence in international language testing, outlining both its transformative potential and the challenges it introduces for test design and governance. Chapter 2 narrows the focus to fairness in the scoring of short-answer listening items, presenting a structured, data-driven methodology for refining scoring keysets. By systematically identifying acceptable response variation, the approach enhances scoring fairness without compromising reliability.

This focus on the interaction between human judgement and automated processes continues in Chapter 3, which reports the first phase of LANGUAGECERT's multi-stage validation programme for its automated writing scoring system. Drawing on large-scale operational data, the study demonstrates strong alignment between automated and human scores and provides evidence of stable, construct-relevant scoring behaviour. Chapter 4 extends the discussion of automation into the predictive domain, exploring the use of large language models to simulate test-taker performance and estimate item difficulty prior to live administration. While findings suggest that aggregate test behaviour can be approximated with reasonable accuracy, variability at the individual item level highlights the current limitations of such approaches.

Section 2: Measurement, Calibration & Validation

The second section, *Measurement, Calibration & Validation*, turns to the evidential basis underpinning assessment systems. Chapter 5 investigates the relationship between linguistic complexity and test-taker performance in high-stakes writing assessment, providing empirical support for the interpretation of scores across CEFR levels. The results indicate that higher-performing candidates generally produce more complex language, reinforcing the construct validity of the assessment.

Chapter 6 examines the role of expert judgement in establishing assessment scales, comparing expert-derived difficulty estimates with those obtained through pairwise comparison and Rasch calibration. The close alignment between these approaches supports the continued centrality of expert judgement within test development processes. Chapter 7 builds on this focus on calibration by presenting an iterative anchoring procedure for adaptive reading and listening tests. The findings demonstrate how systematic

refinement of anchor items leads to stable and robust measurement frameworks, even within large and evolving item pools.

Section 3: Comparative Studies & Learning Contexts

The final section, *Comparative Studies & Learning Contexts*, situates LANGUAGECERT assessments within a broader ecosystem of tests and learning environments. Chapter 8 presents a construct-based comparison between LANGUAGECERT Academic and IELTS Academic, identifying substantial alignment in the underlying abilities assessed while also noting differences in task design and scoring practices that may influence performance. Chapter 9 complements this analysis with a statistical concordance study, demonstrating strong correlations between the two tests and providing equipercentile linking functions to support score interpretation across scales.

The volume concludes with Chapter 10, which examines a mobile-based, CEFR-aligned learning and assessment system designed around bite-sized formative testing. Findings from a large-scale pilot indicate strong learner engagement, measurable proficiency gains, and positive user perceptions, pointing to the potential of digitally mediated, formative assessment to support language development across diverse contexts.

Taken together, the studies in this volume reflect a central dynamic within contemporary language assessment: the integration of technological innovation with rigorous measurement practice. Across automated scoring, simulation-based modelling, calibration methodologies, and cross-test comparability, the contributions collectively underscore that quality in assessment is an ongoing process of empirical validation, methodological refinement, and principled adaptation.

SECTION 1: CALIBRATION/ VALIDATION STUDIES

Chapter 1: AI in Large-Scale International Language Testing: Challenges and Opportunities

Michael Milanovic

Abstract

The rapid proliferation of artificial intelligence systems capable of generating human-like text presents unprecedented challenges and opportunities for large-scale international language assessment. This paper examines the implications of AI for high-stakes language testing used in university admissions, employment, and migration contexts, with particular attention to validity, fairness, and security. It argues that AI is not merely a technological disruption to test delivery but a fundamental challenge to construct definition itself, as AI-mediated communication becomes increasingly normalised in professional and academic domains. Traditional models of communicative language ability, premised on individual cognitive and linguistic competence, may require reconceptualisation to accommodate emergent human-AI collaborative competencies. The paper proposes a framework distinguishing core human language competencies from higher-order collaborative intelligence, suggesting that different stakeholder groups may prioritise these

constructs differently. Fairness concerns arising from inequitable access to AI tools are also examined, alongside evolving security threats including item harvesting and deepfake-based fraud. Promising adaptation strategies are identified, including AI-generated item banks, process-oriented assessment, and technology-integrated task design. The paper concludes by outlining three potential paradigms for the future of language assessment and calls for systematic empirical research to guide the field's response to accelerating technological change.

Keywords: AI in language assessment, human–AI communicative competence

Background

In this paper, I'm going to be exploring the impact of AI on language assessment. My focus will be on international exams for study, employment and migration although many of my observations may be more generally relevant. For the most part, I'll be talking about AI systems that affect language use such as ChatGPT, Deepseek, Claude etc.

The language testing landscape is changing continuously and now is a particularly interesting time as we see how AI is shaping up to have a big impact. Back in 1982 when I first came to China we were in a paper-based, multiple-choice, classical test analysis world, a tradition developed in the 1950s and still alive and well today in many contexts. Communicative language testing was just starting to have an impact and ELTS, developed in the 1970s (the mother of IELTS), became the best-known communicative language test. In the 1980s, while TOEFL was testing hundreds of thousands of candidates a year worldwide, ELTS was testing around ten thousand a year and in its 15-year life only two test forms were ever used. Researchers were looking at construct validity (the extent to which a test is measuring the language skills it is supposed to measure), worrying about the unitary competence hypothesis and focusing on test reliability.

In the 1990s, everyone was modernising, developing and improving item banking systems, calibrating test items and test forms using Item Response

Theory (IRT). People were looking at how tests could be delivered by computer but generally in a local PC context as the internet was so unstable and restricted. In 1997, TOEFL launched its first CB test. In the international context, high stakes testing for study and migration dominated the language testing world. Test delivery was starting to become a global industry particularly in China and India, reflecting the rapid growth of these economies and many others around the world. From an academic perspective, the 1990s represented a period of consolidation and theoretical development in language testing, moving away from largely psychometric approaches towards more holistic, communicative, and socially conscious testing practices. Many contemporary concerns in language assessment today – particularly around validity, authenticity, and the social impact of testing – were developed in the 1990s.

In China, the 1990s were a time of modernisation of English language testing methods. A greater focus on communicative language ability became a priority with the development and introduction of the Public English Language Testing System (PETS). This was a joint project funded by the governments of China and the UK between 1997 and 2000. At the time there were 11 English language tests in general use in China and the PETS five level system, drawing heavily on the communicative language testing movement and the newly developed Common European Framework of Reference (CEFR). PETS was intended to replace this set of tests with a single five-level system; I think, however, that we ended up with 16 tests as opposed to 11. I was quite heavily involved in the development of the CEFR as well as the PETS tests.

The next decade saw a massive increase in international assessment both in national educations systems, where English became the most widely taught language in the world and most used in international study and migration. Whereas, in 2000, less than a million people were taking IELTS and TOEFL combined, this number rose to around five million by 2010. Much greater focus turned to managing the rapidly increasing test candidature, to test security, cheating, item harvesting, cramming and the impact this had on test validity and integrity. TOEFL edged towards fully automated CB assessment while IELTS continued to focus on authenticity and positive impact and

remained paper-based for the most part. It is not clear how much this difference in focus made to test takers but during this decade we saw IELTS overtake TOEFL and become the most widely used test for international study and migration. Technology played a role in test delivery and test security. An excellent example of this is China's massive network of CCTV monitoring of high stakes tests.

In the background, from an academic perspective, testing was starting to migrate to computer-based delivery. Most existing pencil and paper test tasks were adapted for computer-based use and there was a focus on the automated marking of writing and speaking. The language testing field continued to professionalise with a particular focus on assessment literacy and standards. Frameworks such as the CEFR were introduced and there was a focus on validity and validation research that continued to build on Messick's (1989) validity framework and Kane's (2013) argument-based approaches.

The first major disruption in the decade came with the introduction of the Pearson Test of English (PTE) in 2009. PTE was the first fully automated test to be used in the context of international study and migration. Its innovative impact was not really on test materials, which remained somewhat traditional, if not rather old fashioned, but rather on the fact that for the first time, the human being was removed entirely from the equation when it came to marking across all four skills. It is not clear that the test was very well received amongst teachers and students and in terms of candidature, for most of the next decade, it remained a distant third behind IELTS and TOEFL. However, it became increasingly popular in countries such as Australia and India as we approached 2020 because of its availability and accessibility. While its delivery remained anchored in the traditional high stakes test centre network context, it could be made available several times a day in places where demand was high.

In the 2020s, there have been significant developments, and the pace of change is accelerating. Firstly, the concept of testing at home and being remotely proctored was widely adopted in several contexts. This was driven largely by the pandemic. LANGUAGECERT pioneered online proctored tests (OLP), taken by thousands worldwide, including in China. Secondly, we have

seen AI grow from relatively simple operations, largely inconsequential to language testing, to systems which challenge the very fabric of language assessment.

International testing organisations collectively administer millions of high-stakes international language tests annually across diverse global contexts - high stakes tests that determine university admissions, immigration opportunities, and professional certification, and there is a growing pressure to offer these tests in an online proctored (OLP) environment. LANGUAGECERT was the first organisation to run secure high stakes OLP language tests in 2019, and you just have to look at the subsequent introduction of remotely proctored online tests by Pearson (Express), ETS (Home), IDP (Envoy), the British Council (Aptis Remote), and others to see how much of a growing trend OLP is.

AI technologies have the potential for negative impact as they present significant challenges to testing systems, while at the same time also offering innovation opportunities. They represent a technological disruption that goes beyond previous innovations in computer-based testing. Such innovations were relatively low key and revolved largely around delivery. Item types in listening, reading and writing didn't really change much and when it came to speaking, assessments focused on pronunciation, repetition, read aloud and so on rather than the communicative use of the language. If anything, use of technology has typically played a restrictive role in testing because it couldn't do anything very creative.

AI models on the other hand, are quite different because they can generate human-like text, they can answer comprehension questions with high accuracy, create barely identifiable deep fakes in oral interview contexts and even simulate conversational exchanges. I'm going to look at how AI might impact large-scale international assessments, paying attention to validity, fairness, and security. I'm also going to touch on some potentially relevant adaptation strategies.

Validity Considerations

Large-scale international language testing organisations are required to show that their scores accurately reflect relevant language abilities and support appropriate decision-making by universities, employers, and immigration authorities worldwide. In other words, they need to demonstrate that the tests are measuring what they are supposed to measure: that they have construct validity.

As AI systems like ChatGPT, Deepseek and others become increasingly prevalent, they are starting to change the communicative competence construct. The workplace integration of AI tools is accelerating rapidly: according to Microsoft's (2024) Work Trend Index (taken by 31,000 people across 31 countries), 75% of knowledge workers now use AI at work, with 46% having adopted it within just the past six months. 90% of users report that AI helps them save time, 85% that it helps them focus on their most important work, and 84% that it helps them be more creative.

Similarly, PwC's (2024) *Global Workforce Hopes & Fears Survey* (taken by 56,600 individuals across 50 countries and territories), found that more than 70% of generative AI users agree that AI tools create opportunities to be more creative at work and improve the quality of their work. These findings indicate that AI-mediated communication is becoming normalised in professional contexts, suggesting that traditional constructs will need to evolve as the target language use domain changes fundamentally. This represents a qualitatively different shift from previous technological advances. Electric typewriters may have increased speed and word processors improved spelling accuracy, but AI is fundamentally changing how we work with and manipulate knowledge and language.

What does it mean in the assessment context? Will the assessment of reading comprehension assessment, for example, need to evolve over the coming years from information extraction to include information orchestration?

Until now, reading comprehension has been handled by us, human beings, alone. We can see that AI is changing the way we work with language in a fundamental way. In the traditional reading construct, a reader encounters a

text, processes lexical items, builds mental models, makes inferences, and critically evaluates content, all within an individual's cognitive architecture.

In an AI-enhanced reading construct, a reader works with AI to rapidly process, cross-reference, synthesise, and evaluate information from multiple sources simultaneously. The cognitive load shifts from decoding to directing and critically assessing AI-mediated interpretations. This is clearly a different construct. If our testing processes do not change to take this sort of thing into account, we won't be assessing what people actually do in the tests that we use.

For example, when a test-taker can use AI to generate written responses, the relationship between test performance and the underlying construct becomes tenuous. Have you noticed how people's apparent writing ability has changed in the last year or so? I certainly have. In the past I found myself spending a lot of time correcting what people wrote when they were doing minutes, writing reports and so on. There were always issues with grammar, structure, argumentation, vocabulary etc. Then suddenly, when AI was able to transcribe recorded meetings and write minutes, I stopped having to correct minutes. The language proficiency of the people responsible for the minutes and reports hadn't changed. AI had taken over. Two of the most influential language testing researchers of recent years, Bachman and Palmer's (2010), proposed a model of communicative language ability that emphasises the integration of language knowledge with strategic competence. This becomes difficult to assess accurately when AI generates or substantially enhances responses. Writing presents a complex challenge.

If we accept AI as part of communicative competence, writing ability could evolve in a couple of directions:

1. Writing becomes the ability to effectively prompt, guide, and refine AI output to achieve communicative goals. The testing focus shifts from text generation to text orchestration - knowing how to get AI to produce appropriate content and how to shape it for specific audiences and purposes.

2. Alternatively, we develop multiple constructs: "independent writing ability" (purely human) and "AI-assisted communicative competence" (human-AI

collaboration), each serving different purposes and contexts and each assessed in different ways.

So, we might need at least two assessment levels:

Level 1: Core Human Competencies represent the foundational language abilities that remain necessary regardless of technological context. These include basic linguistic knowledge, fundamental comprehension processes, and core strategic competencies. These competencies align with traditional assessment constructs and remain essential for effective communication in any context and do not change assessment practices.

Level 2: Collaborative Intelligence represents emergent competencies that arise from effective human-AI collaboration. This includes the orchestration of AI capabilities to achieve communicative goals that exceed individual human capacity, the development of new discourse conventions for human-AI interaction, and the creation of hybrid communicative competencies that leverage both human and artificial intelligence. This involves the development of new assessment practices reflecting this emerging new construct.

These two levels require multiple construct definitions serving different purposes. Universities evaluating academic writing readiness might focus primarily on Level 1 competencies, ensuring that students possess the foundational abilities necessary for independent scholarly communication. Employers assessing workplace communication skills might emphasise Level 2 competencies, recognising that professional contexts increasingly involve human-AI collaboration. The distinction echoes that made in the 1980s by Cummins (1979) between BICS (Basic Interpersonal Communicative Skills) and CALP (Cognitive Academic Language Proficiency).

Immigration authorities evaluating language proficiency for social integration might require demonstration of core human competencies while acknowledging that settlers will likely use technology-enhanced communication in their daily lives. Different stakeholders can thus receive different but related information from stratified assessment approaches.

So, the challenge relates to the construct definition itself. Traditional definitions of language proficiency have emphasised individual cognitive and

linguistic capabilities. However, in a world where AI-assisted communication becomes normalised, language proficiency construct itself is likely to need reconsideration.

For large-scale international assessment providers, these challenges suggest a need to reexamine what they are measuring. Bachman and Palmer (2010) outlined how test usefulness depends on several qualities including authenticity, reliability, and construct validity - all dimensions affected by AI use. Their framework reminds us that assessment must reflect the target language use domain, which itself is evolving as AI becomes more integrated into communication practices.

Does the rapid development in AI tools mean that we are on the brink of a completely new way of testing that is not about responding correctly but rather about tracking the process needed to respond correctly and deploying technology to do so? AI's potential to identify and track cognitive processes represents perhaps the most transformative possibility for language assessment. Rather than simply measuring outputs, we could assess the underlying cognitive architecture that generates performance. Understanding processes required to generate performance provides much richer information about likely performance across different contexts than any single performance sample.

Fairness

The global scope of major international language tests makes fairness considerations particularly complex when it comes to AI even though many of these considerations already exist. Large assessment providers operate across diverse economic, technological, and educational contexts from major urban centres with advanced infrastructure to remote regions with limited connectivity. As Kunnan (2018) who has written extensively in his research on the concept of fairness in language testing, has long argued, fairness in language assessment requires providing equal opportunities for all test-takers to demonstrate their abilities.

The integration of AI into test preparation (and potentially test-taking) raises significant fairness concerns across all skill areas. Technological access and literacy are not distributed equally across socioeconomic groups. Technology can either mitigate or exacerbate existing inequalities depending on implementation.

For objectively marked reading and listening tests, AI tools can potentially provide advantages through enhanced test preparation, automated analysis of practice tests, and even real-time assistance in identifying patterns or extracting information. Khalifa and Weir (2009) emphasised how familiarity with test formats and strategies affected reading test performance; AI tools that provide sophisticated practice opportunities could create significant advantages for those with access.

More than twenty years ago, Shohamy (2001) argued in her influential work on the power of tests that language assessments already function as gatekeeping mechanisms with profound social consequences. AI technologies have the potential to make these mechanisms even more opaque and potentially unjust if they are not implemented with careful attention to fairness considerations. This is a particularly foggy scene.

Security

For objectively marked reading and listening tests, security challenges focus on preventing information retrieval and answer sharing. AI tools capable of rapidly retrieving information, recognising patterns in test questions, or even memorising answers from previous administrations pose significant challenges for high-stakes assessments. The implications for objectively marked tests are particularly concerning because these tests often rely on item banks that are reused across administrations. Item harvesting - the systematic collection of test items for distribution or sale – is much easier using AI tools. Such technologies can potentially accelerate the harvesting process by automatically recording, categorising, and storing test content. For international testing organisations, this represents both a security threat and

a financial challenge. If item harvesting is successful without AI, just think what can be done with it!

LANGUAGECERT has developed a number of ways to use AI to address security challenges in both high stakes test centres and OLP. These include enhanced identity verification, response pattern analysis, keystroke monitoring, similarity detection in writing tasks and deep fake detection in speaking tests to mention but a few. The challenges nonetheless continue to increase. It would make sense for language test providers to share their information in this area.

On the other hand, AI has already made real time item generation possible although it needs to be improved but we can envisage a time, in the not-too-distant future, (a year or two at most would not appear unrealistic) where AI creates not only test items, but creates unique test items for each administration making memorisation unfeasible. Similarly, it can create infinite variations of core item types measuring at appropriate difficulty levels. Getting the right difficulty level as well as the right content is very important. Auto generation of test items is already a reality and although imperfect now, with continuously refined prompting it will be in common use relatively soon. Our experience at LANGUAGECERT is relevant in this area and may apply to others in international language assessment. Where 12 months ago auto generation of test items was fully successful only 4% of the time, this year it has risen to around 50%. AI technology used in the production of good items both from the perspective of content and difficulty will be a reality quite soon. It is interesting to consider the effect such progress will have on language assessment.

Adaptation Strategies

Several adaptation strategies are emerging. First, assessment providers are having to reconceptualise what they measure.

The cognitive processes and knowledge structures needed for success in target language environments, increasingly include technology-mediated communication. With objectively marked reading and listening tests for example, this reconceptualisation might focus on higher-order cognitive processes that remain distinctively human. Khalifa and Weir (2009) distinguished between careful reading processes (extracting complete meanings) and expeditious reading processes (quick, efficient information retrieval) - with AI potentially better at the latter.

Test design will need to evolve to incorporate technology explicitly. This requires viewing technology as a component of contemporary language use and hence communicative competence. Just as language tests in the past evolved to incorporate authentic materials and communicative tasks, tests of the future might need to evolve to incorporate AI-mediated communication as part of the assessed construct. For example, we might set writing tasks in an AI simulated environment where the task is not to do the writing per se but to adapt the AI generated output to satisfy relevant communication needs.

Automated test question generation will become widely used, either by allowing for the creation of enormous quantities of test materials or by dynamically creating test items in both computer adaptive and linear contexts we will be able to maintain security (through larger item pools and reduced item exposure) while providing more precise measurement and possibly more tailored assessment.

Future Research

As we navigate these challenges, several research directions emerge and perhaps some of you will investigate these with us.

First, we need systematic investigation of how AI affects test performance across different skill areas and proficiency levels. While theoretical analyses provide important frameworks, empirical research is essential for understanding the real-world impacts of AI on assessment outcomes. Controlled studies comparing AI-assisted and independent performance across different test formats and populations would provide valuable evidence to inform adaptation strategies.

Second, we need validation research for innovative assessment approaches designed to maintain validity in AI-enhanced environments. New item types, integrated task formats, and technology-mediated assessment approaches require validation studies to ensure they effectively measure the intended constructs and provide fair opportunities for diverse test-taker populations.

Third, cross-cultural research on attitudes toward AI in assessment contexts is essential for understanding how different cultural perspectives might influence the acceptance and implementation of various adaptation strategies. International testing organisations serve diverse global populations; understanding cultural variations in technology acceptance and ethical perspectives is crucial for developing appropriate policies. The only thing is, given the pace at which AI technology is moving forward, will we have enough time for validation research, or will AI do the validation research for us?

Conclusion

In this paper, I have outlined how AI might impact large-scale international language assessment across all skill domains. While the challenges are substantial, thoughtful adaptation rather than resistance offers the most promising path forward. The future of international language assessment depends on our ability to evolve while maintaining the core principles that have always guided our field.

This suggests a couple of ways of looking at the future:

Parallel Constructs

We maintain both "traditional language proficiency" (purely human) and "AI-enhanced communicative competence" as separate, valid constructs serving different purposes. Universities might require traditional proficiency for academic writing, while employers might value AI-enhanced competence for workplace communication.

Hierarchical Integration

We develop a model where traditional skills form the foundation that enables effective AI collaboration. You need strong reading comprehension skills to effectively evaluate AI summaries; you need writing ability to craft effective prompts and refine AI output. AI competence becomes an advanced layer built on fundamental skills. The integration of the two levels is then addressed in relation to user needs.

Paradigm Replacement

We fully embrace AI as part of language competence and redesign assessments around human-AI collaborative communication. Traditional skills like independent reading comprehension become as obsolete for assessment as handwriting has largely become for writing assessment.

As we stand at this crossroads, it is clear that international language assessment cannot remain static. Whether we choose to preserve traditional constructs, layer AI competence onto foundational skills, or redesign our frameworks entirely, the decisions we make now will shape the integrity and relevance of language testing for decades to come. The real challenge is three-fold: technical, ethical and educational - ensuring that assessment continues to empower individuals rather than exclude them, that it reflects real-world communicative practice without sacrificing fairness, and that it keeps pace with the speed of technological change without abandoning the rigorous validation that underpins its credibility. If we succeed, AI will not diminish the value of language assessment but instead expand its scope, helping us to capture more fully the complex, evolving ways in which humans use language to learn, to work, and to live together.

References

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice. Oxford University Press. https://doi.org/10.7916/D8CV4HB8

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, interpersonal language skill and bilingual education. Working Papers on Bilingualism, (18), 197-205.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

Khalifa, H., & Weir, C. J. (2009). Examining reading: Research and practice in assessing second language reading. Cambridge University Press.

Kunnan, A. J. (2018). Evaluating language assessments. Routledge. https://doi.org/10.4324/9780203803554

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Microsoft. (2024). 2024 Work Trend Index: AI at work is here. Now comes the hard part. Microsoft Corporation. https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part

PwC. (2024). Global Workforce Hopes & Fears Survey 2024. PwC. https://www.pwc.com/gx/en/issues/workforce/hopes-and-fears.html

Shohamy, E. (2001). The power of tests: A critical perspective on the uses of language tests. Routledge. https://doi.org/10.4324/9781315837970

Chapter 2: Bridging Human and Automated Scoring: Improving Fairness in Short-Answer Listening Items

David Coniam and Leda Lampropoulou

Abstract

This paper presents a methodology for the analysis and subsequent refinement of Short-Answer, open-ended Items (SAIs) in listening assessment after the initial trialling stage, as implemented in LANGUAGECERT's development workflow. Using purpose-designed software, test-taker responses are classified into *Exact Same*, *Similar*, or *Totally Different* matches against an item keyset. The *Similar* category is particularly informative, as it identifies high-scoring test taker responses that were marked as incorrect under the initial keyset, prompting targeted review. The software also reports mean item scores and inter-rater agreement (Cohen's Kappa, κ), which give an indication as to whether keyset amendment or item revision is warranted. The approach is illustrated with two SAI examples from a trial listening test, showing how analysis led to specific adjustments. The structured review process enhances fairness while maintaining reliability.

Keywords: listening, short-answer items, fairness, keyset

Introduction

Listening tests typically combine machine-scorable formats with constructed-response tasks to balance practicality and construct coverage (Ockey, 2020). These fall under limited-response formats, which constrain length or form of the answer. Limited-response includes (a) selected-response items (e.g., multiple choice, true–false, multiple matching) and (b) constructed-response items (e.g., gap-fill, short answer). Within the latter category, Short-Answer, open-ended Items (SAIs) require brief responses, typically one to three words. Like other limited-response formats, SAIs are operationally scored dichotomously (correct/incorrect). In this study, the analyses additionally classify responses as *Exact Same*, *Similar*, or *Totally Different* to inform potential post-trial keyset revisions, after which scoring remains dichotomous.

While much item analysis research focuses on multiple-choice formats, where distractor performance provides rich diagnostic data, constructed-response formats such as SAIs often receive less systematic review. If, post-analysis, item facilities and item discriminations are within acceptable limits, the item is retained; if not, it is simply discarded.

This paper has two purposes. First, it introduces a practical methodology for analysing and refining SAIs after trialling, using a purpose-developed piece of software – the *Short Answer Marker* (SAM). Second, it explores the feasibility of further automating SAI marking within this trialling framework. At present, responses are auto-scored against a keyset and then reviewed by human markers. The present paper shows how analysis of *Similar* responses - valid paraphrases initially marked incorrect - can motivate calibrated keyset expansion, thereby supporting more consistent, fair, and potentially fully automatable scoring.

Previous SAI Research

Research on SAIs spans several adjacent assessment domains, including reading comprehension, vocabulary testing, and subject-matter examinations. Although these contexts differ from listening assessment, they collectively illuminate the psycholinguistic, construct-representation, and scoring challenges associated with short-answer formats.

Several studies demonstrate that SAIs impose greater cognitive and linguistic demands than multiple-choice formats. Akhavan Masoumi and Sadeghi (2020), examining vocabulary assessment, found that multiple-choice items inflated performance relative to SAIs due to cueing, while SAIs better discriminated between learners by requiring recall rather than recognition. Similar findings emerge in reading research: Liao and Lee (2024) reported that high-proficiency readers benefit more from SAIs, as the format draws on metacognitive strategies not activated when selecting from pre-formulated options.

Advances in natural language processing have stimulated interest in automated SAI scoring. Studies in automatic question generation and short-answer evaluation (e.g., Riza et al., 2023; Ziai et al., 2012) highlight both the feasibility and the limitations of machine-scoring systems. Finally, recent work in medical education underscores the value of very-short-answer questions (VSAQs) as an alternative to multiple-choice items. These studies consistently report reduced cueing, improved discrimination, and positive perceptions of authenticity (Sam et al., 2018; Brenner et al., 2024). Although these assessments are not listening-based, their findings reinforce a key principle: when constructed-response formats require concise, information-specific output, the design of the marking scheme, and specifically the completeness of permissible responses, is central to fairness and reliability.

Together, this research establishes the empirical and theoretical rationale for SAIs in listening assessment, while also identifying conditions under which scoring must be carefully validated. The methodology developed in the current paper directly addresses these conditions by offering a structured means to detect and credit legitimate paraphrases during trialling.

Short-Answer Open-Ended Items – Test Quality Features

In evaluating the role of SAIs in listening assessment, three interrelated features of test quality are central: validity, reliability, and fairness (Bachman & Palmer, 1996).

Validity

SAIs can strengthen construct representation by requiring test takers to extract and produce essential information from the spoken input, mirroring the selective, purposeful retrieval involved in authentic listening (see Geranpayeh & Taylor, 2013). This avoids unconstrained paraphrasing while still ensuring that successful performance reflects comprehension rather than option recognition. Research from the adjacent domain of reading indicates that short-answer tasks can elicit processing more akin to real-world comprehension than multiple-choice formats and that they demand higher levels of academic language proficiency to comprehend input and construct appropriate responses (Weigle et al., 2013). Evidence from very-short-answer formats further shows that self-constructed responses test recall, understanding, and application more effectively than MCQs, yield stronger psychometric indices, and reduce opportunities for guessing or option-elimination strategies (Puthiaparampil & Rahman, 2020). In a similar vein, very-short-answer questions (VSAQs) have been found to reduce cueing effects, improve discrimination, and to be judged by students to better reflect real-world practice (Sam et al., 2018). Recent evidence also highlights that students and faculty value constructed-response short-answer formats for their authenticity, deeper learning benefits, and feedback potential, even while acknowledging challenges of feasibility and reproducibility (Brenner et al., 2024). In listening, authenticity is further enhanced when items demand retrieval of key information from spoken input (Buck, 2001).

Reliability

Constructed responses require explicit marking criteria and consistent application across responses. Operationally, SAIs can achieve acceptable reliability when marking schemes are clearly specified and standardisation and monitoring procedures are in place. Evidence from progress testing shows that short-answer formats, often assumed to be less reliable than multiple-choice questions, can in fact yield high reliability when supported by model answers and clear marking instructions (Rademakers et al., 2005). In this study, response classification and accompanying statistics, such as mean item scores and inter-rater agreement (Cohen's κ), provide diagnostic feedback on scoring consistency. These indicators guide targeted keyset adjustments and subsequent re-checks, helping to ensure stable and reproducible scoring outcomes.

Fairness

Following Kunnan's (2000) framework, fairness encompasses equitable access to the construct and minimising construct-irrelevant variance. For SAIs, this includes providing clear instructions and crediting legitimate paraphrase. In listening assessment, fairness also requires attention to delivery factors (e.g., audibility, acoustic conditions) to avoid group-differential effects (Buck, 2001). The procedure outlined here addresses fairness directly by detecting high-ability responses that fall outside the initial keyset and incorporating them, where appropriate, into scoring policy. This ensures that test takers are not disadvantaged by valid but unanticipated paraphrase.

Together, these considerations highlight how SAIs, when analysed systematically and supported by software tools, can enhance the validity, reliability, and fairness of listening assessment.

Developing Short-Answer Open-Ended Items

Most limited-response item types are primarily receptive: test takers demonstrate comprehension by selecting among given alternatives. SAIs differ in requiring test takers to produce a response, rendering them more productive in nature and, consequently, more communicative, as they engage real-world language processing (Buck & Tatsuoka, 1998). At LANGUAGECERT, SAIs form a dedicated section of all listening tests, designed to enhance their communicative authenticity.

The development of SAIs requires careful attention to item setting, given that short constructed-response formats are less controlled than selected-response tasks and demand particularly meticulous design (Hughes, 2010; Green, 2020). Input prompts must be concise and unambiguous (Alderson et al., 1995), and response length should be constrained through explicit instructions (e.g., "Answer in no more than three words"). To support reliability, items undergo rigorous setting, with a key for each SAI listing a clearly defined and as exhaustive as possible set of acceptable answers developed by expert item writers (Fulcher & Davidson, 2007; Lane et al., 2016).

Field testing is particularly crucial for high-stakes exams, even when conducted with a relatively small but representative sample (see Bachman & Palmer, 1996). Such trials provide initial evidence of item functioning and highlight whether keysets require refinement. As Buck (2001) observes, although short-answer items are relatively easy to construct, producing good ones is far more challenging: items need to be pretested, and poorly functioning items identified and either revised or discarded. For multiple-choice items, post-trial analysis can be relatively straightforward, as option statistics readily indicate problematic distractors. For SAIs, however, refinement is less systematic: identifying viable alternatives often involves time-consuming manual inspection of hundreds of responses, which can make the process unfocused and inconsistent.

The current paper outlines a more structured methodology for SAI analysis, implemented at LANGUAGECERT at the 200-test taker investigation stage. Using purpose-designed software, responses – in particular frequent incorrect answers from high-performing test takers - are systematically compared against the keysets. The software produces item means and inter-rater agreement (Cohen's Kappa), offering diagnostic insights that inform targeted keyset adjustments. This approach provides a more efficient and evidence-based alternative to traditional manual reviews.

Marking SAIs

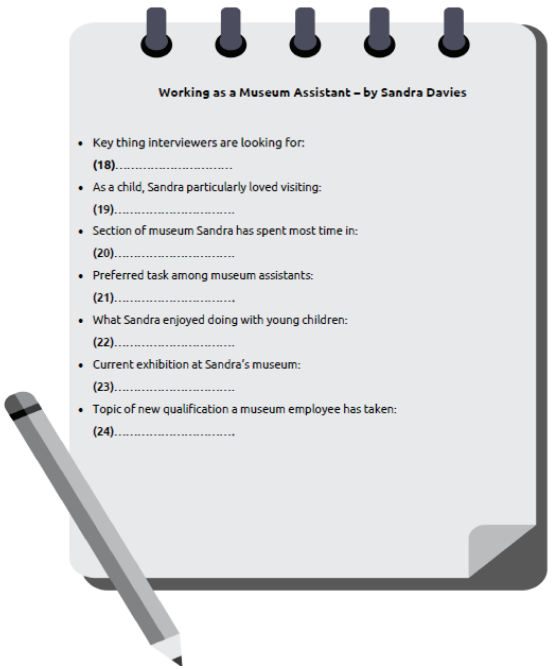
In developing and marking SAIs for a given test, a number of principles are paramount to ensuring accuracy and fairness. First, items must be developed so that a relatively limited set of expected responses can be specified. Second, an extensive keyset should be constructed to encompass as wide a gamut of acceptable responses as possible. Third, as explained earlier, while SAIs are auto-marked against the keyset, all responses are also clerically checked before scores are finalised. Clerical markers are therefore trained and standardised to recognise both the limits of the specified keyset and instances where responses, though not included in the keyset, are clearly acceptable. While the focus is on the content of test taker responses, allowances should be made to accommodate some of the vagaries of English spelling, capitalisation and punctuation. The marking specifications thus define, with as much clarity as possible, what constitutes a satisfactory response.

The LANGUAGECERT (LC) System: the LC Academic and LC General Listening Tests

Against the backdrop of the research outlined above, the current study investigates test taker performance on the SAIs in the Listening Tests of the LC Academic (LCA) and LC General (LCG) tests, the two assessments in the LC System. Although LCA and LCG target different language group domains, they share the same test format for the specific listening task. Consequently, the processes and rationale discussed in this paper apply to both LCA and LCG (LCA/G) listening tests and are used during routine trialling. Each LCA/G listening test comprises 30 items, of which 23 are MC items and seven are SAIs. In the SAI section, test takers need to provide limited responses to complete a prompt. This is usually one or two words, with an allowance of up to three, which is detailed in the task instructions. Each item carries one mark, giving a total possible score of 30. Table 1 presents a sample SAI task from the LCG test, showing the original keyset.

As the table illustrates, expected responses range from two to three words. Some items include bracketed variations, resulting in a relatively extensive set of acceptable alternatives, though not all are required for a correct response.

Table 1: *Sample Listening question paper and original keyset for short-answer open-ended items*

Question Paper	Keyset
Listening Part 3 You will hear someone talking. Complete the information on the notepad. Write short answers of **one to thre** words. You will hear the presentation twice. You have 30 seconds to look at the notepad.  **Working as a Museum Assistant – by Sandra Davies** • Key thing interviewers are looking for: **(18)**..... • As a child, Sandra particularly loved visiting: **(19)**..... • Section of museum Sandra has spent most time in: **(20)**..... • Preferred task among museum assistants: **(21)**..... • What Sandra enjoyed doing with young children: **(22)**..... • Current exhibition at Sandra's museum: **(23)**..... • Topic of new qualification a museum employee has taken: **(24)**.....	**PART 3** 18. (an) interest in knowledge 19. art galleries 20. (the) local history (section/one) 21. making (new) displays 22. (organising) (a) treasure hunt 23. (the) world maps (exhibition) 24. (diploma in) visitor services

Note. It should be noted that the item in Table 1 was selected for its usefulness in illustrating the analysis process, rather than for being representative of average item functioning. Its characteristics allow the SAM outputs, keyset structure, and potential omissions to be demonstrated more transparently than a typical item.

Table 2 shows the same keyset extended into every possible response that a human marker could consider, expanding implicit variants to explicit marking keys.

Table 2: *Keyset made explicit*

Item	Original	Made explicit
18	(an) interest in knowledge	interest in knowledge an interest in knowledge
19	art galleries	art galleries
20	(the) local history (section-one)	local history the local history local history section the local history section local history one the local history one
21	making (new) displays	making displays making new displays
22	(organising) (a) treasure hunt	treasure hunt a treasure hunt organising a treasure hunt organising treasure hunt
23	(the) world maps (exhibition)	world maps the world maps exhibition world maps exhibition the world maps
24	(diploma in) visitor services	visitor services diploma in visitor services

The items show a degree of diversity in terms of what responses may be accepted as correct: item 19 has only one valid response, whereas item 20 allows for six alternatives. As mentioned above, human markers are trained and standardised such that they accept capitalisation and generally ignore punctuation such as apostrophes or dashes, unless these are critical to the answer.

Research Questions

The current paper addresses two research questions:

1. How comparable are human and computer markings of SAIs when both use the same restricted keyset?
2. How can discrepancies between human and computer markings be reduced by incorporating selected responses by high-scoring test takers into the keyset?

The Study

In this section, the trialling process as applied to one dataset is illustrated: the SAI data are analysed using both classical item statistics and the Short Answer Marker (SAM) software. All items were auto-marked and subsequently clerically checked. A dataset of 200 randomly selected test takers was compiled, with a view to emulating recommended practice for pretesting samples (Buck, 2001).

All LC objective tests undergo classical item analysis, reporting, amongst other statistics, item facility (IF) and item discrimination (ID, measured by the point–biserial correlation). For the present LCG listening test, items were coded dichotomously (correct/incorrect). Table 3 presents the analysis for the 200 test takers under examination, for the seven SAIs in the listening task described in Table 2.

Table 3: *Item Analysis for sample*

Item	18	19	20	21	22	23	24
IF	0.41	0.29	0.59	0.39	0.34	0.54	0.23
ID	0.6	0.44	0.53	0.65	0.51	0.6	0.39

The results indicate that the items show acceptable discrimination values overall, although their facilities vary. Falvey et al. (1994) define good item facility as being between 0.30 and 0.80, and good item discrimination as being above 0.30. Items 20, 21, and 23 fall comfortably within the desired facility range, with the remaining items performing less strongly. Item 19 (IF = 0.29) sits at the borderline of acceptable facility, while item 24 (IF = 0.23) would generally be considered too difficult.

In addition to these classical indices, the SAM software provides a more qualitative perspective by categorising test-taker responses into three sections: *Exact Same, Similar*, or *Totally Different*. SAM first marks each item as correct or incorrect against the keyset supplied (as in Table 2 above, for example). Correct answers matching the keyset fall under Exact Same (as in the keyset options, that is). Incorrect responses are further classified as *Similar* or *Totally Different*.

In this analysis, *Exact Same* and *Totally Different* categories are ignored. *Exact Same* are correct; *Totally Different* responses can be disregarded since explorations have shown that responses in this category are indeed totally dissimilar to any of the keys.

It is the *Similar* category which provides the key input for consideration, because it most often contains plausible variants that the mark scheme omitted. Two types of analysis are produced: (a) responses sorted by score, and (b) frequencies of all *Similar* responses, with special attention to those from high-scoring test takers (i.e., those scoring 25 or above out of 30, the maximum possible score on the listening test). High-scoring test takers' responses are highlighted since it is this set of test takers who are most likely to provide possible acceptable responses that merit consideration for keyset

expansion. Tarrant and Ware (2008) state –in the context of MC items, though the argument generalises for most objective test types– that higher-achieving test takers appear more likely than borderline ones to be disadvantaged by flawed items.

The most frequent responses of high-scoring test takers (although the two categories are not necessarily the same) allow the assessment development team to identify recurring patterns that may signal the need for keyset refinement.

Table 4 presents the SAM analysis for the seven items alongside human marking. The human marking is in Row 2, with SAM markings beneath. Two statistics are provided to aid in evaluating which items may need attention: item means and Cohen's Kappa to assess agreement between the computer- and human-marked versions of the test (see Landis & Koch, 1977).

Items 22 and 24 are highlighted since these two items are explored in greater detail due to the larger discrepancies between human-correct and SAM-*Exact Same* means.

Table 4: *Exact Same and Similar responses analysed*

Category	18	19	20	21	22	23	24
Human marking as *Correct*	0.41	0.29	0.59	0.39	0.34	0.54	0.23
SAM *Exact Same*	0.39	0.28	0.57	0.37	0.22	0.54	0.18
SAM *Similar*	0.32	0.29	0.06	0.08	0.40	0.12	0.42
SAM *Totally Different*	0.30	0.44	0.38	0.56	0.39	0.35	0.41

As can be seen, with the exceptions of items 22 and 24, differences between human and SAM Correct means are minimal (≤2 percentage points). Discrepancies occur because humans may: (a) disregard non-critical punctuation such as hyphens or apostrophes, (b) accept unlisted but plausible responses based on training, or (c) overlook typing or orthographic anomalies

(e.g., stray symbols or spelling variants) that the system flags as incorrect. In nearly all cases, SAM's mean is slightly lower than the human mean, reflecting its stricter adherence to the original keyset.
While small discrepancies (2–3%) may be an acceptable margin of error, items 22 and 24 showed notably larger differences: 12% and 5% respectively. These gaps are one indicator that the keysets for these items may be incomplete.

To further assess agreement, Cohen's Kappa was calculated for each item. The results are presented in Table 5.

Table 5: *Cohen's Kappa for human vs. SAM marking*

Item	18	19	20	21	22	23	24
Kappa	0.95	0.96	0.94	0.96	0.70	0.98	0.83

In Kappa analyses, values of 0.8 and above indicate high agreement (Landis & Koch, 1977). While most items meet this threshold, including item 24, item 22 falls below with $\kappa = 0.70$. Given the high-stakes context, a cutoff stricter than 0.8 needs to be considered. This issue is revisited in the Discussion section.

Amending the Keyset

With item means and Kappa values available to inform decision making, attention now turns to an analysis of items 22 and 24, where discrepancies between human and SAM scoring were largest. The analysis focuses on responses in the *Similar* category, examined in two ways: (a) responses produced by high-scoring test takers, and (b) the most frequent *Similar* responses overall.

Similar Responses

SAM identifies *Similar* responses according to three criteria: (a) Contains the Whole Key, (b) Contains Part of the Key, and (c) *Levenshtein Distance*, the latter measuring the minimal number of edits needed to transform one string into another (Miller et al., 2009). Each *Similar* response is accompanied by the test-

taker's total score, and those scoring 25 or above (out of 30) are highlighted, as higher-ability test takers are more likely to produce valid but initially unanticipated variants (Tarrant & Ware, 2008).

Tables 6 and 7 present unedited examples of *Similar* responses for items 22 and 24. Spelling, punctuation, and case reflect the raw test-taker data. These examples illustrate how SAM surfaces potential omissions or ambiguities in an item's design. Although such issues do not arise frequently, this dataset was selected because it provides a clear demonstration of how the procedure operates; in most routine trialling cycles, the analysis confirms that the existing keyset is already sufficiently comprehensive.

Table 6 presents a sample of Similar responses produced for item 22.

Table 6: *Item 22 Similar responses to the key*

Response	Detail	Score
organise a treasure hunt	Contains Part of the Key	29
organizing a treasure hunt	Contains Part of the Key	29
organising a treasure fund	Contains Part of the Key	28
drawing posters and organised a treasure hunt	Contains Part of the Key	27
Treasure hunts	Contains the Whole Key	27
Organizing a tresure hunt.	Contains Part of the Key	27
drawing posters, treasure hunt	Contains the Whole Key	27
posters and treasure hunts	Contains the Whole Key	27
organizing a treasure hunt	Contains Part of the Key	27
Doing a treasure hunt dressed as a gorilla	Contains the Whole Key	26
Tresure hunt	Levenshtein Distance	19
treature hunt	Levenshtein Distance	19
organising something	Contains Part of the Key	19

These responses reveal two systematic patterns: (a) plural forms such as treasure hunts occur frequently among high scorers, (b) test takers needed to use more than three words, and (c) SAM's grouping makes these patterns easy to detect during trialling. Per LANGUAGECERT policy, both American and British English spellings (e.g., organise/organize) are always acceptable and recognised as correct.

Table 7 shows *Similar* responses produced for item 24. Only responses with scores of 25 or more have been included in the table.

Table 7: *Item 24 Similar responses to the key: (diploma in) visitor services*

Response	Details	Score
diploma of visiting services	Contains Part of the Key	29
diploma in visitors' services	Contains Part of the Key	29
diploma of visitor services	Contains Part of the Key	29
diploma in visitors services	Contains Part of the Key	28
deploying visitors' services	Contains Part of the Key	28
diploma in visitors' services	Contains Part of the Key	28
Diploma in visiting services	Contains Part of the Key	28
a diploma in visiting services	Contains Part of the Key	28
diploma in visiting services	Contains Part of the Key	28
a diploma in visitor services	Contains the Whole Key	28
diploma in visiting services	Contains Part of the Key	27
Diploma in visiting services	Contains Part of the Key	27
Diploma on visitor services.	Contains Part of the Key	27
diploma in visitor's services	Contains Part of the Key	27
diplomating visitors services	Contains Part of the Key	27
vistor services	Levenshtein Distance	27
diploma in visiting services	Contains Part of the Key	27
diploma in visitor's services	Contains Part of the Key	27
diploma	Contains Part of the Key	26

Response	Details	Score
Diploma in Visiting Services	Contains Part of the Key	26
diploma in visitor's services	Contains Part of the Key	26
diploma in digital services	Contains Part of the Key	26
diploma of visitor services	Contains Part of the Key	26
diploma of visiting services	Contains Part of the Key	26
diploma of visitor services	Contains Part of the Key	25
diploma in visiting services	Contains Part of the Key	25
diploma in museum service	Contains Part of the Key	25

High-scoring test takers often produced variants such as diploma of visitor services or diploma in visitors' services, which reflect attested usage in English. Table 8 summarises these variants.

Table 8: *Possible acceptable responses*

diploma of visiting services
diploma in visitors' services
diploma of visitor services
diploma in visitors services

Some of responses above exceed the instructed response-length limit. Under operational marking rules, any response exceeding the stated limit is automatically incorrect, just as selecting multiple answers in a single-response MC item results in an incorrect score. These longer variants are therefore not candidates for inclusion in the keyset. They are presented here only because SAM identifies them as structurally related to the keyed answer, making them useful for illustrating how the analytical process operates.

When such patterns emerge in trial data, the implication is not that these responses should be credited but that the item itself may warrant revision. In routine development this may prompt the assessment team to adjust the stem

or rewrite the item to ensure that the intended response can be produced within the stated word limit and without ambiguity.

Most Frequent Responses

During trialling at LANGUAGECERT, *Similar*-category outputs are routinely reviewed both to confirm keyset completeness and to identify any plausible implicit variants. In addition to high-scorer responses, SAM also outputs the most frequent *Similar* responses across the sample. Given the relatively small dataset (N = 200), recurrence is limited but nonetheless informative. Tables 9 and 10 present detail on items 22 and 24 respectively.

Table 9: *Item 22, most frequent Similar responses*

Response	N
treasure hunts	6
organizing treasure hunts	6
posters, treasure hunt	2
treature hunt	2

Table 10: *Item 24, most frequent Similar responses*

Response	N
diploma in visiting services	6
diploma	5
diploma of visiting services	4
diploma of visitor services	3
diploma in visitor's services	3
diploma in visitors' services	2
diploma in visitors	2
diploma in visitor service	2

For item 22 (Table 9), the dominant *Similar* response was "treasure hunts", accompanied by longer variants such as *organizing a treasure hunt*. The latter shows show that test takers understood the intended meaning but were unable to express it within the one- to three-word limit specified for the task. Because variants such as organising a treasure hunt exceed the response-length requirement, they cannot be credited and therefore are not candidates for inclusion in the keyset. Instead, this pattern indicates that the original stem ("What Sandra enjoyed doing with young children") may not have sufficiently constrained the form of the expected answer. A more targeted stem, such as "What Sandra enjoyed organising with young children", would guide test takers more directly toward the intended response ("treasure hunt") while keeping it within the allowed word limit.

For item 24 (Table 10), frequent alternatives such as "diploma of visiting services" and "diploma in visitors' services" illustrate how SAM surfaces structurally related variants, but these examples also show that recurring discrepancies may point to issues with item wording rather than to omissions in the keyset. As with item 22, such findings may prompt item revision rather than key expansion. Ultimately, decisions regarding whether to adjust the keyset or update the item stem lie with the assessment development team.

Extending the Keyset

Drawing on these analyses, as per LANGUAGECERT's routine trialling procedures, the results for items 22 and 24 were shared with the Assessment Development team. The purpose of this stage is to review the performance of the items and to consider whether the stem should be revised or whether the keyset should be amended or extended to incorporate additional acceptable variants that fall within the task constraints. Only after this review are any changes finalised.

For the purposes of this study, a revised keyset was produced for items 22 and 24 (Table 11), which was then used exclusively to re-mark the pilot performances; it was not applied to any live operational data.

Table 11: *Original and revised keysets for items 22 and 24*

Item	Original keysets	Revised keysets
22	organising a treasure hunt organising treasure hunt treasure hunt a treasure hunt	organising a treasure hunt organising treasure hunt organising treasure hunts treasure hunt a treasure hunt treasure hunts
24	diploma in visitor services visitor services	diploma in visitor services visitor services diploma in visitors services diploma of visitor services diploma in visitor's services diploma in visitors' services

Using the revised keysets, all 200 responses were re-marked, and item means and Kappas recalculated. This procedure aligns with the LC trialling policy that test takers should receive credit for any acceptable response (Buck, 2001). Item means and Kappas are presented in Table 12.

Table 12: *Means and Kappas for items 22 and 24 after second marking*

	1st marking	2nd marking	1st marking	2nd marking
Item	22	22	24	24
Human marking	0.34	0.34	0.24	0.24
SAM marking	0.22	0.32	0.18	0.23
Cohen's Kappa	0.7	0.87	0.83	0.83

As can be seen from Table 12, the revised keysets substantially improved alignment between SAM and human marking. For item 22, SAM's mean increased by ten percentage points, reducing the discrepancy with human marking to two points, and Kappa rose from 0.70 to 0.87. For item 24, SAM's mean improved by five points, leaving only a one-point gap with human marking, though Kappa remained unchanged at 0.83.

Discussion

The current study has examined how short-answer open-ended items (SAIs) may be fruitfully and systematically analysed and refined during the test-trialling window. The findings illustrate that if assessment managers receive targeted feedback after the administration of a trial test with SAIs, it becomes possible to revise the markscheme of permissible responses or, where necessary, the items themselves. Incorporating this process promotes fairness for test takers, particularly high performers whose legitimate responses may otherwise be excluded, while also enhancing the reliability of automated scoring. This SAM-supported review is used by LANGUAGECERT during routine trialling of new listening items to calibrate items and keysets before live administration.

The SAM software proved effective in supporting this analysis by (a) categorising responses into *Exact Same, Similar*, and *Totally Different* groups, and (b) providing item means and Kappa coefficients to compare human and computer marking. Together, these outputs supply assessment developers with both quantitative and qualitative evidence for decision making. This echoes findings from medical education, where progress testing with short-answer questions achieves high reliability when supported by examples and explicit marking instructions, challenging the assumption that open-ended formats are inherently less reliable than multiple choice (Rademakers et al., 2005).

As an internal working standard, if item means between human and computer marking differ by no more than three percentage points and Kappa exceeds 0.8, the keyset may be considered robust and no further action required. Conversely, if means diverge by more than three points and Kappa falls below 0.8, the keyset and, where appropriate, the item wording, should be re-examined. In such cases, SAM's analyses of *Similar* responses, sorted by high-scoring test takers and by frequency, provide structured insights that can inform this review. The resulting findings are then shared with the Assessment Development team, who determine whether to extend the keyset, revise the stem, or take no further action. Once revisions are agreed, rerunning SAM on

the trial dataset allows developers to verify whether agreement improves to acceptable levels.

In this way, the procedure not only refines keysets but also provides a replicable framework for monitoring the fairness and reliability of SAI scoring. By foregrounding responses from high-achieving test takers, the process reduces the risk that valid but unanticipated answers are unfairly penalised, while ensuring that automated marking remains consistent with human judgment. In practice, though, it may also assist in ensuring that the keyset is already sufficiently comprehensive, with revisions only occasionally needed when systematic omissions are identified.

Conclusion

This paper has reported a study into the marking of short answer open-ended items (SAIs). SAIs contribute to the communicative value of listening assessment by requiring test takers to produce brief, targeted responses rather than selecting from pre-formulated options. They have also been suggested to place greater demands on language proficiency, as test takers must comprehend input and construct acceptable responses rather than relying on recognition or elimination strategies (Weigle et al., 2013). Comparable evidence from very-short-answer formats in medical education shows that such items avoid the cueing effects of multiple choice, improve discrimination, and are judged by learners to better reflect real-world practice (Sam et al., 2018). Ensuring that their keysets are comprehensive and fair is therefore essential. LANGUAGECERT incorporates this review into its trialling workflow for new listening tests to ensure fairness, communicative authenticity, and alignment between human and automated scoring.

By applying the Short Answer Marker (SAM) software, response data from high-scoring test takers and frequent alternative variants can inform targeted keyset review. This process reduces discrepancies between human and automated scoring, improves reliability, and safeguards fairness for test takers whose valid paraphrases may otherwise be overlooked. In most trialling cycles, however, this procedure primarily serves to confirm the accuracy and

completeness of the keyset, with amendments only required in the case of systematic omissions.

While the analysis focused on the LANGUAGECERT General test, such analysis and subsequent amendments should be considered by any examinations body where fairness is a guiding principle. The process described here, supported by purpose-built software, enables the empirical refinement of keysets and, where necessary, revision of items. This is particularly important for test takers whose responses may be valid yet initially unaccounted for. By incorporating this structured trialling process, LC ensures that its assessments maintain high standards of validity and reliability. Additionally, during live testing, clerical markers check all SAIs, following strict guidelines so that acceptable responses are always marked as correct.

Analyses such as these conducted in this study serve a number of purposes: first, they inform test designers as to the type of almost-correct-yet-incorrect answers provided by high-scoring test takers, that should be considered for keyset expansion. Second, they provide insights to test takers into things they should be mindful of when completing limited-response items, what to aim for and the kind of errors to avoid.

Beyond its application in the LCA/G Listening Tests, the methodology outlined in the current paper is applicable across a variety of assessments that incorporate SAIs as a test type and can have broader implications for the field of language assessment. As technology-driven marking systems become increasingly prevalent, integrating a structured review process in the trialling phase, as demonstrated in this study, can bridge the gap between human judgment and automated precision. This not only benefits large-scale test providers but also offers a viable framework for smaller examination boards that may lack the resources for extensive pre-testing. By incorporating such methodologies into their marking and quality assurance processes, providers can improve the fairness and accuracy of their own assessments, ultimately strengthening the credibility of their programmes.

Ultimately, as the landscape of language assessment continues to evolve, stakeholders, whether global or local, must recognise both the opportunities and the challenges of using short-answer formats, and commit to principled,

evidence-based practice. Embedding data-driven review in the trialling of SAIs, as outlined here, offers a replicable framework for refining keysets, thereby promoting fairness, reliability, and transparency in language testing.

References

Akhavan Masoumi, G., & Sadeghi, K. (2020). Impact of test format on vocabulary test performance of EFL learners: The Role of Gender. *Language Testing in Asia, 10(1),* 2.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice: *Designing and developing useful language tests*. Oxford University Press.

Brenner, J. M., Fulton, T. B., Kruidering, M., Bird, J. B., Willey, J., Qua, K., & Olvet, D. M. (2024). What have we learned about constructed response short-answer questions from students and faculty? A multi-institutional study. *Medical Teacher, 46*(3), 349-358.

Buck, G. (2001). *Assessing listening*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157. https://doi.org/10.1177/026553229801500201

Falvey, P., Holbrook, J., & Coniam, D. (1994). *Assessing students.* Longman Asia Limited.

Fulcher, G., & Davidson, F. (2007). Language testing and assessment. Routledge.

Geranpayeh, A., & Taylor, L. (2013). Examining listening: Research and practice in assessing second language listening. *Studies in Language Testing v.35*. Cambridge University Press.

Green, A. (2020). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge.

Hughes, A. (2010). Testing for language teachers. Cambridge University Press. https://doi.org/10.1017/CBO9780511732980

Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. Cambridge University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33(1)*, 159–174. https://doi.org/10.2307/2529310

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). Handbook of test development (Vol. 2, pp. 3-18). Routledge.

Liao, R. J. T., & Lee, K. (2024). The Interplay between Metacognitive Knowledge, L2 Language Proficiency, and Question Formats in Predicting L2 Reading Test Scores. *Education Sciences, 14*(4), 370. https://doi.org/10.3390/educsci14040370

Miller, F. P., Vandome, A. F., & McBrewster, J. (2009). Levenshtein distance: Information theory, computer science, string (computer science), string metric, Damerau?Levenshtein distance, spell checker, hamming distance. Alpha Press.

Ockey, G.J. (2020). Assessment of Listening. In C.A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, https://doi.org/10.1002/9781405198431.wbeal0048.pub2

Puthiaparampil, T., & Rahman, M. M. (2020). Very short answer questions: A viable alternative to multiple choice questions. *BMC Medical Education, 20*, 141. https://doi.org/10.1186/s12909-020-02057-w

Rademakers, J., ten Cate, T. J., & Bär, P. R. (2005). Progress testing with short answer questions. *Medical Teacher, 27*(7), 578–582. https://doi.org/10.1080/01421590500062749

Riza, L. S., Firdaus, Y., Sukamto, R. A., Wahyudin, & Abu Samah, K. A. F. (2023). Automatic generation of short-answer questions in reading comprehension using NLP and KNN. *Multimedia Tools and Applications, 82(27)*, 41913-41940. https://doi.org/10.1007/s11042-023-15191-6

Sam, A. H., Field, S. M., Collares, C. F., van der Vleuten, C. P. M., Wass, V. J., Melville, C., Harris, J., & Meeran, K. (2018). Very-short-answer questions: Reliability, discrimination and acceptability. *Medical Education, 52(4)*, 447–455. https://doi.org/10.1111/medu.13504

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Medical Education, 42(2), 198-206. https://doi.org/10.1111/j.1365-2923.2007.02957.x

Weigle, S. C., Yang, W., & Montee, M. (2013). Exploring reading processes in an academic reading test using short-answer questions. *Language Assessment Quarterly, 10*(1), 28–48. https://doi.org/10.1080/15434303.2012.750660

Ziai, R., Ott, N., & Meurers, D. (2012). Short answer assessment: Establishing links between research strands. Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, 190–200. Association for Computational Linguistics.

Chapter 3: Operational validation evidence for the LANGUAGECERT automated writing scoring system: Phase 1

David Coniam, Leda Lampropoulou and Vlasis Megaritis

Abstract

This paper reports Phase 1 of a multi-stage validation programme for the LANGUAGECERT Automated Writing Scoring System, focusing on its large-scale operational behaviour on the LANGUAGECERT Academic Writing Test (LCAWT). The purpose of this phase is to evaluate whether the automated scoring system operates coherently and reliably under authentic testing conditions, and whether its behaviour aligns sufficiently with trained human examiners to support progression to subsequent validation phases.

The study draws on a dataset of 2,394 test takers, each completing two writing tasks assessed against four analytic criteria: Task Fulfilment, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, and Organisation and Coherence. Automated scores were compared with scores produced by a composite of trained and quality-assured human markers, consistent with

LANGUAGECERT's operational marking procedures. Analyses included descriptive statistics, correlational analyses, and Many-Facet Rasch Measurement modelling to examine score correspondence, marker behaviour, and facet stability across tasks, criteria, and prompts.

Results indicate strong agreement between automated and human scores at the total-score level (Spearman's ρ = 0.87 for both tasks), with similarly high correspondence across most analytic criteria. Rasch analyses showed acceptable fit for marker, criterion, and question facets, and discrepancy rates relative to operational quality-assurance thresholds were low (approximately 1.5% per task), substantially below typical human–human discrepancy levels.

Taken together, the findings demonstrate that the automated scoring system exhibits stable, interpretable, and construct-aligned behaviour under operational conditions. In line with its intended evidential role, Phase 1 establishes a robust operational foundation for subsequent precision- and fairness-focused phases of the validation programme, rather than constituting a standalone or definitive validation of automated scoring performance.

Keywords: LANGUAGECERT Academic Writing Test, Automated Writing Assessment, Human–Machine Score Agreement, Operational Validation

Introduction

Automated scoring systems are increasingly used in high-stakes language assessment, where their deployment must be supported by a coherent body of validation evidence. Such evidence is rarely provided by a single study; rather, it typically accumulates across multiple investigations that address different aspects of system performance under varying conditions.

Within this context, the present paper forms part of a broader validation programme for the LANGUAGECERT Automated Writing Scoring System. The work reported here constitutes Phase 1 of this programme and focuses on the system's large-scale operational behaviour on the LANGUAGECERT Academic Writing Test (LCAWT), using authentic test data and established analytic approaches.

The purpose of this study is to examine whether, under realistic operational conditions, the automated scoring system operates within an acceptable performance range when compared with trained and quality-assured human markers. The focus is on large-scale score behaviour, including stability, consistency, and alignment with the intended writing construct, rather than on fine-grained estimates of human rater variability or subgroup-level effects. As such, the findings are intended to provide foundational operational evidence that informs subsequent stages of validation, rather than a comprehensive evaluation of automated scoring performance in isolation.

The next section outlines the validation framework within which this study is positioned and clarifies the evidential role of the present analyses.

Validation Framework and Study Positioning

The validation of automated scoring systems intended for use in high-stakes language assessment is typically supported by evidence accumulated across multiple, complementary investigations. Different stages of validation address distinct questions, ranging from large-scale operational behaviour to more tightly controlled examinations of scoring precision and fairness.

Within this broader context, the validation of the LANGUAGECERT Automated Writing Scoring System is being undertaken through a staged programme of related studies, each contributing a different form of evidence to the overall validity argument. The programme currently comprises three phases at different stages of development.

Phase 1, reported in the present paper, constitutes a foundational operational study and is now completed. Its purpose was to examine the large-scale behaviour of the automated scoring system when applied to authentic test data and compared with trained and quality-assured human markers. The emphasis at this stage is on score stability, consistency, and construct-aligned behaviour across tasks, criteria, and prompts under realistic scoring conditions.

Phase 2 is a multi-rater study, for which a detailed study plan has been developed. This phase is designed to examine agreement between the automated scoring system and multiple independent human raters, situating automated–human agreement within the broader distribution of human–human variability. Evidence from Phase 1 informed the design parameters of this work, but did not constitute its sole motivation.

Phase 3 extends the validation programme to questions of subgroup behaviour and fairness. This phase examines automated scoring performance consistency across relevant test-taker subgroups and contributes fairness-related evidence required for responsible use in high-stakes assessment contexts.

The present paper (Phase 1 study) should therefore be interpreted in relation to its specific evidential role within this broader programme. It provides foundational operational evidence that supports, but does not replace, subsequent precision- and fairness-focused validation work, and should not be read as a standalone or definitive evaluation of automated scoring performance.

Background

The use of computers to score student essays dates back to the 1960s with the work of Page (1966), whose Project Essay Grade *(PEG)* modelled relationships between surface linguistic features (e.g., sentence length, word count, and punctuation) and human-assigned scores using regression analysis. Although primitive by contemporary standards, PEG represented a critical conceptual leap: that aspects of writing quality could be quantified and predicted algorithmically.

From the 1980s onward, advances in computational capacity and the availability of larger linguistic corpora contributed to increased research activity in computer-based testing and assessment (Chapelle & Douglas, 2006). During this period, the educational potential of computational tools was increasingly recognised within broader Computer-Assisted Language

Learning (CALL) frameworks. Warschauer and Healey (1998), for example, anticipated the emergence of an Intelligent CALL phase in which technology-supported systems would provide adaptive guidance and personalised feedback. Developments in natural language processing (NLP) and artificial intelligence (AI) have since enabled aspects of this vision to be realised within automated writing assessment.

Evolution of Analytical Techniques in Automated Scoring

Throughout the 2000s and 2010s, research and commercial applications of automated writing assessment (AWA) systems expanded substantially. Ramineni and Williamson (2013) documented a range of AWA systems, each employing different combinations of linguistic analysis and statistical modelling. These approaches included rule-based feature modelling (e.g., later iterations of PEG; Page, 2003), semantic vector methods such as Latent Semantic Analysis (e.g., the Intelligent Essay Assessor; Foltz, Laham, & Landauer, 1998), and NLP-driven feature extraction approaches used in systems such as e-rater and IntelliMetric (Burstein, 2003).

Validation of AWA systems has typically been framed around multiple dimensions reflecting both psychometric and practical considerations. These dimensions commonly include statistical agreement with human scores, fairness across test-taker groups, relationships with external performance indicators, and the broader consequences of automated scoring use. Table 1 summarises these dimensions as articulated by Ramineni and Williamson (2013), which continue to inform contemporary validation practice.

Table 1: *Criteria for assessing AWA system performance (adapted from Ramineni & Williamson, 2013)*

Dimension	Description
Statistical agreement between computer- and human-derived scores	The degree of correspondence between machine-generated and human-assigned scores, typically measured through correlations, kappa coefficients.
Fairness	The extent to which scores are consistent and unbiased across different groups of test takers, such as those differing by gender or first language.
Relationship with external variables	The strength of association between automated scores and other relevant indicators of performance, such as results on related test components or English class grades.
Consequential validity	The practical and ethical implications of using automated scores in place of human scores.

By the mid-2010s, NLP techniques had become integral to the design of nearly all AWA systems. The growing sophistication of language models enabled systems to move beyond surface feature counting to capture aspects of coherence, cohesion, and rhetorical organisation. Systems such as ETS's E-rater (Burstein, 2003) were used operationally in large-scale assessments such as the TOEFL, functioning either autonomously or within hybrid human–machine marking models. (Attali & Burstein, 2006; Ramineni & Williamson, 2013).

Integration of AI-Based Scoring in Large-Scale English Language Assessment

Over the past decade, the use of AI-based automarking has transitioned from experimental to operational in a number of large-scale English language assessments. Adoption across the sector, however, has been neither uniform nor unconditional. Testing organisations differ markedly in the extent to which automated scoring is integrated into operational decision-making, reflecting varying institutional assessments of risk, accountability, and evidential sufficiency.

In high-stakes contexts, automated scoring is most commonly implemented within hybrid or human-in-the-loop frameworks, in which automated scores are complemented by human oversight and review. Such models are designed to combine the efficiency and consistency of algorithmic scoring with the interpretive judgement, contextual sensitivity, and ethical accountability of trained human examiners.

While advances in AI-based scoring have produced increasingly strong alignment with human ratings under controlled conditions, fully automated and unsupervised deployment remains comparatively limited in high-stakes writing assessment. This reflects ongoing concerns relating to construct representation, edge-case behaviour, and the handling of atypical or severely deficient responses, areas in which human judgement continues to play a critical role. As a result, sector-wide practice tends to emphasise progressive integration rather than wholesale replacement of human scoring.

Importantly, a number of high-stakes English language assessments continue to rely exclusively on human marking for writing, reflecting differing institutional positions on automation and risk management. Across the sector as a whole, however, a consistent emphasis can be observed on reliability, fairness, transparency, and human oversight wherever automated scoring is used. Automated scoring is thus increasingly positioned as a complement to expert human judgement rather than a wholesale replacement.

From NLP to Large Language Models

The emergence of transformer-based Large Language Models (LLMs) such as ChatGPT has marked a transformative phase in automated writing assessment. Whereas earlier systems relied on manually engineered linguistic features and regression modelling, LLMs can infer discourse-level, pragmatic, and rhetorical relationships directly from vast text corpora. This allows for more nuanced evaluation of content relevance, coherence, and stylistic appropriateness.

Recent studies have shown that LLM-based approaches can achieve levels of agreement with human raters comparable to, and in some cases exceeding, earlier NLP-based systems, particularly when guided by explicit rubrics and exemplar material (e.g., Mizumoto & Eguchi, 2023; Xiao et al., 2024). Related work across multiple languages suggests that such models may generalise scoring behaviour beyond English, although concerns regarding bias, construct representation, explainability, and governance remain active areas of research (Kostić et al., 2024; Kwon et al., 2023).

The Role of AI Feedback and Future Directions

Beyond summative assessment, AI-based systems have increasingly been explored for *formative feedback generation*. Prior research has examined the use of automated systems to provide targeted guidance on vocabulary, grammar, and discourse structure (Ramineni & Williamson, 2013; Mansour et al., 2024; Shi & Aryadoust, 2024). This capacity aligns with the long-standing pedagogical vision of *Intelligent CALL* outlined by Warschauer & Healey (1998). While feedback generation lies outside the scope of the present study, it represents a potential extension of automated scoring systems in future work.

LANGUAGECERT Automated Writing Scoring: Analytic and Modelling Framework

The LANGUAGECERT Automated Writing Scoring System adopts a hybrid modelling approach that integrates linguistically motivated features with contemporary deep learning techniques. The design reflects a balance between construct-relevant linguistic analysis and contextual language representation, with the aim of supporting reliable operational scoring while enabling transparent validation. The process is described briefly below.

Scripts identified as unsuitable for automated analysis are excluded at this stage. These include responses that are excessively short (fewer than 85 unique words), blank or score-zero submissions, and texts that exhibit characteristics of non-language output, such as a very low proportion of valid English words or unusually high repetition. The remaining scripts are then partitioned into training and evaluation sets. This filtering step reflects standard operational quality-control practice and is distinct from the validation analyses reported in this study.

A range of readability, lexical, and syntactic metrics is subsequently computed. Readability measures include indices such as Flesch Reading Ease and SMOG, while lexical and syntactic features encompass average sentence length, indicators of vocabulary diversity and complexity, and part-of-speech (POS) n-gram distributions (uni- to tri-gram). These features are selected to align with aspects of writing performance explicitly referenced in the LCAWT rating scale and to support construct-aligned score modelling.

In parallel, contextual representations are derived using Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019). A pre-trained BERT model is fine-tuned on scored writing scripts to produce dense embeddings that capture semantic and stylistic characteristics not readily represented by surface linguistic features alone.

Score prediction is performed using a regression-based learning framework (XGBoost; Chen & Guestrin, 2016). For each analytic criterion, the modelling process allows for different combinations of available feature sets, including contextual embeddings, handcrafted linguistic features, and POS-based representations. In practice, most deployed models incorporate all feature types, though the framework does not require a fixed fusion of components for all criteria. Model performance during development is evaluated using standard predictive indices, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Quadratic Weighted Kappa (QWK), which provide complementary information about average prediction error, sensitivity to large deviations, and agreement with human scores.

The system underwent two major development iterations. Version 1 was trained on data collected up to August 2024, while Version 2 was retrained and re-evaluated using additional data collected between September and December 2024. Table 2 summarises the training and testing configurations for each version. All analyses reported in the present study are based on Version 2 of the model.

Table 2: *Automarker training and testing parameters*

Automarker version	Time frame	Training scripts	Testing scripts
Version 1	Up to Aug 2024	1,838	460
Version 2	Sept-Dec 2024	2,232	558

A hybrid design was selected over an end-to-end generative language model for three primary reasons. First, the inclusion of explicit linguistic features provides interpretable signals aligned with the construct definitions of the LCAWT rating scale, supporting transparent validation work, even where BERT-derived representations typically carry greater predictive weight. Second, the use of contextual embeddings enables modelling of semantic and stylistic nuance while remaining comparatively stable and data-efficient, compared with large generative models. Third, the XGBoost regression-based stage offers predictable behaviour under controlled conditions, reducing the risk of drift, hallucination, or unexplainable variance associated with generative LLM scoring. Taken together, this architecture balances modern contextual

modelling with the operational reliability and controlled interpretability required in high-stakes assessment.

The Current Study

To contextualise the analyses that follow, this section outlines the assessment instrument, dataset, and analytic design used to compare automated and human-generated scores.

The LANGUAGECERT Academic Writing Test

The present study draws on scripts from the LANGUAGECERT Academic Writing Test (LCAWT), a high-stakes, multi-level test designed to assess the ability to understand, use, and communicate effectively in written English for academic purposes. The LCAWT lasts approximately 50 minutes and comprises two tasks: (1) a structured response to visual input (e.g., a graph or table) of 150-200 words, and (2) an argumentative academic essay of approximately 250 words.

Writing performance is assessed using analytic marking criteria aligned with the descriptors of the Common European Framework of Reference for Languages (CEFR). The four criteria are Task Fulfilment (TF), Accuracy and Range of Grammar (GRA), Accuracy and Range of Vocabulary (VRA), and Organisation and Coherence (OC). Performance on each criterion is rated on a nine-point scale (0–8), yielding a maximum composite score of 32. Each task is double-marked by trained human examiners, and the final task score represents the mean of the two ratings. While the LCAWT reports results across CEFR levels A1 to C2, its primary purpose is essentially to indicate readiness for tertiary-level study; hence, most test-taker scores fall between B1 and C1 levels.

Sample

The dataset comprises 2,394 LCAWT test takers, each contributing one script for each of the two writing tasks. Across the dataset, each task was administered using 11 distinct prompts, yielding a total of 22 unique writing prompts. Minor variation in the number of scripts per task reflects routine data-cleaning procedures applied prior to analysis. Five scripts from Task 1 were excluded because they did not meet minimal analysable criteria (e.g. truncated or empty responses). No candidates were removed in full, and the total number of test takers therefore remains unchanged.

Markers

Two sources of scores were compared in the study. Marker 1 represents the finalised human scores produced through LANGUAGECERT's routine operational marking process. These scores were generated by multiple trained and accredited human markers, all of whom marked in accordance with standard procedures. Where applicable, scores were subject to routine moderation and sign-off processes before being finalised. The resulting human scores therefore reflect operationally valid outcomes rather than individual examiner judgements.

Marker 2 refers to the automated writing scoring system described above, which generated criterion-level and total scores for the same scripts. Comparisons in the present study are thus between operationally finalised human scores and automated scores produced independently by the system.

Research Questions

The broad research hypothesis of the study is that the LANGUAGECERT automarker would demonstrate statistical qualities comparable to those reported among human markers, reflecting the intended evidential role of Phase 1.

Specifically:

1. Correlations between automarker and human total scores (out of 32 for the four criteria) will be above 0.80.
2. Rasch fit statistics for marker, criterion and question facets will fall within the range 0.50–1.50.
3. Discrepancies between automarker and human ratings will remain below 10%.

Analysis

This section analyses test-taker writing performance as scored by two sources: the operationally finalised human scores (Marker 1) and the automated scoring system (Marker 2). Analyses are presented in stages, beginning with descriptive statistics, followed by correlational analyses and Many-Facet Rasch Analysis (MFRA).

Descriptive statistics are first reported for both score sources at the criterion level and for total composite scores. Inferential analyses then examine the degree of association between human and automated scores. Given the ordinal nature of the 0–8 rating scales and the focus on association rather than mean differences, Spearman's rank correlation coefficient (ρ) is used as the primary measure of correspondence (Lumley, 2002).

In addition, MFRA is employed to examine score behaviour across multiple facets simultaneously. MFRA has been widely used in performance assessments such as writing to model variability associated with test takers, raters, tasks, and rating criteria (Engelhard, 1992; McNamara, 1996). In the present study, MFRA is used to examine marker behaviour and score consistency across test-taker, task, and criterion facets within a unified measurement framework.

Descriptives

Table 3 presents descriptive statistics for two writing tasks by score source. The maximum possible score for each task is 32.

Table 3: *Descriptive statistics for human and automarker scores*

Task	Task 1			Task 2	
	Human marker	Automarker		Human marker	Automarker
Number	2,389	2,389		2,394	2,394
Mean	18.10	17.65		18.03	17.66
SD	6.24	5.37		6.17	5.35
Minimum	1	4		2	4
Maximum	32	28		32	27

Across both tasks, mean scores produced by the human and automated scoring sources differed by less than one point. Human mean scores were highly consistent across tasks (18.10 and 18.03), while the automarker's mean scores were virtually identical (17.65 and 17.66), indicating stable scoring behaviour across task types.

Correlations

Table 4 presents Spearman's ρ correlations between human and automated scores at the total-score level and for each analytic criterion. Following Hatch and Lazaraton's (1991) descriptors for inter-rater reliability, correlations of 0.80 or above are interpreted as strong, 0.50–0.79 as moderate to strong, and around 0.50 as moderate.

Table 4: *Correlations (Spearman's ρ) between human and automarker scores by task and criterion*

Task	Total	TF	GRA	VRA	O&C
1	0.87	0.81	0.84	0.83	0.75
2	0.87	0.81	0.84	0.84	0.76

At the total-score level, correlations between human and automated scores were strong for both tasks (ρ=0.87). Criterion-level correlations were similarly high, with strong correlations (ρ ≥ 0.80) across three of the four criteria. The slightly lower, though still substantial, correlations on Organisation and Coherence (ρ ≈ 0.75) are consistent with prior findings that discourse-level constructs present greater challenges for automated scoring models.

Interestingly, Task Fulfilment, despite its reliance on prompt-specific information, also demonstrated correlations above 0.80, indicating that the automarker was able to infer topic relevance with reasonable accuracy. Overall, the descriptive and correlational results indicate that the automarker performs comparably to an experienced human marker in terms of scores awarded, range of scores, and correlations between criteria.

Organisation and Coherence correlations (ρ≈0.75) fall below the 0.80 threshold, consistent with prior research on discourse-level constructs. For Phase 1, correlations above 0.70 are acceptable given strong total-score correspondence. However, this relatively weaker performance will be looked at in Phase 2 through multiple independent raters.

Many Facet Rasch Analysis

In performance assessments such as writing tests Many-Facet Rasch Analysis (MFRA) has become a widely accepted approach for modelling multiple sources of variability (e.g., facets for test takers, raters, tasks, and criteria) (Engelhard, 1992; Weigle, 1998, 2002). In MFRA, the measurement scale derived by application of a unified metric such as the Rasch model means that various phenomena – marker severity, question difficulty etc – can be modelled and compensated for (McNamara, 1996; Weir, 2005). In MFRA, the

measurement scale derives from the probability of observed ratings given facet locations; thus, situational factors -here, test-taker ability, question difficulty, and marker severity- are explicitly modelled in constructing the overall measurement picture. The present study specified a four-facet design: markers, test takers, task prompts (labelled 'Questions' in FACETS output), and marking criteria. The human marker facet represents an operationally finalised composite of multiple trained examiners rather than an individual rater, reflecting standard LANGUAGECERT marking and moderation procedures. Analyses were conducted with FACETS v4.1.8 (Linacre, 2024).

MFRA offers advantages over purely classical statistics by calibrating all facets onto a single unidimensional latent scale, enabling direct comparison across marker severity, task difficulty, and criterion behaviour (Eckes, 2015). In line with previous LANGUAGECERT work (e.g., Coniam et al., 2021a; Papargyris & Yan, 2022), fit is reported as a key diagnostic. Fit relates to how well obtained values match expected values and is divisible into related, if slightly different, categories. The most widely used is the infit mean square (MnSq) statistic. Infit values around 1.0 indicate expected variation, with values between approximately 0.5-1.5 commonly taken as acceptable in operational assessment contexts (Lunz & Stahl, 1990). High infit suggests excess, unsystematic variation (misfit); very low infit indicates over-predictability (overfit).

Data preparation

For MFRA purposes, scripts from both tasks were combined into a single dataset and subjected to routine preprocessing prior to calibration. The resulting FACETS analysis comprised 39,240 criterion-level rating observations, distributed across four analytic criteria, two scoring sources, 22 task prompts, and 2,394 test takers. This dataset represents the complete set of valid rater–response interactions retained for multifaceted calibration.

Model Fit and Facet Maps

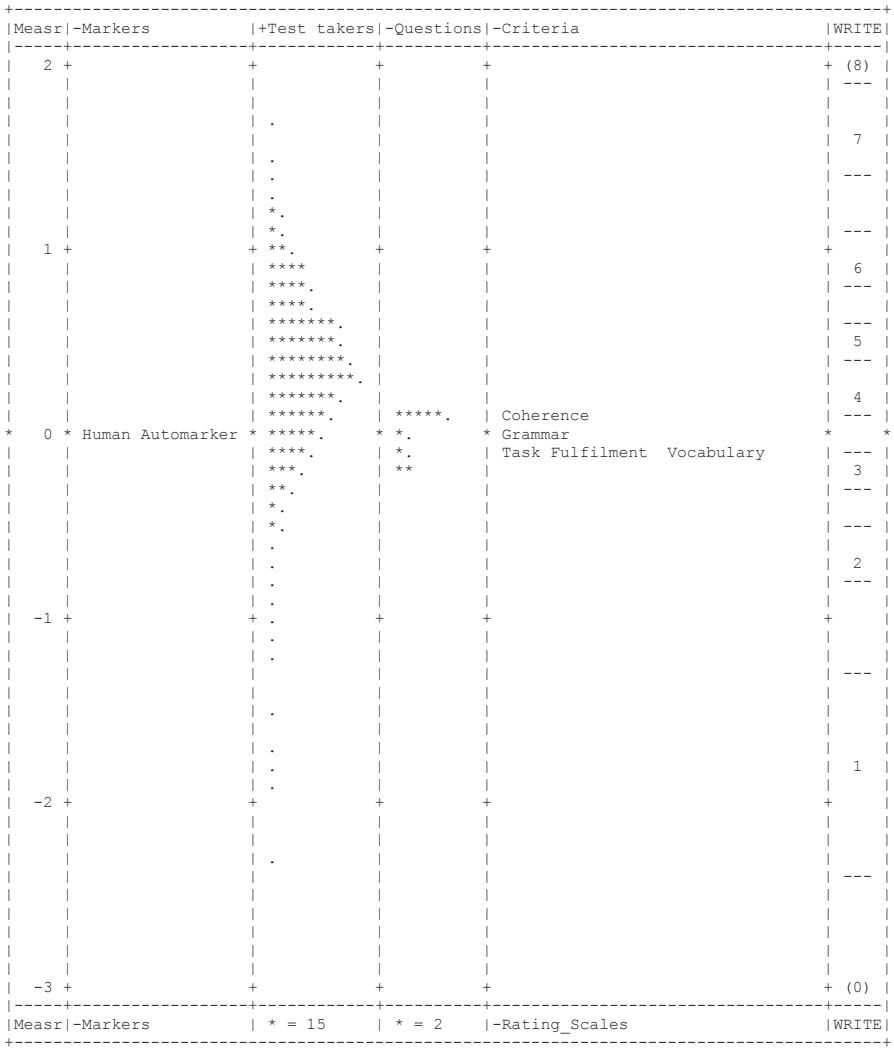
Overall model fit was examined prior to interpretation of individual facets. Following Linacre (2002), satisfactory fit is indicated when no more than approximately 5% of standardised residuals exceed ±2 and no more than 1% exceed ±3. In the present analysis, 39,240 valid responses contributed to parameter estimation. Of these, 21 responses (0.05%) exceeded ±2, and 79 responses (0.20%) exceeded ±3, well within acceptable limits.

As some raw scores included decimal points, all scores were multiplied by 10 to meet FACETS input requirements. This transformation affects the scale of reported scores, as in some cases (total score, observed score, fair average), the output appears 10 times larger than it actually is, however, fit and measure are not affected.

Figure 1 presents the Wright map displaying the relative locations of all facets. Stricter markers are located higher up the ruler; more lenient markers down the ruler. More able test takers are located higher up the scale, less able test takers further down the scale. Marking criteria shown higher are harder (i.e., test takers are awarded lower scores on these criteria); criteria which are lower are easier.

As can be seen from the map, test takers show a spread of approximately three logits, while marker, criterion, and task facets cluster closely around the zero-logit line, indicating broadly comparable behaviour across these facets. Task prompt difficulty spans a narrow range (approximately −0.21 to +0.15 logits), with infit values close to 1.0, suggesting minimal variation in prompt difficulty.

Figure 1: *Facet maps*



Marker Facet

Table 5 presents the MFRA statistics for the two scoring sources.

Table 5: *Marker Measurement Report*

Total Count	Obsvd Average	Fair(M) Average	Model Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Discrm	PtBis	N Markers
19344	44.04	43.86	.01	.00	.82	-9.0	.80	-9.0	1.34	.40	2 Automarker
19896	44.00	44.31	-.01	.00	1.18	9.0	1.15	9.0	.67	.39	1 Human
19620.0	44.02	44.08	.00	.00	1.00	.0	.97	.0		.39	
390.3	.03	.32	.01	.00	.25	12.7	.24	12.7		.01	

Model, Sample: RMSE .00 Adj (True) S.D. .01 Separation 3.18 Strata 4.58 Reliability .91
Model, Fixed (all same) chi-squared: 11.1 d.f.: 1 significance (probability): .00

Marker reliability was high (0.91), indicating consistent internal ranking of scripts. Both the human and automated markers were located close to the zero-logit line, suggesting comparable overall severity. Infit statistics for both sources fell comfortably within accepted operational thresholds.

Task Prompt Facet

In previous MFRA analyses, Tasks 1 and 2 showed very comparable statistics: both were located around the zero-logit line indicating that, as facets, they were operating very similarly. Consequently, the analysis below presents a composite analysis of the questions that constitute both Tasks 1 and 2. Table 6 reports Task Prompt Measurement statistics (labelled as “Questions” in FACETS output).

Task prompt reliability was high (0.99), and all prompts demonstrated infit and outfit values within the 0.5–1.5 range, indicating good fit to the model. Severity ranges from about –0.21 to +0.15 logits, i.e., < 0.5 logit spread. The narrow spread of prompt difficulty supports the interpretation that prompts functioned equivalently across the dataset.

Table 6: *Task Prompt Measurement Report*

Fair(M) Average	Model Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Question
42.28	.06	.01	1.21	6.7	1.17	5.3	57408
43.59	.02	.01	1.17	5.3	1.13	4.0	59651
42.78	.04	.01	1.16	4.9	1.11	3.3	59075
41.74	.08	.01	1.14	4.7	1.11	3.6	58195
41.23	.09	.01	1.12	3.6	1.08	2.5	74112
46.95	-.09	.01	1.09	2.3	1.10	2.6	67259
42.47	.05	.01	1.04	1.2	.98	-.5	57595
40.81	.11	.01	1.03	.9	.99	-.1	58140
49.79	-.19	.01	.99	-.2	.99	-.2	70394
50.32	-.21	.01	.98	-.5	1.00	.0	66924
40.34	.12	.01	.99	-.1	.96	-1.3	71701
43.63	.01	.01	.98	-.8	.95	-1.8	72344
42.13	.06	.01	.98	-.7	.94	-1.8	59565
40.02	.13	.01	.92	-2.7	.91	-3.2	58194
41.90	.07	.01	.94	-2.2	.91	-3.2	59782
47.48	-.11	.01	.90	-2.9	.89	-3.0	69580
49.24	-.17	.01	.89	-4.4	.90	-3.9	70134
47.65	-.12	.01	.89	-2.9	.88	-3.1	67370
39.54	.15	.01	.89	-3.9	.87	-4.5	59776
42.24	.06	.01	.84	-5.7	.81	-6.9	72343
48.66	-.15	.01	.80	-5.5	.79	-5.7	67356
44.04	.00	.01	1.00	-.1	.97	-.9	
3.51	.12	.00	.12	3.7	.11	3.4	

Model, Sample: RMSE .01 S.D. .12 Separation 11.18 Strata 15.23 Reliability .99
Model, Fixed (all same) chi-squared: 2376.7 d.f.: 20 significance (probability): .00

Discrepancy Analysis

Discrepancy analysis examined cases in which differences between human and automated scores exceeded LANGUAGECERT's routine quality-assurance threshold, triggering review by a Chief Examiner. Under normal operational conditions, approximately 10% of scripts marked by two human raters require such review.

Table 7 below presents the figures for Tasks 1 and 2 where the differences between human and automarker scores met the criteria for third-marker review under this policy.

Table 7: *Task discrepancy figures between the two markers*

Task	Raw figure	Percentage discrepancy
1	37/2,389	1.55%
2	38/2,394	1.59%

In the present dataset, discrepancies meeting the review threshold occurred in approximately 1.5% of scripts for each task. These figures indicate that automated scores aligned closely with operationally finalised human scores and generated substantially fewer review-triggering cases than are typically observed in human–human marking.

Taken together, the results suggest that, under operational conditions, the automated scoring system exhibits stable and coherent scoring behaviour, consistent with its intended role in Phase 1 of the validation programme. However, it important to emphasise that Marker 1 represents quality-assured operational scores after LANGUAGECERT moderation procedures. Approximately 10% of scripts underwent Chief Examiner adjudication; others represent concordant double-marking. The 1.5% automarker discrepancy rate therefore reflects disagreement with moderated outcomes, not pre-moderation human-human variance. Pre-moderation variability is not analysed in Phase 1. It is however a key issue and will be examined in Phase 2's independent multiple-rater design.

Limitations

The scope of the present study is intentionally bounded by its role within a staged validation programme. While the automated scores are compared against operationally finalised human scores produced through LANGUAGECERT's standard marking and moderation processes, the design does not incorporate multiple independent human ratings of each script. As a result, the study does not estimate the full distribution of human inter-rater variability or define a human benchmark beyond the operational composite.

This reflects the specific evidential purpose of Phase 1, which is to examine large-scale operational behaviour, score stability, and construct-aligned performance under realistic scoring conditions. Questions that require independent rater replication—such as fine-grained estimates of rater severity differences or direct comparison of automated–human agreement against the full range of human–human variability—lie outside the scope of this phase.

Subsequent phases of the validation programme address these complementary questions through designs that incorporate multiple independent human raters and more controlled rating conditions. The present findings should therefore be interpreted as foundational operational evidence that supports, but does not replace, later precision- and fairness-focused validation work.

Conclusions

This study reported Phase 1 of a staged validation programme for the LANGUAGECERT Automated Writing Scoring System, focusing on its operational behaviour on the LANGUAGECERT Academic Writing Test (LCAWT). The analyses drew on a large, operationally representative dataset comprising 2,394 test takers, each completing two writing tasks. Automated scores were compared with operationally finalised human scores produced through LANGUAGECERT's standard marking and moderation processes, reflecting authentic scoring conditions in high-stakes assessment.

Across both tasks, automated and human scores demonstrated strong correspondence at the total-score level, with correlations of 0.87. Many-Facet Rasch Analysis indicated acceptable fit for marker, criterion, and task prompt facets, with stable behaviour across scoring dimensions. Discrepancy rates relative to routine quality-assurance thresholds were low, and substantially below those typically observed in human–human marking under comparable conditions.

Taken together, the findings indicate that the automated scoring system operates coherently within the intended assessment framework and produces score patterns that are interpretable alongside human judgements under operational conditions. In line with the evidential purpose of Phase 1, these results provide foundational operational evidence regarding score stability, construct alignment, and system behaviour at scale. They do not constitute a complete validation of automated scoring performance, nor do they address questions requiring independent human rater replication or subgroup-level fairness analysis.

These complementary questions are addressed through subsequent phases of the validation programme, including controlled multi-rater precision studies and analyses of subgroup behaviour. The present study establishes an appropriate empirical foundation for this continued work by demonstrating that the automated scoring system behaves predictably and proportionately within a human-in-charge assessment model. In this way, the phased approach supports a cumulative and methodologically coherent validation strategy, with the current study serving as a necessary first stage rather than a definitive endpoint.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. Journal of Technology, Learning, and Assessment, 4(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650/1492

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

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

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics. (pp. 4171-4186).

Eckes, T. (2015). Introduction to many-facet Rasch measurement. Frankfurt am Main: Peter Lang.

Foltz, P.W., Laham, D., and Landauer, T.K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2), http://imej.wfu.edu/articles/1999/2/04/index.asp

Hatch, E. & Lazaraton, A. (1991). The Research Manual. Heinle and Heinle: Boston, MA.

Kostic, M., Witschel, H. F., Hinkelmann, K., & Spahic-Bogdanovic, M. (2024). LLMs in Automated Essay Evaluation: A Case Study. Proceedings of the AAAI Symposium Series, 3(1), 143-147. https://doi.org/10.1609/aaaiss.v3i1.31193

Kwon, S. Y., Bhatia, G., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Beyond English: Evaluating LLMs for Arabic grammatical error correction. arXiv preprint arXiv:2312.08400.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. Journal of Applied Measurement, 3(1), 85-106.

Linacre, J. M. (2024) FACETS computer program for many-facet Rasch measurement. Beaverton, Oregon: Winsteps.com.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19(3), 246–276. https://doi.org/10.1191/0265532202lt231oa

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

Mansour, W., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can Large Language Models Automatically Score Proficiency of Written Essays?. arXiv preprint arXiv:2403.06149.

Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 48 , 238–243.

Page, E. B. (2003). Project essay grade: PEG. In: M. D. Shermis & J. C. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 43–54). Hillsdale, NJ: Erlbaum

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

Ramineni, C., & Williamson, D.M. (2013). Automated essay scoring: Psychometric guidelines and practices. Assessing Writing, 18, 25-39.

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. ReCALL (2024), 36(2), 187–209.

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. Language Teaching, 31(2), 57–71. https://doi.org/10.1017/S0261444800012970

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.

Weir, C. J. (2005). Language Testing and Validation. Palgrave Macmillan. https://doi.org/10.1057/9780230514577

Xiao, C., Ma, W., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. arXiv preprint arXiv:2401.06431

Chapter 4: Exploring LLM Simulations for Pre-Operational Prediction of Reading Test Item Difficulty

David Coniam, Michael Milanovic and Leda Lampropoulou

Abstract

This paper reports on an exploratory simulation-based investigation into the potential use of a large language model (LLM) to predict item facility values for an objective reading test. The study used a 30-item multiple-choice reading test from the LANGUAGECERT Academic (LCA) suite, for which robust empirical data from 671 test takers were available. These item statistics provided a baseline against which the LLM's predictive performance could be evaluated.

Four progressively constrained simulation scenarios were designed. In each scenario, the LLM was instructed to respond to the test items as if it were test takers at specified CEFR proficiency levels, generating multiple responses per level. The resulting datasets were analysed using classical test statistics and compared with empirical item facility values derived from live administrations. Later scenarios incorporated increasingly realistic CEFR-level distributions and, in the most constrained condition, a detailed CEFR-aligned syllabus.

Under constrained conditions, the model produced test-level statistics that closely approximated empirical results, with a mean facility discrepancy of 0.02 in the most aligned scenario. However, variation remained at the individual item level, with a small number of items displaying notable divergence from observed values. These findings suggest that while simulation-based LLM approaches may approximate aggregate test behaviour, consistent item-level precision remains challenging.

The study highlights both the potential and the limitations of simulation-based psychometric prediction. Further refinement of prompting strategies and hybrid modelling approaches may enhance precision, while complementary data-driven methods grounded in calibrated item banks may offer a more stable basis for operational implementation.

Keywords: Large Language Models (LLMs), item facility prediction, simulation-based modelling, LANGUAGECERT Academic

Introduction

The rapid emergence of Large Language Models (LLMs) such as ChatGPT-4 has generated much interest in their potential applications to many fields. The English language assessment team at LANGUAGECERT has been exploring such applications since 2022. There has been research on areas related to test development, such as automated item generation (Dourda & Jones, 2026; Jones & Dourda, 2026) or the development of preparation and practice materials (Milanovic & Jones, 2026) as well as areas related to test security such as deep fake technology, item harvesting and autogenerated responses. While LLM tools have demonstrated remarkable capabilities in generating human-like responses, completing complex linguistic tasks, and even passing standardised examinations, their use in the technical domain of psychometric prediction remains comparatively unexplored.

At LANGUAGECERT, attention has recently focused on a programme of work investigating the extent to which LLMs might be used to predict item statistics normally accessible only after pretesting or live administration. This paper

presents findings from an initial exploratory study that considered the extent to which an LLM (GPT-4) could meaningfully approximate item performance statistics for language assessment, specifically focusing on item facility (IF) values for a reading comprehension test.

The motivation for this research stemmed from a range of practical challenges in language test development. These include the need to pretest items before operational deployment, to generate vast quantities of test materials to mitigate item harvesting and malpractice, and to maintain stable test difficulty while scaling production. Traditional pretesting is resource-intensive, requiring appropriate test-taker samples, pilot test administrations, and post-hoc statistical analyses. If LLM-based approaches could provide reasonably accurate preliminary estimates of item performance, they might serve as an additional screening tool, helping test developers identify potentially problematic items before investing in full-scale pretesting or they may even contribute to the generation of items with preliminary predicted statistical values prior to empirical validation.

The study reported here used a simulation-based methodology, asking an LLM (GPT-4) to respond to test items as if it were test takers at different Common European Framework of Reference (CEFR) levels. By generating multiple responses at each proficiency level and analysing the resulting response patterns, the study examined whether predicted item facility values approximate empirically observed data from actual test administrations. Four progressively refined scenarios were explored, each introducing additional constraints and contextual information intended to improve prediction accuracy.

The investigation used a 30-item reading test from the LANGUAGECERT Academic (LCA) suite, for which robust empirical data from 671 test takers were available. The availability of confirmed item statistics provided a solid baseline against which the LLM's predictive performance could be evaluated. Through systematic manipulation of instructions, sample distributions, and background information provided to the model, the study examined whether an LLM can predict item difficulty, as well as what types of scaffolding and constraints might be necessary to optimise its performance in this task.

Literature Review: AI in English language assessment

While earlier generations of "artificial intelligence" (AI) have been used in educational assessment since the late 1960s, interest in AI applications intensified in the early 2020s, particularly following the emergence of Large Language Models (LLMs). In language assessment, this has accelerated research and operational experimentation across multiple parts of the assessment lifecycle, from item development to scoring and delivery.

One of the most established applications of AI in language assessment is automated marking and scoring, particularly for productive skills such as writing and speaking. Automated essay scoring systems have been studied extensively in relation to reliability, construct representation, and operational deployment (Attali & Burstein, 2006; Coniam et al., 2025; Williamson et al., 2012). Similarly, research on automated spoken-response scoring has demonstrated how speech recognition, acoustic features, and supervised learning models can support large-scale speaking assessment, while emphasising the need for careful validation and monitoring (Zechner et al., 2009; Bernstein et al., 2010). Across both writing and speaking contexts, the assessment literature consistently frames automated scoring not as a replacement for psychometric principles, but as a system that must be evaluated within established measurement frameworks.

Another important strand of technology-enabled assessment concerns computerised adaptive testing (CAT) and related models of dynamic test assembly. In CAT, item selection occurs during test administration based on real-time ability estimates derived from item response theory, allowing the test to adapt to the test taker's performance (Wainer, 2000; Van der Linden & Glas, 2010). Similarly, linear-on-the-fly test assembly (LOFT) approaches construct equivalent test forms dynamically from large, calibrated item banks using optimisation algorithms (Van der Linden, 2005). These models demonstrate that "on-the-fly" delivery is feasible and operationally robust when grounded in empirically estimated item parameters and well-defined measurement models. Crucially, both adaptive selection and dynamic

assembly rely on stable, pre-calibrated item characteristics derived from field data.

A developing strand of AI in language assessment is automated test item generation (AIG). Traditionally, AIG has relied on structured item and cognitive models to generate large item banks while maintaining construct alignment and psychometric control (Gierl & Lai, 2012, 2016). With the emergence of LLMs, research has increasingly explored their use across pre-generation, generation, and post-generation stages of the AIG process (Dourda & Jones, 2026; Tan et al., 2025). A recent review of 60 empirical studies found that models such as T5, BERT, and GPT variants are widely used to generate items across domains and languages; however, evidence regarding key measurement properties, such as difficulty, discrimination, and reliability remains limited (Tan et al., 2025). This distinction between linguistic generation and psychometric validation remains central in high-stakes assessment contexts.

Taken together, these strands of research illustrate both the potential and the constraints of AI integration in language assessment. Despite advances in automated scoring, item generation, and adaptive delivery, these applications share a common feature: they rely on empirically derived data and established measurement frameworks. Automated scoring models are trained on large corpora of human-rated responses, and adaptive testing depends on pre-calibrated item parameters estimated from field data. By contrast, comparatively little research has examined whether LLMs can approximate psychometric item properties prior to pretesting. While simulation-based approaches may offer preliminary insights, the extent to which they can produce stable and operationally reliable item-level estimates remains an open empirical question. Given the operational importance of pretesting in large-scale language assessment, further empirical investigation of this issue is warranted.

Research Question

The broad research question being pursued involved the extent to which an LLM such as GPT-4 would be able to predict item facility values for objective reading test items in the context of test takers at different CEFR levels. Specifically, the study sought to determine whether a simulation-based approach, in which the model is instructed to respond as test takers at different proficiency levels, can produce item facility estimates that are sufficiently close to empirically observed values to be of practical use in test development.

The Study

As outlined above, the study adopted a simulation-based methodology, asking GPT-4 (a large language model developed by OpenAI) to respond to a set of objective reading test items as if it were test takers at different CEFR proficiency levels. By generating multiple responses at each level and analysing the resulting response patterns using classical test statistics, the study examined whether AI-generated data could yield item facility values comparable to those derived from live test administrations.

The investigation reported in this paper was conducted in mid-2025. It used a 30-item multiple-choice reading test from the LCA suite, for which robust empirical data from 671 test takers were available. These empirically derived item statistics provided a stable baseline against which the model's predictions could be evaluated. The test is level-agnostic, covering CEFR levels B1–C2, with scoring aligned to both CEFR and the LANGUAGECERT Global Scale.

To explore the feasibility of simulation-based prediction under different conditions, four scenarios were designed, each introducing progressively tighter constraints intended to improve the plausibility of the simulated data. Across all scenarios, the LLM was instructed to respond to each item multiple times as if it were a test taker at specified CEFR levels. The resulting response

sets were analysed using classical test statistics, in line with standard LANGUAGECERT procedures for reading tests.

The four simulation scenarios are summarised in Table 1, which outlines the constraints applied, sample sizes, and CEFR levels represented in each condition.

In scenarios 1 and 2, the model was asked to generate responses with minimal constraints. In Scenario 1, it responded as if it were test takers at CEFR levels A1 through C2, producing 50 responses per level. Scenario 2 restricted responses to CEFR levels A2 through C1, again with equal numbers of responses per level. These scenarios were intended to establish baseline behaviour and to observe how the model distributed responses across proficiency levels in the absence of realistic population constraints.

Scenario 3 introduced a more realistic test-taker distribution. In this condition, the composition of the requested response sets was designed to resemble the CEFR distribution observed in the baseline live test. The LLM was instructed to generate responses reflecting the actual CEFR distribution observed in the live test sample (below-B1 to C2), with response counts matched to the empirical dataset. This scenario aimed to examine whether aligning the simulated sample more closely with real test-taker populations would improve the correspondence between predicted and observed item statistics.

Scenario 4 built on Scenario 3 by adding a further layer of constraint. Its sample was specified as for Scenario 3. In addition to using the empirical test-taker distribution, the LLM was provided with a CEFR-aligned syllabus detailing relevant vocabulary, grammar, syntax, topics, and language functions, available to potential test takers from the LANGUAGECERT website. The intention was to determine whether supplying explicit linguistic and proficiency-level information would enable the model to produce response patterns more closely aligned with empirical data.

Table 1: *Four GPT analysis scenarios*

Scenario	Constraints	Sample size	CEFR levels to respond to
1	No constraints. The model should respond to each item as if it were an A1, A2, B1, B2, C1, C2 test taker.	300 50 responses to each level	A1 to C2
2	No constraints. The model should respond to each item as if it were an A2, B1, B2, C1 test taker.	200 50 responses to each level	A2 to C1
3	The model should respond to each item as if it were a below-B1, B1, B2, C1, C2 test taker.	671 below-B1:45 B1:199 B2:234 C1:146 C2:47	below-B1, B1, B2, C1, C2
4	Syllabus containing vocabulary, topics, functions, grammar pertinent to specific CEFR levels uploaded for GPT to digest - The model should respond to each item as if it were a below-B1, B1, B2, C1, C2 test taker.	671 below-B1:45 B1:199 B2:234 C1:146 C2:47	below-B1, B1, B2, C1, C2

For all scenarios, the model was provided with the full set of test items; however, it was not given access to the item keys or any information about correct answers. Responses were recorded in structured Excel templates corresponding to each CEFR level, and item-level and test-level statistics were subsequently calculated and compared with the live test data.

The detailed instructions provided to the model for Scenario 3 are illustrated in Figure 1. The instructions were broadly similar across all scenarios, with variations primarily in the specified sample sizes. In Scenario 4, the LLM was additionally informed that a CEFR-aligned syllabus would be uploaded and that it should digest this information before generating revised predictions.

Figure 1: *Instructions for GPT*

I want to get input as to how English language reading test items work.

I am going to upload to you a 30-item Reading Test, and an Excel Response Template containing a grid for responses.

In the Excel file there are 5 worksheets. Each worksheet contains space for responses to all 30 items by test takers as CEFR levels below-B1 (b-B1), B1, B2, C1, C2 respectively.

Record your answers to each item on the appropriate Excel worksheet.

Can you respond to each of the 30 Reading Test items the number of times below according to the level of below-B1, B1, B2, C1, C2 test takers.

Level	N
Below-B1 (b-B1)	45
B1	199
B2	234
C1	146
C2	47

When you respond to the test items, follow the restrictions below

Test part	Items	Permissible responses
1A	1-6	A-D
1B	7-11	A-C
2	12-17	A-H
3	18-24	A-D
4	25-30	A-D

Make sure you give thought to each item. Do not just answer randomly.

Following completion of each simulation condition, the model produced response datasets that were analysed using classical test statistics to evaluate alignment with empirically observed item behaviour. An example of the response dataset generated for one CEFR level is shown in Table 2, illustrating the structure of the simulated data used for subsequent statistical analysis.

Table 2: *GPT's predicted responses*

	A	B	C	D	E	F	G
1	TestTakerID	1	2	3	4	5	6
2	B2_001	C	D	B	D	C	D
3	B2_002	C	D	A	C	C	D
4	B2_003	C	D	B	D	C	D
5	B2_004	C	B	B	B	D	D
6	B2_005	C	D	B	D	C	A
7	B2_006	C	D	B	D	C	D
8	B2_007	D	D	B	B	C	D
9	B2_008	C	B	B	A	C	D
10	B2_009	C	D	B	D	C	D
11	B2_010	C	D	B	D	C	D
12	B2_011	C	D	B	D	D	D

Analysis

Following the completion of GPT's predictions for each scenario, all output underwent classical test statistics (CTS) analysis. The analysis focused first on test-level statistics, followed by a more detailed examination of item-level performance, with item facility values used as the primary basis for comparison.

Overall test statistics

This section reports test-level statistics for the live test alongside the four GPT-generated scenarios. Table 3 presents test and group mean scores, standard deviation (SD), and reliability (alpha). The live test statistics are shown in the second column in red font and serve as the empirical baseline. The LANGUAGECERT Academic Reading test targets CEFR levels B1 to C2, although some test takers obtain scores below B1; these are referred to here as "below-B1".

Table 3: *Test-level statistics*

Test type	Live test	GPT			
Scenario		1	2	3	4
Levels assessed	Below-B1 to C2	A1 to C2	A2 to C1	below-B1 to C2	below-B1 to C2
Sample size		50 per level	50 per level	as per live test	as per live test
Other					with Syllabus
Item total	30	30	30	30	30
Sample total	671	300	200	671	671
Test mean	17.03	17.59	21.09	19.39	16.63
Test mean%	56.77	58.63	70.3	64.62	55.43
SD	5.91	11	7.17	6.27	5.96
SD%	19.70	36.67	23.90	20.90	19.87
Cronbach's α	0.84	0.97	0.92	0.85	0.85

Scenarios 1 and 2 were demonstrably out of scope and therefore offer no value for operational test development. In scenario 1, the overall mean appeared superficially close to that of the live test; however, this was effectively due to the inclusion of extreme A1 and C2 groups. Scenario 2, which excluded these extreme

groups, yielded a substantially higher mean score, reflecting the absence of lower-ability test takers.

Scenarios 3 and 4 showed closer alignment with the live test at the test level. In particular, Scenario 4, where GPT was provided with a CEFR-aligned syllabus, produced a mean score (55.43%) that is the closest to the live test mean (56.77%).

Reliability estimates (Cronbach's alpha) were extremely high for Scenarios 1 and 2. This is hardly surprising given the artificial stratification with equal sample sizes across A1 to C2. The standard deviations (SD) in these scenarios are also markedly inflated (e.g., SD= 11 in Scenario 1 compared with 5.91 in the live test), primarily due to the inclusion of extreme proficiency groups in equal proportions. In contrast, alphas and SDs for Scenarios 3 and 4 were more comparable to those of the live test, suggesting that, at an aggregate level, the simulated data can approximate global test characteristics when more realistic distributions are used.

To further exemplify mean scores, detail on group (i.e., CEFR level) performance is presented in Table 4.

Table 4: *Group distributions*

	Live test		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Group	N	Mean %	N	Mean %	N	Mean %	N	Mean %	N	Mean %
A1	–	–	50	4.8	–	–	–	–	–	–
A2	45	21.34	50	20.87	50	33.93	45	25.85	45	21.85
B1	199	39.10	50	51.07	50	68.13	199	46.21	199	38.64
B2	234	57.38	50	81.00	50	86.53	234	68.21	234	55.71
C1	146	76.65	50	94.73	50	92.60	146	86.03	146	76.96
C2	47	92.75	50	99.27	–	–	47	95.39	47	90.43
Means	671	56.77	300	58.63	200	70.3	671	64.62	671	55.43

Drawing on the overall group means for the live Reading Test, the whole group mean for all 671 test takers was 56.77%.

Under Scenario 1, A1 test takers scored almost zero and C2 test takers virtually 100%, an unlikely distribution in operational contexts. In Scenario 2, removing A1 and C2 test takers produced a more plausible distribution, though performance remained inflated at higher levels.

Scenario 3's group mean scores were reasonably close to the live test whole group although ability appeared to be overestimated somewhat at the C1 level.

Scenario 4 produced group mean scores that were reasonably close to those observed in the live test, with differences generally within a few percentage points. This suggests that increased constraint and contextual information can improve test-level and group-level approximation.

Item Statistics

While test-level results provided an initial indication of plausibility, item-level behaviour was the critical focus of this study. Item facility (IF) values were therefore examined in detail across all scenarios. Item discrimination indices were above the accepted benchmark of 0.3 (Falvey et al., 1994) in most scenarios; however, for clarity and relevance, the discussion below focuses primarily on IF values. Table 5 presents IF values for the live test alongside those generated under the four ChatGPT scenarios.

Table 5: *Item analyses*

	Live test	Scenario 1	Scenario 2	Scenario 3	Scenario 4
		A1-C2	A1-C2	below B1-C2	below B1-C2
		50 per level	50 per level	distribution as per live test	distribution as per live test / with syllabus
Item	**IF**	**IF**	**IF**	**IF**	**IF**
R1	0.86	0.79	0.86	0.55	0.86

	Live test	Scenario 1	Scenario 2	Scenario 3	Scenario 4
R2	0.84	0.74	0.90	0.55	0.86
R3	0.72	0.72	0.86	0.54	0.83
R4	0.74	0.68	0.85	0.57	0.83
R5	0.56	0.69	0.86	0.55	0.85
R6	0.38	0.69	0.82	0.57	0.84
R7	0.63	0.66	0.82	0.57	0.72
R8	0.68	0.69	0.82	0.56	0.70
R9	0.51	0.65	0.80	0.50	0.72
R10	0.89	0.66	0.81	0.52	0.73
R11	0.82	0.61	0.79	0.59	0.70
R12	0.40	0.68	0.73	0.57	0.71
R13	0.49	0.62	0.77	0.56	0.45
R14	0.54	0.58	0.72	0.54	0.51
R15	0.27	0.61	0.72	0.54	0.48
R16	0.55	0.59	0.71	0.54	0.52
R17	0.31	0.58	0.66	0.50	0.54
R18	0.50	0.58	0.68	0.54	0.51
R19	0.56	0.50	0.64	0.53	0.53
R20	0.60	0.54	0.63	0.57	0.36
R21	0.62	0.52	0.64	0.52	0.33
R22	0.44	0.52	0.67	0.57	0.31
R23	0.40	0.52	0.62	0.55	0.32
R24	0.57	0.53	0.52	0.55	0.37
R25	0.57	0.46	0.57	0.53	0.35
R26	0.28	0.40	0.56	0.57	0.31
R27	0.79	0.44	0.51	0.55	0.34
R28	0.53	0.44	0.57	0.54	0.37
R29	0.51	0.46	0.50	0.58	0.33
R30	0.47	0.45	0.46	0.55	0.35
MEAN	0.57	0.59	0.70	0.55	0.55

In Scenarios 1 and 2, GPT's predictions imposed a clear and artificial facility gradient across the test, with items becoming progressively more difficult as the test progresses. While such a pattern may appear intuitively reasonable, it was not reflected in the empirical data, where item difficulty was distributed more irregularly. Scenario 3 produced a different but still implausible pattern, with item facilities clustering closely together across items. This lack of variation again contrasted with the live test data and suggests that aligning simulated sample distributions alone is insufficient to capture realistic item behaviour.

Scenario 4 showed greater variation and, at face value, appeared to approximate the empirical pattern more closely. Table 6 compares IFs for the live test and Scenario 4 directly. The final column provides the difference between each set of IFs.

Table 6: *IFs for the Live Test and for Scenario 4.*

	Live Test	GPT Scenario 4	Difference
Item	**IF**	**IF**	**(Live – GPT)**
R1	0.86	0.86	0.00
R2	0.84	0.86	-0.02
R3	0.72	0.83	-0.11
R4	0.74	0.83	-0.09
R5	0.56	0.85	-0.29
R6	0.38	0.84	-0.46
R7	0.63	0.72	-0.09
R8	0.68	0.70	-0.02
R9	0.51	0.72	-0.21
R10	0.89	0.73	0.16
R11	0.82	0.70	0.12
R12	0.40	0.71	-0.31
R13	0.49	0.45	0.04
R14	0.54	0.51	0.03
R15	0.27	0.48	-0.21
R16	0.55	0.52	0.03

	Live Test	GPT Scenario 4	Difference
Item	**IF**	**IF**	**(Live – GPT)**
R17	0.31	0.54	-0.23
R18	0.50	0.51	-0.01
R19	0.56	0.53	0.03
R20	0.60	0.36	0.24
R21	0.62	0.33	0.29
R22	0.44	0.31	0.13
R23	0.40	0.32	0.08
R24	0.57	0.37	0.20
R25	0.57	0.35	0.22
R26	0.28	0.31	-0.03
R27	0.79	0.34	0.45
R28	0.53	0.37	0.16
R29	0.51	0.33	0.18
R30	0.47	0.35	0.12
MEAN	0.57	0.55	0.02

While some items showed relatively small differences, only two items exhibited discrepancies exceeding 0.4. The overall mean difference between live and simulated IF values was 0.02, indicating strong aggregate alignment (0.57 vs 0.55). However, variation at the individual item level remained evident.

Across the test as a whole, item-by-item correspondence varied, with some items closely aligned and others displaying moderate divergence. In particular, the model tended to overestimate facility for earlier items and underestimate facility for later items, again reflecting an imposed progression that was not supported by empirical evidence. Although the aggregate mean IF for Scenario 4 (0.55) is close to that of the live test (0.57), this overall similarity masked substantial item-level inaccuracies.

Discussion

This paper presented an exploratory, simulation-based analysis of whether an LLM can predict item facility values for an objective reading test. Four progressively constrained scenarios were examined, with findings that reveal both the potential and the persistent limitations of this approach.

The most constrained scenario, which used realistic test-taker distributions and a detailed CEFR-aligned syllabus, produced test-level statistics closely aligned with empirical values (mean IFs of 0.55 vs 0.57 for the live test), with an overall mean discrepancy of 0.02 across items. Yet aggregate correspondence masked considerable item-level variability: two items showed discrepancies exceeding 0.4, and several others displayed moderate divergence.

A recurring pattern in the less constrained scenarios was the model's tendency to impose an artificial facility gradient, predicting items as increasingly harder across the test, a pattern absent from the empirical data, where difficulty was distributed more irregularly. Even with extensive scaffolding – vocabulary lists, grammatical and syntactic specifications, and topic descriptors for each CEFR level – the model could not fully capture the complex interaction between item characteristics and heterogeneous test-taker behaviour that underlies observed facility values. Strong linguistic competence, it seems, does not straightforwardly translate into stable psychometric modelling of population-level performance.

Two directions appear most promising for future work. The first concerns methodological refinement: improvements to prompting strategies, conditioning mechanisms, or hybrid modelling approaches may enhance item-level precision. The second shifts orientation more substantially. Rather than relying on simulated test-taker responses, data-driven methodologies trained on large corpora of calibrated items with confirmed psychometric properties could offer a more stable empirical foundation A system of this kind might also yield interpretable diagnostic output, identifying which item features drive predicted difficulty or discrimination values, and thereby supporting the

attribution of reliable statistical characteristics to items produced through automated authoring.

These findings underscore the importance of rigorous methodological evaluation when integrating AI tools into language assessment. Simulation-based approaches show meaningful promise at the aggregate level, but require substantial refinement before they can inform operational decisions at the item level.

Conclusion

This study investigated whether LLM-based simulation of CEFR-level test-taker behaviour could yield reliable predictions of item facility values for an objective reading test. Under the most constrained conditions, aggregate alignment was strong, with a mean discrepancy of just 0.02 between predicted and observed values. Item-level precision, however, remained inconsistent: while most predictions fell within a reasonable range of empirical estimates, a small number of items showed notable divergence — a degree of variability that falls short of the stability required for high-stakes operational use.

The broader implication is that direct simulation of test-taker behaviour through LLMs is a methodologically demanding undertaking, one where current capabilities are better suited to approximating test-level tendencies than to modelling individual item behaviour with the reliability that assessment contexts demand. Progress will likely depend on advances in constraint design and prompting, alongside closer integration with empirically validated item banks and established psychometric frameworks. As AI capabilities continue to develop, their effective integration into language assessment will depend on sustained alignment with established principles of validity, reliability, and empirical calibration.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. Journal of Technology, Learning, and Assessment, 4(3), 1–30.https://ejournals.bc.edu/index.php/jtla/article/view/1650

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. Language Testing, 27(3), 355-377. https://doi.org/10.1177/0265532210364404

Coniam, D., Lampropoulou, L., & Megaritis, V. (2025). Operational validation evidence for the LANGUAGECERT automated writing scoring system: Phase 1. LANGUAGECERT.

Dourda, C., & Jones, C. (2026). AI-assisted item generation in high-stakes language assessment: A Design-Based Study of Human-AI Item Development. LANGUAGECERT.

Falvey, P., Holbrook, J., & Coniam, D. (1994). Assessing students. Hong Kong: Longman.

Jones, C., & Dourda, C. (2026). Making the invisible visible: how human-AI collaboration transforms item development practices in language testing. LANGUAGECERT.

Milanovic, M., & Jones, C. (2026). Mobile-based English language learning through bite-sized formative CEFR-aligned assessment. LANGUAGECERT.

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2025). A review of automatic item generation techniques leveraging large language models. International Journal of Assessment Tools in Education, 12(2), 317–340. https://doi.org/10.21449/ijate.1602294

Van der Linden, W. J. (2005). Linear models for optimal test design. Springer. https://doi.org/10.1007/0-387-29054-0

Van der Linden, W. J., & Glas, C. A. (Eds.). (2010). Elements of adaptive testing (Vol. 10). Springer.

Wainer, H. (2000). Computerized adaptive testing: A primer (2nd ed.). Lawrence Erlbaum Associates.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13.

https://doi.org/10.1111/j.1745-3992.2011.00223.x

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of nonnative spontaneous speech in tests of spoken English. Speech Communication, 51(10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009

SECTION 2: MEASUREMENT, CALIBRATION & VALIDATION

Chapter 5: The Relationship between Language Complexity and Test-taker Achievement on a High-Stakes Test of Writing

Leda Lampropoulou, Irene Stoukou and David Coniam

Abstract

The current study explored the relationship between language complexity and test-taker proficiency at Common European Framework of Reference for Languages (CEFR) levels B1, B2, and C1 from scripts extracted from a high-stakes examination – the LANGUAGECERT Academic Writing Test. The purpose of the study involved exploring the extent to which test takers obtaining higher grades on a writing test actually produce more advanced-level writing skills, providing possible validity evidence for a high-stakes test of writing. 2,746 scripts, produced for Task 1 and Task 2 of the LANGUAGECERT Academic Writing test, which had been graded as CEFR levels B1, B2, or C1, were passed through the computational linguistic analytic tool Coh-Metrix 3.0. The scripts

were analysed against nine categories of lexical, syntactic and discourse level features in 62 individual subcategories. ANOVAs were conducted on each of the 62 subcategories against CEFR level.

In total, 68% (42 of the 62) of the subcategories followed the expected progression. Texts produced by B2 test takers showed a greater number of higher-level linguistic features than B1 texts, and C1 texts in turn showed a greater number of higher-level linguistic features than those of B2 test takers. The B1 to B2 to C1 progression on the 42 subcategories was also seen to be significant. The conclusion drawn is that with over two thirds of linguistic features affirming that higher ability test takers do produce more complex texts than lower-ability test takers, the LANGUAGECERT Academic Writing Test performs as it is intended to, grading test takers appropriately.

Keywords: language complexity, writing proficiency, coh-metrix, LANGUAGECERT Academic

Introduction

The Common European Framework of Reference for Languages (CEFR) provides detailed descriptions of language proficiency across the six different levels (A1 to C2). As learners progress through these levels, their language ability is expected to show increased lexical, syntactic and discourse complexity.

Against this backdrop, the relationship between linguistic complexity and L2 proficiency has been a longstanding focus of research. The underlying assumption is that the use of more complex lexical, syntactic and discourse structures within a text indicate more advanced-level writing skills (Crossley, 2020).

Crossley (2020) presents a cogent overview of the interactions between lexical, syntactic and discourse structures and L2 proficiency. While he highlights how more proficient L2 writers generally illustrate greater ability in their control of lexical, syntactic and discourse structures, he nonetheless urges that the

relationship between the features in terms of complexity and L2 proficiency is dependent upon various contextual factors and individual differences.

The focus of Crossley's (2020) study is essentially argumentative writing, where "L2 writing is often used as a proxy of language ability in both standardised tests and language acquisition studies." In the current paper, this assumption is related directly to LANGUAGECERT's Academic test of English (LCA).

The LCA is designed as a multi-level test, aligned with the CEFR, spanning levels B1 to C2. As outlined in the LANGUAGECERT Academic Qualification Handbook (Version 7.0) (which is provided at the following link: https://www.LANGUAGECERT.org/en/language-exams/english/LANGUAGECERT-academic/-/media/d4f4c8da7ef24dd6be4666c7729308d5.ashx), the design of the LANGUAGECERT Academic test allows test takers' performance to be mapped against a continuous proficiency scale, rather than being confined to a single level. The test is underpinned by a complex set of constructs that lay out what is intended to be assessed at each CEFR level in terms of lexis, grammar, syntax and discourse features.

The analytic rating scale descriptors that the markers use reflect a clear hierarchy, or cline, of linguistic and discourse competence, ensuring that achieving a score tied to a higher CEFR level requires greater mastery of complex structures, a wider range of vocabulary, and more sophisticated organisation of ideas as test takers move from B1 to C2. Thus, while all test takers engage with the same task types, the expected level of linguistic range and control, syntactic complexity, and discourse management increases progressively at each successive CEFR level.

Building on the background presented above, it is reasonable to assume that texts produced by test takers are similarly graduated in terms of complexity and difficulty. Nonetheless, it is this construct that the current study examines and attempts to validate through a detailed exploration of test-taker lexical, syntactic and discourse level features. The premise underpinning the current study and driving the research question is, to reiterate the thesis just stated, that texts produced by test takers at higher CEFR levels (namely C1) will illustrate greater complexity than texts produced by test takers at lower CEFR levels (namely B1).

As a lead-in to the current study, an overview of research will now be provided as to how language proficiency supports the progression of linguistic complexity across CEFR levels. As an elaboration, some key research findings will first be provided under each heading.

Lexical Complexity

Lexical complexity is commonly analysed to assess linguistic ability. It generally encompasses several aspects of the vocabulary used in a text, including lexical diversity (the number of unique words), lexical density (the ratio of content words to function words), and lexical sophistication (the proportion of advanced words) (Crossley, 2020). Lexical richness and sophistication develop as learners move up the CEFR scale, and research into lexical diversity and lexical development has consistently reported that lexical complexity increases as learners progress from lower to higher proficiency levels – see González, 2017; Crossley et al., 2011; Zareva, 2007.

Read (2000) illustrates how higher CEFR learners (B2 and above) use more low-frequency words and academic or technical terms. This reflects Nation's model of vocabulary depth (2001), which shows that advanced learners develop not just more vocabulary but also a deeper understanding of word nuances, collocations, and connotations.

Syntactic Complexity

Syntactic complexity refers to the sophistication and variety of syntactic forms produced in language, and is considered a key aspect of L2 writing development and an indicator of more advanced writing skills. Research shows that syntactic complexity increases significantly as learners progress through the CEFR levels. There is considerable empirical evidence in the second language acquisition (SLA) literature of a strong link between the (syntactic) complexity of learners' L2 productions and their overall level of L2 development and proficiency (Ortega, 2003; Vyatkina, 2013; Wolfe-Quintero, et al., 1998).

In Martínez's (2018) analysis of syntactic complexity of writing at different proficiency levels, complexifications at sentence, clause and phrase levels of syntactic organisation revealed high correlations between scores and virtually all complexity metrics.

Discourse Complexity

Discourse complexity refers to the ability to construct and sustain coherent and cohesive extended texts. It encompasses the effective use of organisational patterns, logical sequencing of ideas, and appropriate cohesive devices, all of which have been shown to improve with increasing language proficiency.

Crossley and McNamara (2011) and Schiftner-Tengg (2022), for example, examined the use of discourse markers and connective devices as a sign of discourse complexity. They illustrate how learners at lower CEFR levels rely on simple cohesive devices while more proficient learners (B2-C2) use a broader range of discourse markers to create more nuanced and sophisticated text structures.

In the context of text structure and coherence, Weigle (2002) reported how C1-C2 level writers create more structurally complex texts, with a clear opening, development, and conclusion, using cohesive ties that guide the reader through the progression of an argument. Lower-level writers (A1-B1), in contrast, struggle with maintaining coherence over extended discourse and rely on simple, sequential structures.

At higher CEFR levels, learners show the ability to handle complex argumentation and extended discourse. In research into argumentation and elaboration, for example, Byrnes and Manchón (2014) demonstrated how B2-C1 learners maintained topic continuity and developed arguments over longer stretches of writing, employing higher-level discourse strategies such as counter-arguments and concessions.

Lu (2011) reported how higher proficiency writers use a wider range of clause types, more sophisticated discourse markers, and varied lexical choices. This

contrasts with lower-level learners, whose writing tends to be more formulaic, and syntactically simple.

In findings related directly to assessment, Green (2012) and Hawkins & Filipovic (2012) report how learners in higher CEFR bands consistently outperform lower-level learners in tasks involving the use of complex syntax, vocabulary, and extended discourse.

Research across syntactic, lexical, and discourse domains aligns with the CEFR's descriptions of language proficiency, confirming that as learners move through levels from A1 to C2, their ability to handle more complex grammar, richer vocabulary, and extended discourse improves. This progression is empirically supported by findings in SLA and assessment studies.

The Current Study

Building on the research reported above, the current study investigates test-taker performance at CEFR levels B1, B2, and C1 on the LANGUAGECERT Academic Writing Test (LCAWT).

The LANGUAGECERT Academic Writing Test

The LCAWT is a multilevel test designed to assess the ability to understand, use, and communicate effectively in English within an academic setting. The LCAWT lasts about 50 minutes, in which test takers are expected to produce two pieces of expository writing: one of 150-200 words and a second of around 250 words. While the LCAWT generates scores ranging from A1 to C2, its purpose is essentially to indicate readiness for tertiary-level study; consequently, the majority of test-taker grades emerge at B1, B2, and C1 levels. Three sample test-taker scripts from the second, longer, writing task are provided in Appendix C: a B1 test taker scoring 13/32; a B2 test taker scoring 19/32; and a C1 test taker scoring 27/32 in the task.

The Writing tasks are marked against rating scales aligned to the descriptors of the CEFR. These rating scales are: *Task Fulfilment; Accuracy and Range of*

Grammar; Accuracy and Range of Vocabulary; and *Organisation and Coherence*. Test-taker performance on each rating scale is rated on a nine-level scale (0–8), yielding a maximum possible score of 32 per task. To illustrate the *Accuracy and Range of Vocabulary* criterion, the Task 2 markscheme outlines specific qualitative descriptors at different score levels. For a mid-level mark, candidates "use a range of vocabulary, using simple items accurately and attempting more complex forms," with "some errors in usage/spelling/word formation which normally do not impede meaning, but can cause some re-reading." In the top band, candidates "use a wide range of vocabulary, including less-common items, with fluency and sophistication, and to give style," making "very few errors in usage or spelling which only occur as slips," and "can convey precise meaning." Such descriptors exemplify the increasing lexical sophistication and accuracy expected across proficiency levels.

Writing scripts are independently scored by two trained human markers via LANGUAGECERT's marking environment. Markers assess scripts electronically, entering scores for each criterion. The final score for each writing task is calculated as the average of the two markers' ratings across all four criteria. If a significant discrepancy arises between the markers' scores, the script is referred to a Chief Examiner, whose decision is final.

To ensure scoring consistency and validity within the multi-level framework, inter-rater and intra-rater reliability analyses identify markers whose scoring may be inconsistent, leading to retraining or closer monitoring. Chief Examiners also second-mark a random sample of 10% of scripts to ensure marking quality, alongside targeted re-marking of tests.

Sample

The sample for the current study comprises 2,746 scripts—1,373 scripts from Task 1 and 1,373 scripts from Task 2—produced by approximately 1,373 test takers who sat the LCAWT between mid-2023 and mid-2024. Given that most grades fall within the CEFR levels B1, B2, and C1, the scope of the analysis and discussion in the current study is limited to these three CEFR levels. The distribution of the productions between levels is presented in Table 1 below.

Table 1*: CEFR sample spread*

CEFR level	N of scripts
B1	1,131
B2	1,241
C1	374

Methodology and Analysis

The analysis in the current study involves the use of language feature analysis via the computational linguistic analytic tool *Coh-Metrix 3.0* (Graesser et al., 2004).

Coh-Metrix has come to be an accepted and well-established computational tool in the analysis of written text. It has been used in a variety of contexts; some of its uses in previous research are provided below.

- to assess college students' writing quality (Li & Liu, 2017; Riazi, 2016)
- to predict quality in argumentative writing (MacArthur et al., 2019)
- to investigate quality in the written output of both first- and second-language writers (Crossley & McNamara, 2011)
- to classify written assessments and automate text evaluation (McNamara et al., 2014; McNamara et al., 2015)
- to explore how textual features influence judgments of writing quality (Crossley et al., 2011)
- to explore the development of writing in novice writers (Shpit, 2022)
- to explore developments in text cohesion in writing (Wen, 2023)

Coh-Metrix indices appear as eleven sections, although the sections are not systematically arranged into groups which directly denote lexical, syntactic or discourse level features. Table 2 presents the current section order as it is laid out in Coh-Metrix 3.0 (Graesser et al., 2004).

Table 2: *Coh-Metrix sections*

Coh-Metrix section	Description
Text Descriptives	*Basic counts and rates (e.g., number of words, sentences, syllables, type-token ratio)*
Text Easability	*Factors derived from factor analysis, such as narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion*
Readability	*Traditional readability indices like Flesch Reading Ease and Flesch-Kincaid Grade Level*
Referential Cohesion	*Measures of how often words and concepts are repeated or overlap across sentences (e.g., stem overlap, noun overlap)*
Lexical Diversity	*Type-token ratio, word frequency, and other metrics of lexical variation*
Connectives	*Frequency and type of logical connectives (e.g., causal, temporal, adversative)*
L2 Readability	*Adapted readability indices aimed at second-language learners*
Syntactic Complexity	*Measures like mean number of words before the main verb, number of modifiers, and syntactic pattern densities*
Syntactic Pattern Density	*Specific syntactic structure frequencies (e.g., infinitive forms, passive voice)*
Word Information	*Word concreteness, imagability, familiarity, meaningfulness (from psycholinguistic norms)*
Latent Semantics Analysis (LSA)	*Measures of semantic similarity between sentences and across the text, based on LSA vector space*

Across the 11 sections, 108 measures, or subcategories, are reported.

As Latifi and Gierl (2021) note, incorporating all sections and subcategories would possibly 'overshadow' actual informativeness, while the relevance and informativeness of certain sections or subcategories of features are context-dependent; they consequently reduce the categories they use to those which provide the most information. A similar move has been taken in the current study, where the dataset has been reduced to nine sections comprising the 62

most informative subcategories. These are listed in Appendix A. Table 3 provides a summary. It should be noted that the sections have been rearranged such that lexical, syntactic and discourse level features have been grouped together.

Table 3*: Categories used in the current study*

Level	Categories	No.
Lexical	Word Information	11
N=18	Lexical Diversity	4
	Latent Semantics	3
Syntactic	Syntactic Complexity	7
N=15	Syntactic Pattern Density	8
Discourse	Text Easability	8
N=29	Referential Cohesion	10
	Situation Model	8
	Readability	3
	Total	62

With certain Coh-Metrix categories, a higher score is indicative of an easier text, while, with other categories, a higher score is indicative of a more complex text. Such detail is provided in the rightmost column of Appendix A.

Research Questions

The research question being pursued in the study may be framed from two perspectives.

To what extent do Coh-Metrix scores on the three CEFR levels of B1, B2, C1:

(1) increase across CEFR levels (B1 to C1) in categories associated with greater linguistic complexity?

(2) decrease across CEFR levels (B1 to C1) in categories associated with greater linguistic simplicity?

Statistical Analysis

Each of the 62 Coh-Metrix subcategories was subjected to a one-way ANOVA using the software JASP (JASP Team, 2024), comparing Coh-Metrix scores across the three CEFR levels (recoded as B1=1; B2=2; C1=3). Assumptions of normality and homogeneity of variances were checked prior to analysis. For ease of readability in each category in Appendix B, a visual colour-coding scheme is used. A consistent decrease in score from B1 to B2 to C1 indicates texts coded as linguistically 'easier' across levels. Conversely, a consistent increase in score from B1 to B2 to C1 indicates texts which are coded as linguistically 'more demanding' across levels.

In the context of the research questions above, this means that yellow scores across Coh-Metrix subcategories indicate more complex texts being produced from B1 to C2, while green scores across subcategories indicate easier texts being produced from B1 to C2. This coding and detailed results can be found in Appendix B.

Results

Table 4 presents a picture of the number of Coh-Metrix subcategories where scores followed the expected progression across CEFR levels (B1 to B2 to C1). The fourth column ("Directional Progression [B1 to B2 to C1]") indicates the number of subcategories showing this trend. The fifth ("Significance [p<.001]") identifies how many of these trends were statistically significant based on the one-way ANOVAs. The final column combines the results of the fourth and fifth columns, showing the number and percentage of subcategories in which both progression and statistical significance were observed.

Table 4: *Progression across the CEFR levels vis-à-vis Coh-Metrix categories*

Level	Category	Subcats	Directional Progression (B1 to B2 to C1)	Significance (p<.001)	Progression & Significance
Lexical N=18	Word Information	11	9	8	73%
	Lexical Diversity	4	4	4	100%
	Latent Semantics	3	2	2	67%
	Subtotal	18	15	14	77.8%
Syntactic N=15	Syntactic Complexity	7	4	3	43%
	Syntactic Pattern Density	8	7	4	50%
	Subtotal	15	11	7	46.7%
Discourse N=29	Text Easability	8	7	6	75%
	Referential Cohesion	10	10	10	100%
	Situation Model	8	4	3	38%
	Readability	3	2	2	67%
	Subtotal	29	23	21	72.4%
Grand total		*62*	*49*	*42*	*67.7%*

As can be seen, in 42 of the 62 subcategories (i.e., 67.7%), the direction followed the B1 to B2 to C1 progression and was statistically significant. This pattern was particularly pronounced in the *Lexical* domain, where 14 of 18 indices (77.8%) showed both progression and significance. The *Syntactic* domain showed a lower rate (46.7%), while the *Discourse* domain showed a strong alignment (72.4%), particularly in Referential Cohesion, where all 10 subcategories exhibited both progression and significance.

Discussion

A discussion is provided below on each level of delicacy, summarising the details of the different categories and the subcategories beneath them.

Lexical Level

The *Lexical* level explored features related to vocabulary sophistication and variation, across 18 Coh-Metrix subcategories.

In the *Word Information* category, eight of the eleven subcategories, including content word familiarity, concreteness, imagability, meaningfulness, and hypernymy for nouns and verbs, showed a significant increase from B1 to C1.

All four *Lexical Diversity* subcategories (e.g., type-token ratios and measures of textual and vocabulary density) also exhibited consistent and statistically significant upward trends.

In the *Latent Semantic* category, two of the three measures (e.g., sentence-to-sentence overlap), showed similar significant progression.

In total, 14 out of 18 Lexical level indices (77.8%) aligned with both the expected directional increase and statistical significance, pointing to a strong association between CEFR level and lexical development in the produced scripts.

Syntactic Level

The *Syntactic* level explored sentence structure and grammatical complexity of texts, across 15 syntactic subcategories.

Regarding *Syntactic Complexity*, three of the seven subcategories, including noun phrase modifiers, minimal edit distance (i.e., the average distance between syntactically related words), and sentence syntax similarity, showed a significant upward trend across CEFR levels.

As for *Syntactic Pattern Density*, five subcategories – adverbial, prepositional, gerund, infinitive and agentless passive voice density – followed the B1 to B2 to C1 progression and were significant.

Overall, 7 of the 15 *Syntactic* level indices (46.7%) demonstrated both directional progression and statistical significance. This lower rate compared to lexical and discourse measures may suggest that syntactic development in writing scripts is more gradual or variable, and may not be fully captured by surface-level syntactic features alone.

Discourse Level

The *Discourse* level explored how information is organised at the level of sentences and larger units, such as paragraphs and whole texts, as well as the effect of cohesion and coherence. There were 29 subcategories investigated at this level.

Concerning *Text Easability*, six of the eight measures, followed the B1 to B2 to C1 progression and were significant. These were: narrativity, word concreteness, referential cohesion, deep cohesion, verb cohesion and connectivity.

As regards *Referential Cohesion*, all ten measures exploring overlap across sentences in terms of nouns, arguments, stems, content words and anaphor followed the B1 to B2 to C1 progression and were significant.

With *Situation Model* category, three of eight indices, reflecting the extent to which texts support mental model construction, showed significant increases.

With *Readability*, two of three subcategories, including the Flesch-Kincaid Grade Level, also progressed significantly with CEFR level.

In total, 21 of the 29 *Discourse* level indices (72.4%) demonstrated both directional progression and statistical significance. These findings highlight that discourse-level features, particularly those related to cohesion and textual integration, are strongly associated with increases in writing proficiency across CEFR bands.

Conclusion

The current study explored the relationship between lexical, syntactic and discourse features and CEFR proficiency levels in the high-stakes LANGUAGECERT Academic Writing Test, focusing on texts rated at B1, B2, and C1. The purpose of the study was to provide external validation, or triangulation, of the assumption that higher CEFR-level grades reflect greater linguistic complexity, and thus, higher language ability in writing.

LANGUAGECERT's English language writing assessments are grounded in CEFR-aligned constructs that define expectations for lexis, grammar, syntax, and discourse features at each level. The associated rating scales similarly describe an increasing level of complexity across CEFR bands. It is therefore expected that performance should reflect this progression in measurable ways. The current study sought to validate this expectation by analysing linguistic features in test-taker scripts using the computational tool Coh-Metrix 3.0, a widely recognised instrument for text analysis.

The research hypothesis driving the study was that texts produced by test takers achieving higher CEFR levels (such as C1) would illustrate greater complexity than texts produced by test takers at lower CEFR levels (such as B1).

A total of 2,746 test-taker scripts, produced for both Task 1 and Task 2 across 62 Coh-Metrix subcategories spanning lexical, syntactic, and discourse domains, were analysed. One-way ANOVAs were conducted to examine differences across proficiency levels. Results showed that 49 out of 62

subcategories (79.0%) displayed a directional increase from B1 to B2 to C1, and 42 subcategories (67.7%) demonstrated statistically significant differences.

At the lexical level, 14 of the 18 measures (77.8%) emerged as following the B1 to B2 to C1 progression and as being significant.

At the syntactic level, 7 of the 15 measures (46.7%) emerged as following the B1 to B2 to C1 progression and as being significant.

At the discourse level, 21 of the 29 measures (72.4%) emerged as following the B1 to B2 to C1 progression and as being significant.

These findings provide considerable empirical support for the construct validity of the LANGUAGECERT Academic Writing Test: more proficient candidates produce texts that are measurably more complex, particularly in terms of vocabulary use and discourse cohesion.

The conclusion that may be drawn is that with over two thirds of lexical, syntactic and discourse level features affirming that higher-ability test takers do produce more complex texts than lower-ability test takers, the LANGUAGECERT Academic Writing test functions as intended, effectively distinguishing among levels of proficiency. While the notion might appear somewhat simplistic that a multilevel test should assign more able test takers to higher levels and lower-ability ones to lower levels, the current research confirms that this is the case. The thesis possibly also needs to be considered from the opposite perspective: if the categories did not in general follow the expected progression, this might well give cause for concern.

While the study provides strong validation evidence, it is limited to three CEFR levels and the written modality. Further research would aim to cover the whole gamut of the six-level CEFR scale and extend the analysis to include the speaking test. Additionally, while this study employed a quantitative approach, further research could complement it with qualitative analysis of linguistic features in candidate texts (e.g., through detailed discourse analysis or expert ratings), to explore how complexity manifests in more nuanced ways.

Overall, the findings from this large-scale analysis offer strong support for the construct validity of the LANGUAGECERT Academic Writing Test and reinforce its alignment with CEFR proficiency levels. By applying computational linguistic tools to authentic test-taker output, this study contributes to both the validation of high-stakes assessment and the broader understanding of language development. As demand grows for transparent, evidence-based language testing, this research demonstrates the value of integrating automated linguistic analysis into the validation process, helping ensure that test scores are both interpretable and meaningful for stakeholders.

Acknowledgement

We are grateful to the developers of Coh-Metrix for permitting us access to the Coh-Metrix 3.0 software.

References

Byrnes, H., & Manchón, R. M. (Eds.). (2014). *Task-Based Language Learning: Insights from and for L2 Writing*. John Benjamins Publishing Company. https://doi.org/10.1002/tesq.388

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3), 415-443.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(2/3), 170–191. http://doi.org/10.1504/IJCEELL.2011.040197

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29, 243–263. https://doi.org/10.1177/0265532211419331

González, M. C. (2017). The contribution of lexical diversity to college-level writing. *TESOL Journal*, 8(4), 899-919. https://doi.org/10.1002/tesj.342

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193–202. https://doi.org/10.3758/BF03195564

Green, A. (2012). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge University Press.

Hawkins, J. A., & Filipović, L. (2012). *Criterial FEATURES in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge University Press.

JASP Team (2024). *JASP* (Version 0.19.0)[Computer software].https://jasp-stats.org/

Latifi, S., & Gierl, M. (2021). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*, 38(1), 62-85. https://doi.org/10.1177/0265532220929918

Lu, X. F. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Li, X., & Liu, J. (2017). Automatic essay scoring based on Coh-Metrix feature selection for Chinese English learners. In T. Huang, R Lau, Y Huan, M Spaniol, & C Yuen. (Eds.), *International symposium on emerging technologies for education* (pp. 382–393). Springer. https://doi.org/10.1007/978-3-319-52836-6_40

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic

writers, and how do those features change with instruction? *Reading and Writing*, 32(6), 1553–1574. https://doi.org/10.1007/s11145-018-9853-6

Martínez, A. C. L. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, 35, 1-11. https://doi.org/10.1016/j.asw.2017.11.002

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.

Riazi, A. M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing*, 28, 15–27. https://doi.org/10.1016/j.asw.2016.02.001

Schiftner-Tengg, B. (2022). Analysing discourse coherence in students' L2 writing: Rhetorical structure and the use of connectives. In Berger, A., Heaney, H., Resnik, P., Rieder-Bünemann, A., Savukova, G. (Eds), *Developing advanced English language competence: A research-informed approach at tertiary level* (pp. 237-255). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-79241-1_23

Shpit, E. I. (2022). The use of Coh-Metrix by individual Russian novice writers for developing self-assessment and self-correction skills. *International Journal of English for Specific Purposes*, 3(1), 6-33.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus*. Modern Language Journal,* 97(S1), 11–30. https://doi.org/10.1111/j.1540-4781.2012.01421.x

Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press.

Wen, J. (2023). The changes in text cohesion of senior high school students measured by Coh-Metrix as a function of grade level. *International Journal of Social Science and Education Research*, 6(8), 109-119. https://doi.org/10.6918/IJOSSER.202308_6(8).0014

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, Hi: University of Hawaii, Second Language Teaching & Curriculum Center.

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization?. *Second Language Research*, 23(2), 123–153. https://doi.org/10.1177/0267658307076543

Appendix A: Coh-Metrix Categories and Subcategories Used in the Analysis

Table A1: *Lexical level*

		LEXICAL LEVEL	
No.	**Subcategory**	**Category and subcategory gloss**	**Higher score =**
		Word Information	
95	WRDFRQa	CELEX Log frequency for all words	Easier
96	WRDFRQmc	CELEX Log minimum frequency for content words	Easier
97	WRDAOAc	Age of acquisition for content words	More Complex
98	WRDFAMc	Familiarity for content words	Easier
99	WRDCNCc	Concreteness for content words	Easier
100	WRDIMGc	Imagability for content words	Easier
101	WRDMEAc	Meaningfulness, content words	Easier
102	WRDPOLc	Polysemy for content words	More Complex
103	WRDHYPn	Hypernymy for nouns	More Complex
104	WRDHYPv	Hypernymy for verbs	More Complex
105	WRDHYPnv	Hypernymy for nouns and verbs	More Complex
		Lexical Diversity	
48	LDTTRc	Lexical diversity, type-token ratio, content word lemmas	More Complex
49	LDTTRa	Lexical diversity, type-token ratio, all words	More Complex
50	LDMTLDa	Lexical diversity, MTLD, all words	More Complex

		LEXICAL LEVEL	
No.	**Subcategory**	**Category and subcategory gloss**	**Higher score =**
51	LDVOCDa	Lexical diversity, VOCD, all words	More Complex
		Lexical Semantic Analysis (LSA)	
40	LSASS1	LSA overlap, adjacent sentences	Easier
42	LSASSp	LSA overlap, all sentences in paragraph	Easier
46	LSAGN	LSA given/new, sentences	Easier

Table A2: *Syntactic level*

		SYNTACTIC LEVEL	
No.	**Subcategory**	**Category and subcategory gloss**	**Higher score =**
		Syntactic Complexity	
69	SYNLE	Left embeddedness, words before main verb	More Complex
70	SYNNP	Number of modifiers per noun phrase	More Complex
71	SYNMEDpos	Minimal Edit Distance, part of speech	More Complex
72	SYNMEDwrd	Minimal Edit Distance, all words	More Complex
73	SYNMEDlem	Minimal Edit Distance, lemmas	More Complex
74	SYNSTRUTa	Sentence syntax similarity, adjacent sentences.	More Complex
75	SYNSTRUTt	Sentence syntax similarity, all combinations, across paragraphs	More Complex
		Syntactic Pattern Density	
76	DRNP	Noun phrase density, incidence	More Complex
77	DRVP	Verb phrase density, incidence	More Complex
78	DRAP	Adverbial phrase density, incidence	More Complex
79	DRPP	Preposition phrase density, incidence	More Complex
80	DRPVAL	Agentless passive voice density, incidence	More Complex
81	DRNEG	Negation density, incidence	Easier
82	DRGERUND	Gerund density, incidence	More Complex
83	DRINF	Infinitive density, incidence	More Complex

Table A3: *Discourse level*

		DISCOURSE LEVEL	
No.	**Subcategory**	**Category and subcategory gloss**	**Higher score =**
		Text Easability Principal Components (PC)	
12	PCNARz	PC Narrativity	Easier
14	PCSYNz	PC Syntactic simplicity	Easier
16	PCCNCz	PC Word concreteness	Easier
18	PCREFz	PC Referential cohesion	Easier
20	PCDCz	PC Deep cohesion	Easier
22	PCVERBz	PC Verb cohesion	Easier
24	PCCONNz	PC Connectivity	Easier
26	PCTEMPz	PC Temporality	Easier
		Referential Cohesion	
28	CRFNO1	Noun overlap, adjacent sentences	Easier
29	CRFAO1	Argument overlap, adjacent sentences	Easier
30	CRFSO1	Stem overlap, adjacent sentences	Easier
31	CRFNOa	Noun overlap, all sentences	Easier
32	CRFAOa	Argument overlap, all sentences	Easier
33	CRFSOa	Stem overlap, all sentences	Easier
34	CRFCWO1	Content word overlap, adjacent sentences	Easier

		DISCOURSE LEVEL	
No.	**Subcategory**	**Category and subcategory gloss**	**Higher score =**
36	CRFCWOa	Content word overlap, all sentences	Easier
38	CRFANP1	Anaphor overlap, adjacent sentences	Easier
39	CRFANPa	Anaphor overlap, all sentences	Easier
		Situation Model	
61	SMCAUSv	Causal verb incidence	Easier
62	SMCAUSvp	Causal verbs and causal particles incidence	Easier
63	SMINTEp	Intentional verbs incidence	Easier
64	SMCAUSr	Ratio of casual particles to causal verbs	Easier
65	SMINTEr	Ratio of intentional particles to intentional verbs	Easier
66	SMCAUSlsa	LSA verb overlap	Easier
67	SMCAUSwn	WordNet verb overlap	Easier
68	SMTEMP	Temporal cohesion, tense & aspect repetition	Easier
		Readability	
106	RDFRE	Flesch Reading Ease	Easier
107	RDFKGL	Flesch-Kincaid Grade Level	More Complex
108	RDL2	Coh-Metrix L2 Readability	Easier

Appendix B: Analysis of Coh-Metrix Subcategories by CEFR Level

Colour coding key:

Green: Higher scores = easier, advanced writers should score lower

Yellow: Higher scores = more demanding, advanced writers should score higher

Table B1*: Lexical level*

	Subcategory	B1 mean	B2 mean	C1 mean	Significance	Progression & Significance
Word Information	WRDFRQa	3.10	3.06	2.99	0.001	√
	WRDFRQmc	0.87	0.94	0.83		
	WRDAOAc	371.81	375.26	378.25	0.001	√
	WRDFAMc	579.32	576.42	573.40	0.001	√
	WRDCNCc	375.46	374.92	370.65	0.02	√
	WRDIMGc	405.90	405.78	402.68	0.08	
	WRDMEAc	434.43	433.53	431.26	0.001	√
	WRDPOLc	3.70	3.70	3.67		
	WRDHYPn	5.90	6.13	6.39	0.001	√
	WRDHYPv	1.45	1.51	1.58	0.001	√
	WRDHYPnv	1.62	1.69	1.77	0.001	√
Lexical Diversity	**LDTTRc**	0.72	0.74	0.76	0.001	√
	LDTTRa	0.55	0.56	0.58	0.001	√
	LDMTLD	77.82	84.46	98.38	0.001	√
	LDVOCD	75.86	82.50	94.00	0.001	√
Latent Semantics						
	LSASS1	0.22	0.21	0.20	0.001	√
	LSASSp	0.20	0.19	0.18	0.01	√
	LSAGN	0.27	0.28	0.28		

Table B2: *Syntactic level*

	Subcategory	B1 mean	B2 mean	C1 mean	Significance	Progression & Significance
Syntactic Complexity	SYNLE	7.21	6.41	6.47		
	SYNNP	0.70	0.75	0.79	0.001	√
	SYNMEDpos	0.63	0.63	0.63		
	SYNMEDwrd	0.85	0.87	0.88	0.001	√
	SYNMEDlem	0.83	0.85	0.86	0.001	√
	SYNSTRUTa	0.08	0.09	0.10	0.07	
	SYNSTRUTt	0.08	0.08	0.08		
Syntactic Pattern Density	DRNP	411.45	402.70	392.42	0.001	
	DRVP	199.74	200.64	202.19	0.20	
	DRAP	31.92	32.40	34.02	0.001	√
	DRPP	115.76	121.32	119.74		
	DRPVAL	4.90	6.66	7.01	0.001	√
	DRNEG	7.10	6.48	6.44	0.07	
	DRGERUND	18.34	21.05	24.03	0.001	√
	DRINF	17.25	18.87	20.86	0.001	√

Table B3: *Discourse level*

	Subcategory	B1 mean	B2 mean	C1 mean	Significance	Progression & Significance
Text Easability	PCNARz	0.15	-0.07	-0.27	0.001	√
	PCSYNz	-1.31	-0.98	-0.95	0.001	√
	PCCNCz	0.23	0.19	0.14	0.001	√
	PCREFz	0.60	0.22	-0.20	0.001	√
	PCDCz	0.75	0.62	0.59	0.06	
	PCVERBz	0.60	0.37	0.02	0.001	√
	PCCONNz	-2.10	-2.19	-2.25	0.04	√
	PCTEMPz	-0.95	-0.97	-0.96		
Referential Cohesion	CRFNO1	0.58	0.53	0.49	0.001	√
	CRFAO1	0.68	0.64	0.61	0.001	√
	CRFSO1	0.64	0.61	0.58	0.001	√
	CRFNOa	0.54	0.50	0.46	0.001	√
	CRFAOa	0.64	0.60	0.57	0.001	√
	CRFSOa	0.61	0.58	0.56	0.001	√
	CRFCWO1	0.15	0.12	0.10	0.001	√
	CRFCWOa	0.13	0.11	0.09	0.001	√
	CRFANP1	0.37	0.32	0.31	0.001	√
	CRFANPa	0.20	0.14	0.13	0.001	√

	Subcategory	B1 mean	B2 mean	C1 mean	Significance	Progression & Significance
Situation Model	SMCAUSv	20.39	22.62	22.57		
	SMCAUSvp	33.71	35.57	33.73		
	SMINTEp	10.78	10.74	10.56	0.97	
	SMCAUSr	0.71	0.59	0.47	0.001	√
	SMINTEr	2.10	1.77	1.76	0.001	√
	SMCAUSlsa	0.11	0.10	0.09	0.001	√
	SMCAUSwn	0.50	0.51	0.49		
	SMTEMP	0.76	0.76	0.76		
Readability	RDFRE	52.27	52.33	46.68	0.001	√
	RDFKGL	13.13	12.12	12.89		
	RDL2	20.62	17.49	14.71	0.001	√

Appendix C: Sample Test-Taker Scripts

Task 2 Prompt

Read the following statement and write about the topic.

> The internet allows the individual to feel part of a global online community. Some people believe that this community can bring diverse people together and help solve global problems. Others disagree, fearing that it just leads to disagreements that drive people further apart.

Discuss both of these views and give your own opinion.

Write about 250 words.

About the Samples

The following are responses from candidates at three different CEFR levels (B1, B2, C1) in response to the same Task 2 prompt from the LANGUAGECERT Academic Writing Test. Each script is preceded by a set of identifying details in the following format:

<<Level – Total Raw Marks – Gender – Age – L1 – Word Count>>

<<B1 – 14 – Male – 22 – Chinese – 268>>

Nowadays, it is fact that the internet excatly enable us to become a part of the global online community. However, whether the community can bring diverse people together and help solve global problems or can leads to disagreements is a hot issues. My view is that, internet is a useful tool that can let our lifestyle be more efficiently.

Firstly, it is obvious that we can easily to communicate with each other online. This is because there are a large number of softwares that can let us easily to chat each other. Like, wechat, telegram, facebook and so forth. Therefore, all of us from different countries can easily chat each other.

Moreover, we can easily contact the large number of news from different countries online and deal with some global problems. For example, it is convenient for us to by using searching engines like google or watching youtube or bbc news. So, we can rapidly to get the latestly news on time.

That is not to say that internet is always great. There are, of course, disagreements that drive people further apart. For instance, sometimes we ignore chatting by face to face with our friends in the real society. It mean that we will probably lose the relationship between our friends. Therefore, we should improve our relationship with in our community.

In summary, we should correctly to use internet. Meanwile, we should use internet on a regular basis. And the end of the day. Only by doing so can we let the internet be efficient to server us. And we can be together and solve some global problems.

<< B2 – 21 – Male – 21 – Tamil – 250 >>

The internet is considered to be the *utopia* of the current generation. People tend to spend more time on the internet compare to the real-life. This creates a conflict-of-interest between different communities of people with different opinions. In this essay, I will discuss on both the views on the impact of internet on people.

Internet can be both a boon and bane to society. One of the advantages is that the people can connect with more people and create a network of communities based on their interests and opinions. This seems to be a great thing for the people who want to communicate with people from diverse background and exchange their opinions. Some people like this approach as they think that this approach broadens their minds on wide thoughts and learn some new things.

Eventhough, the internet feels like a wonderland, it also has some negative impact on the people, according to some communities. Some people feels that communication only through the internet, seperates the people apart. They think that the people will tend to lack in interactive skill to communicate with other people in real-life. This shows the different opinion of people with different reasons of beneficial approach. Although, there are always a solution for a problem.

In my opinion, people need to interact with other people in real-life, with less time spent on the internet. This approach can not only benefit the interactive knowlege of a person, but also diminishing the characteristic of introvertness in the future days.

<< C1 – 28 – Male – 44 – Greek – 313 >>

The internet is considered by many as the biggest technological advance of the last 30 years that affected most of the human's activities, including every-day communications.

Communications through internet have many forms; simple emails and chat rooms in the first years to instant multimedia messaging, social media, photo and video sharing and even virtual reality rooms in the so-called 'meta-verse'.

The fact that so many people can have instant communication with each other have obvious advantages, Everyone can have access to a large community of people having similar concerns or even better have solutions to similar problems. The bias that the traditional networks (tv, newspapers etc) may have can be overcome by community communications in a freely-run internet. The truth cannot be hidden by goverments or large capital organizations, so the people can easily organise and act as a union to a common goal. In politics, one can access a large audience without spending a serious amount of money or having relations to traditional media (tv, newspapers etc.), so we could consider that even the democracy becomes more accessible to anyone.

On the other hand, there are many disadvantages of the communications that the internet brought to our lives. Mostly because of the anonimity, usually the ideas spreading in the community have actually been the closest to the edges in the political spectrum, bringing people even more far away from each other. Ideas backed by fake news circulated in the unregulated internet have brought populist parties in many countries, inflating the problems that the people had.

Internet is a tool and as a tool it can be used to bring benefits or to do harm. My opinion is that the users of the internet should have a very good training from their schools and their families in order to be able to judge what is beneficiary and what is not and act accordingly.

Chapter 6: The Use Of Expert Judgement And Pairwise Comparison In The Establishment Of An Assessment Scale

David Coniam, Tony Lee and Leda Lampropoulou

Abstract

This paper explores the validation of an assessment by two complementary processes. Expert judgment assigns difficulty values to items on the basis of long-standing knowledge of as assessment scale; pairwise comparison establishes item difficulty values through statistical analysis. Given that expert judgement is accepted as the cornerstone of the accurate setting of test items at particular points on a proficiency scale, the accuracy of the expert judgement is crucial. In the study, the statistical pairwise comparison of test items provides a Rasch-calibrated, external perspective on item difficulty. Twelve experienced item writers rated a set of 23 multiple-choice items drawn from a large set of calibrated items via pairwise comparison principles. In the

analysis, good fit statistics emerged, with difficulty values very comparable to the original expert-proposed values. The study validates how 'expert' item writers can be seen to produce accurately-located items, and that faith may be placed in their judgements.

Keywords: expert judgment, pairwise comparison, CEFR, assessment

Introduction

In any assessment situation, particularly those with high-stakes, a number of factors may be seen to be associated with why test items need to be well set. Test items – and tests as a whole – need to be valid in that they assess what they intend to and how this reflects the target population. Items need to be assessing at the level they have been supposedly set for, to be reliable and to exhibit a lack of bias across different populations. The key to all these factors working harmoniously together lies in the hands of the test setters, the expert item writers whose task involves ensuring that test content is appropriate for target test-takers in terms of difficulty level, cultural context, and linguistic norms.

Against the above backdrop, 'expert judgement' in language assessment (see e.g., Coniam et al., 2022) has a long history as accepted practice in test development both in item writing and in the estimation of item difficulty – which in turn impacts level setting and cut scores. In the case of test setting, the use of the 'expert' is critical. In a study of minimally-trained item writers, Coniam (2009) reported such personnel as achieving a quality setting rate of only approximately 20%; i.e., items that may be defined as having good item statistics (see Falvey et al., 1994). A number of ground rules for the setting of good items was proposed by Haladyna & Downing (1989); many of these also appear in Alderson et al.'s (1995) discussion of the qualities needed of an "expert item writer". To be able to efficiently produce good tests – with good items and an accurate reflection of a given proficiency level – it is therefore clear that test item writers need to be both familiar and experienced with the test they are engaging in, as well as being well-trained.

Expert judgement in language assessment

Claims have long been made about the importance of trained, experienced test setters, item writers, expert judge standards setters (Davidson & Lynch, 2002). Generally, evidence has been provided that 'experts' judgement is accurate (Coniam et al., 2022; Fulkerson et al., 2009; Shin, 2022; Shiotsu, 2010).

Nonetheless, experts' judgements have been shown to not always be accurate in their perceptions of item or test difficulty (Alderson, 1993; Alderson & Kremmel, 2013; Elder et al., 2002). Alderson (1993), for example, found minimal difference in the predictions of item difficulty between experienced and inexperienced item writers. Elder et al. (2002) reported that expert prediction of speaking task difficulty did not always match actual task difficulty.

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

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

A long-used method for standard setting has been the Angoff method (1971), which requires (presumed) expert judges to visualise and estimate the performance of typical – often borderline – candidates on test items. Research has suggested that judge estimations of item performance are not always accurate (Impara & Plake, 1998; Plake & Impara, 2001). Ricker (2006) provides

a cogent review of Angoff methods, concluding that judges' ability "…to conceptualise and hold the concept of a 'minimally-competent candidate' over a prolonged period is a problem that is difficult to overcome."

The pairwise comparison methodology

While expert judgement provides an initial mapping onto an external assessment standard, the question remains whether experts' verdicts reflect the standard accurately. In other words, can the experts truly be judged to be experts?

In an attempt to overcome some of the issues associated with expert judgement, Heldsinger & Humphry (2010) explore a methodology of *pairwise comparison* of examination scripts. While the methodology dates to the work of Thurstone in the field of psychology in the early twentieth century (see Andrich, 1978), the use of *pairwise comparison* is comparatively new to the education field. Its use in language assessment was pioneered in the 1990s – see Pollitt & Crisp (2004).

In essence, the *pairwise comparison* methodology involves getting judges to simply decide, when examining a pair of test items (or scripts), which item (or script) is the better. In a pairwise comparison study, the issue is the 'relative' nature of the outcome. There is no external reference of what constitutes the item difficulty range. Once judgements have been made by many judges across many items, it is then possible to calibrate the items via Rasch, and hence locate the items on a common scale which goes from relatively easy to relatively difficult.

An early study involving the use of the pairwise comparison methodology in an educational context was that of Bramley et al. (1998) with mathematics and English language scripts. Bramley et al. (1998) concluded that applying the pairwise comparison method to judgements related to standards in terms of academic performance was indeed able to produce interpretable results. Wyatt-Smith et al. (2020) note that the pairwise comparison methodology is beginning to see increasing take-up and acceptance in the field of education.

As a starting point, Heldsinger & Humphry (2010) suggest 30 scripts as a workable sample size in terms of the number of scripts that may be reviewed. The core of the method involves judges simply comparing two scripts and deciding which is the better of the two. Such comparisons will ideally be conducted between all unique pairs of scripts in the sample. Once all comparisons have been obtained for all scripts by all judges, the comparisons are then subjected to a Rasch analysis. Such an analysis places all judges and scripts on the equivalent of a single ruler, on a scale that emerges from the analysis (Heldsinger & Humphry, 2010). In essence, the key concept in pairwise comparison is that the more times one script betters another in a comparison, the further apart the scripts will end up in terms of their locations on the scale. The ensuing analysis permits for an evaluation, albeit relative, of judge performance in terms of internal consistency and relative harshness.

A drawback of the methodology is that conducting all possible comparisons between even a comparatively small number of scripts still results in a very large number of comparisons. Rating all possible comparisons would be too much to ask of a judge. Crompvoets et al. (2020) advocate what might be termed a "reduced comparison" model. Rather than comparing every script against every other, a matrix can be constructed such that points of contact are set up between all scripts and all judges.

The current study

The current study utilised a set of 23 multiple-choice discrete reading and usage items. These items had been previously validated and established as test anchor items (Coniam et al., 2021a, 2021b). The items had also been used as calibrated items in a comparability study comparing reading and usage in LANGUAGECERT CEFR-based tests and China CSE-based tests (Coniam, Milanovic & Zhao, 2022).

As would be expected, an assessment body such as LANGUAGECERT, which offers a range of English language qualifications at various CEFR levels, utilises a substantial number of item writers. For the current study, item writers who had been setting objective-type test items for at least ten years were

approached to take part in the study. Twelve accepted, and these are the 12 judges who are the focus of the current study. All were native speakers of English; in their mid-40s to 60s; the median age was 55; all had been involved in English language teaching for at least 15 years; all had professional qualifications up to Masters level; all had been setting items for CEFR-related tests for at least 10 years. The male / female split was 5:7.

Given that all possible combinations of 23 items would result in too many comparisons to ask of judges, a matrix was constructed such that each judge had 43 comparisons to make, with each judgement only requiring approximately 20 seconds. Appendix A presents a sample of test items and the instructions (amended after an initial pilot) which were given to the judges. LANGUAGECERT setters attend regular standardisation sessions; the study was conducted by all 12 judges in mid 2023 as part of one such session.

Test data and the LANGUAGECERT LID scale

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

Table 1: *LID scale*

CEFR level	LID scale range
C2	151-170
C1	131-150
B2	111-130
B1	91-110
A2	71-90
A1	50-70

For analysis and calibration purposes with LANGUAGECERT tests, Rasch logit values are rescaled to a mean of 100 and a standard deviation (SD) of 20 (Coniam et al., 2021a). As mentioned, the 23 items used in the study have had LID scale values proposed by expert judgement, with values subsequently confirmed empirically (Coniam et al., 2021a).

By virtue of the nature of the study, the 23 items had no external reference point; they had no anchor values which could link the items or judges so that direct cross-calibrating might be conducted. The 23 items assessed via pairwise comparisons in the current study are nonetheless analysed using the same LID scale values and analytic process, i.e., a mean of 100 and an SD of 20.

The statistic used to analyse the data involved the pairwise analytic procedure available in the Rasch software Winsteps (Linacre, 2018). A brief outline of the Rasch measurement model is provided below.

The Rasch model

The use of the Rasch model enables different facets 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 (Wright, 1997). This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred as the 'logit') evenly spaced along the ruler. Second, once a common metric is established for measuring different phenomena, Rasch allows for different features to be examined and their effects monitored or controlled. In this context, Rasch Analysis may be seen as a preferred option to Classical Test Analysis statistics in that all facets – judges and items in the current instance – are calibrated onto a single unidimensional latent trait scale (Eckes, 2015).

A key issue in Rasch is that of the 'fit' of the data to the Rasch model; i.e., how well obtained values match expected values. The most widely used statistic is the *infit mean square* statistic. Infit may be seen as the 'big picture' in that it

scrutinises the internal structure of a facet. 1.0 indicates a 'perfect' fit in terms of obtained values matching expected values. Acceptable ranges of tolerance for fit range from 0.5 to 1.5 (Lunz & Stahl, 1990). High infit mean square values indicate rather scattered information within a facet, providing a confused picture about the exact placement of the facet. Very small infit values indicate minimal variation in the rating, providing too little information to make clear and meaningful judgments about the facet.

Research questions

The focus in the current study revolves the degree of match between expert judges' ratings of a set of items against Rasch-calibrated results of the same item-set resulting from pairwise comparisons. If good Rasch statistics and no significant differences are observed between the initial expert-proposed values and the values obtained from judges' pairwise comparisons of items, then the judge-based ratings and pairwise comparison ratings may be perceived as two complementary approaches for item calibration. That is, one, from a subjective judgement angle (judge ratings), and two, from an objective / assessment data-based angle (pairwise comparisons). While judge rating item calibration involves calibration without anchoring to a pre-existing assessment scale, pairwise comparison calibration results need an assessment scale to provide a reality of the assessment. The two processes – one subjective, one objective – may be seen to be complementary rather than adversarial. The current study provides an example of how the two approaches may work together.

The research question pursued in the current study involves establishing the equivalence of expert judgement and pairwise comparison results for the same set of items. The research question being pursued is:

> *How closely expert-based item locations map against pairwise-comparison-derived item locations.*

Analysis

Item quality

As mentioned, the 23 items had expert-proposed LID scale values which had been corroborated empirically. While these values are provided in Table 4, background detail on their calibration is not provided, but may be found in Coniam et al. (2021a).

The focus in the current section is therefore the analysis of the 23 items through the pairwise comparison method. The items are analysed in line with the parameters established for calibrating all LANGUAGECERT items (see Table 1 above). As mentioned, the LID scale ranges from 50–170, with 100 set as the mid-point and a standard deviation of 20.

Table 2 below first presents summary statistics obtained from the pairwise comparisons of the 23 items.

Table 2: *Summary statistics*

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	22.2	44.5	100.00	6.55	.96	-.11	.94	.08
SEM	2.6	.7	8.25	.26	.03	.16	.10	.15
SD	12.4	3.4	39.54	1.23	.16	.78	.46	.73
MAX.	48.0	54.0	178.03	10.46	1.23	1.12	1.93	1.53
MIN.	2.0	40.0	29.65	5.54	.63	-2.03	.27	-1.16
REAL RMSE	6.81	TRUE SD	38.07	SEPARATION	5.59	ITEM	RELIABILITY	.97

Fit statistics were good for the 23 measured items; reliability was high at 0.97. Infit and outfit figures were close to the optimal at 0.96 and 0.94. Standardised infit and outfit values were also small.

Detailed item quality statistics are provided in Appendix B: LID values, infit, PTMA (the Pearson point-measure correlation), facility (P) values and standard errors. In Winsteps, it should be noted, the item facet is negatively oriented.

As can be seen, LID values range across the whole LID scale range. This reflects the picture presented below of previous expert-suggested values.

With the exception of items 21 and 1, which had extreme values considerably outside the normal LID scale range, all items generally appeared to be well calibrated. Item standard errors were likewise within tolerable limits (below 5%) for all items apart from the same two items with extreme locations (21 and 1) which had the highest standard errors (6.95, 7.84).

Infit was good in that all values were within expected values. PTMAs were also high, all being above 0.3 again with the exception of items 21 and 1.

Pairwise comparison and expert-judgement results compared

Having examined background pairwise comparison item quality, the pairwise comparison results are now examined in the context of the expert-judgment-based LID scale values.

The LID scale was initially calibrated on the basis of expert-suggested anchor items (Coniam et al., 2021a). In a follow-up study (Coniam et al., 2021b), a large set of items were calibrated, following which calibration results were examined and reconfirmed by the experts. In this study, there was a broad 75% agreement between expert suggested and subsequent calibrated values.

Table 3 below provides expert-suggested and pairwise comparison values for the 23 items. As mentioned, pairwise comparison item values were negatively oriented in the analysis.

Despite being from different frames of reference, there is a clear relationship between the two sets of values. The correlation between the two datasets was high at 0.76 ($p<0.01$).

Table 3: *Expert judgement and pairwise comparison item values (sorted by pairwise comparison values)*

Item	Expert judgement values	Pairwise comparison values
21	136	33
20	167	42
23	130	45
22	127	47
19	136	66
15	126	74
16	119	76
17	123	83
9	143	86
8	131	90
6	113	96
12	131	108
13	126	110
18	127	117
4	116	119
11	91	126
2	127	131
14	101	132
5	117	136
10	94	136
3	114	141
7	85	144
1	92	176

To make the two datasets directly comparable and to bring values into a common frame of reference, values for both datasets were transformed to Z-scores. Table 4 contains the values from Table 3, the Z-scores, and, in the final column, the differences between the two sets of Z-scores. The data in the table are sorted by Z-score differences. A score range of within plus or minus two standard deviations is generally seen as acceptable since this range accounts

for 95% of a given distribution. In Table 4 below, Z-scores outside this range are highlighted.

Table 4: *Z-score transformations and differences (sorted by Z-score differences)*

Item	EJ LID values	PC LID values	EJ Z values	PC Z values	Z score difference
20	167	42	2.47	-1.53	4.0
21	136	33	0.82	-1.78	2.6
23	130	45	0.5	-1.46	1.96
22	127	47	0.34	-1.41	1.76
19	136	66	0.82	-0.89	1.71
9	143	86	1.19	-0.38	1.57
15	126	74	0.29	-0.69	0.98
8	131	90	0.56	-0.28	0.84
17	123	83	0.13	-0.46	0.6
16	119	76	-0.08	-0.64	0.56
12	131	108	0.56	0.2	0.36
13	126	110	0.29	0.24	0.05
18	127	117	0.34	0.44	-0.09
6	113	96	-0.4	-0.11	-0.29
2	127	131	0.34	0.79	-0.45
4	116	119	-0.24	0.48	-0.72
5	117	136	-0.19	0.92	-1.11
3	114	141	-0.35	1.05	-1.4
14	101	132	-1.04	0.81	-1.85
11	91	126	-1.57	0.67	-2.23
10	94	136	-1.41	0.92	-2.33
7	85	144	-1.89	1.14	-3.02
1	92	176	-1.51	1.97	-3.49

Key: EJ = Expert judgment; PC = Pairwise comparison

As can be seen in Table 4, only one item (item 20) had a Z-score larger than 2 and pairwise comparison value outside the 95th significance area.

Finally, to test whether the two sets of calibrations might be seen as being statistically equivalent, an equivalence t-test for independent samples was conducted. The equivalence independent samples t-test permits the null hypothesis to be tested 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.

The results are provided in Table 5.

Table 5: *Equivalence independent samples T-test*

	Statistic	t	df	p
Rating	Upper bound	9.564	44	< .001
	T-Test	9.557	44	< .001
	Lower bound	9.551	44	1

The equivalence t-test result indicates that the two sets of ratings are within the upper and lower bounds of the 95% equivalence region. On this basis, expert judgments and pairwise comparisons may be taken as equivalent.

Conclusion

This paper has reported on a study exploring item writer accuracy to validate expert setters' previous judgements using pairwise comparison. The pairwise comparison methodology may be viewed as a standalone method in that items are rated with no external reference point such as anchor values or CEFR levels.

The research question was framed as how closely expert-based item locations mapped against pairwise-comparison-derived item locations. Against this backdrop, the study was exploring the use of pairwise comparison with a view to exploring the accuracy of expert judgments in terms of determining item difficulty.

In the study, 12 experienced LANGUAGECERT item-writing judges each made 43 judgements about item difficulty drawn from a matrix of 253 item pairs compiled from 23 items. Item difficulty had previously been established for the 23 items, first by the expert judges and subsequently through empirical calibration of the items from having been administered to thousands of test takers.

In an initial analysis of the pairwise comparison data, good fit and reliability statistics emerged, with LID scale values that were not very different from the original expert-judge-proposed values.

Bringing both sets of data into a common frame of reference by converting them to Z scores resulted in only small differences between expert judgement and pairwise comparison Z score values. Finally, on an equivalence t-test, the two sets of ratings were seen to be within the upper and lower bounds of the 95% equivalence region – further indication that expert judgments and pairwise comparisons may be taken as equivalent.

While pairwise comparison helps to provide assessment data-based objective ordering of items in an assessment instrument via Rasch analysis, alignment with an external / pre-established assessment scale ideally needs to be observable. Similarly, if from another perspective, expert judgement, while helping to align items in an assessment instrument with an external / pre-

established assessment scale, needs to demonstrate that expert judgement decisions have an objective calibration. The current study has illustrated how the two methodologies (pairwise comparison and expert judgement) are complementary ways of fulfilling the requirements in establishing an assessment scale.

The results of the independent pairwise comparison study provide corroboration of the accuracy of the items produced LANGUAGECERT's item writers. The conclusion that may be drawn from the study is that LANGUAGECERT's item writers have a clear grasp of CEFR levels and that items which they set and propose difficulty levels for may be taken as accurate.

A possible limitation of the current study might be seen from the fact that only very experienced item writers participated. A range of experience in item writers might have resulted in weaker relationships between the expert and empirical values. This was not, however, the purpose of the current study. Rather, implications for the results of the current study lie in the importance of using experienced test setters, item writers who have extensive experience of the target levels and test takers. A side aspect of the current study may also be seen from the fact that the results in the study were obtained from the pooling of a number of experts. A single item writer setting test items on their own is not good practice. While there may never be a perfect fit between empirical values obtained through actual test administration and the subjective views of human item writers, the collective judgement of a number of test setters offers the possibility of the development of more accurately targetted test material.

References

Alderson, C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade in language testing* (pp. 46–57). TESOL.

Alderson, J.C., Clapham, C., Wall, D., (1995). *Language test construction and evaluation*. Cambridge University Press: Cambridge

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

Andrich, D. (1978). *Relationships between the Thurstone and Rasch approaches to item scaling*. *Applied Psychological Measurement*, 2, 449–460. [DOI: 10.1177/014662167800200405]

Angoff, W. H. (1971). *Scales, norms, and equivalent scores*. In R. L. Thorndike, (Ed.), *Educational Measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bramley, T., Bell, J. F., & Pollitt, A. (1998). *Assessing changes in standards over time using Thurstone's pairwise comparisons*. *Education Research and Perspectives*, 2, 1-23.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the impact of training in test development principles on improving test quality. *System*, 37(2), 226–242. https://doi.org/10.1016/j.system.2008.11.008

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., Milanovic, M., Pike, N. & Zhao, W. (2022). *The role of expert judgement in language test validation*. *Language Education & Assessment*, 5(1), 18-33. [DOI: 10.1080/26895269.2022.2074581]

Coniam, D., Milanovic, M., & Zhao, W. (2022). *Foreign CEFR and CSE comparability study: An exploration using the Chinese College English Test and the LANGUAGECERT Test of English*. *CEFR Journal – Research and Practice*, 5, 25-44.

Crompvoets, E. A., Béguin, A. A., & Sijtsma, K. (2020). *Adaptive pairwise comparison for educational measurement*. *Journal of Educational and Behavioral Statistics*, 45(3), 316-338. [DOI: 10.3102/1076998620904402]

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement*. Frankfurt am Main: Peter Lang.

Fulkerson, D., Mittelholtz, D. J., & Nichols, P. D. (2009). *The psychology of writing items: Improving figural response item writing*. In Annual meeting of the American Educational Research Association, San Diego, CA.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3(1), 49–68.

Heldsinger, S. & Humphry, S. M. (2010). *Using the method of pairwise comparison to obtain reliable teacher assessments*. *The Australian Educational Researcher*, 37(2), 1–20. [DOI: 10.1007/BF03216869]

Impara, J. C., & Plake, B. S. (1998). *Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method*. *Journal of Educational Measurement*, 35(1), 69-81. [DOI: 10.1111/j.1745-3984.1998.tb00534.x]

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

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

Plake, B. S., & Impara, J. C. (2001). *Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting*. *Educational Assessment*, 7(2), 87-97. [DOI: 10.1207/S15326977EA0702_2]

Pollitt, A., & Crisp, V. (2004, September). *Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions*. In BERA annual conference, UMIST Manchester, England.

Ricker, K. L. (2006). *Setting cut-scores: A critical review of the Angoff and modified Angoff methods*. *Alberta Journal of Educational Research*, 52(1), 53-64.

Shin, D. I. (2022). Item writing and writers. In *The Routledge handbook of language testing* (2nd ed., pp. 341-356). Routledge.

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, 8, 34–52. https://doi.org/10.1080/15434303.2010.536924

Wyatt-Smith, C., Humphry, S., Adie, L., & Colbert, P. (2020). *The application of pairwise comparisons to form scaled exemplars as a basis for setting and exemplifying standards in teacher education*. *Assessment in Education: Principles, Policy & Practice*, 27(1), 65-86. [DOI: 10.1080/0969594X.2019.1682652]

Appendix A: Sample of test items and instructions to judges

Compare each pair of items in Columns B/C and E/F, one pair at a time. Decide which item of the two is the more difficult (i.e., higher up the CEFR scale). You do not have to assign the items a CEFR level. Just decide which in the pair is the more difficult. There are 43 pairs to rate. Each decision should, however, only take 15-20 seconds or so.

In Column D, simply insert the number of the item (from Columns B or E) which you feel is the more difficult of the pair. The first item pair has been done as an example. (The same pair of items may occur than once. Don't worry; just continue to rate.)

Judge Name:

B	C	More diff?	E	F
37. One major benefit of the new system is the shortened manufacturing lead time, _____ making the company more efficient.	A. thereby B. aside C. regardless	37	32. Please note: the final day for _____ expenses is the 15th of each month.	A. submitting B. handing C. placing

B	C	More diff?	E	F
29. He was disappointed she didn't _____ more positively when he told her the news.	A. reflect B. regain C. react		32. Please note: the final day for _____ expenses is the 15th of each month.	A. submitting B. handing C. placing
30. The salary for my new job isn't as high as I'd hoped for but the _____ for promotion are excellent.	A. prospects B. expectations C. calculations		34. The graph on page 10 shows sales over the past four years and _____ for the next two.	A. probabilities B. progressions C. projections

Appendix B: Pairwise comparison item quality statistics (sorted by LID value)

Item	N	LID values	Infit	PTMA	P Value	SE
21	43	33	1.16	0.27	0.93	6.95
20	46	42	0.76	0.56	0.89	5.87
23	43	45	1.0	0.41	0.88	6.0
22	54	47	1.02	0.39	0.89	5.14
19	44	66	0.89	0.62	0.77	4.95
15	44	74	0.96	0.61	0.7	4.65
16	45	76	1.06	0.6	0.69	4.51
17	54	83	0.92	0.73	0.54	4.15
9	41	86	0.92	0.71	0.61	4.6
8	40	90	0.69	0.76	0.6	4.65
6	43	96	1.09	0.58	0.53	4.32
12	45	108	1.03	0.64	0.44	4.18
13	45	110	0.71	0.74	0.42	4.2
18	42	117	1.1	0.58	0.38	4.35
4	44	119	1.16	0.51	0.36	4.25
11	41	126	0.82	0.66	0.32	4.48
2	43	131	1.23	0.48	0.28	4.57
14	45	132	0.63	0.69	0.24	4.59
5	44	136	0.94	0.53	0.23	4.64
10	46	136	0.92	0.56	0.24	4.42
3	44	141	1.05	0.47	0.2	4.71
7	44	144	1.21	0.38	0.18	4.87
1	44	176	0.93	0.25	0.05	7.84

Key: SE: standard error; PTMA: Pearson point-measure correlation

Chapter 7: From Displacement to Stability: Iterative Anchoring in Rasch Calibration of the LTE Adaptive Reading and Listening Tests

David Coniam, Tony Lee and Leda Lampropoulou

Abstract

This paper reports an operational iterative anchoring procedure for equating items in the LANGUAGECERT Test of English adaptive Reading and Listening tests. Using expert-judged anchor values (Reading: 76; Listening: 36) across large item pools (Reading: 489; Listening: 442), we calibrated with Winsteps software and enforced a displacement rule: anchors with displacement greater than 0.5 logits were released and the scale re-estimated. Both skills stabilised after four iterations; 27 of 76 Reading anchors (35.5%) and 10 of 36 Listening anchors (27.8%) remained, with very satisfactory Rasch fit and reliability statistics confirming robustness. Anchor attrition reflects quality control rather than instability: only the most stable items are

retained to preserve the measurement frame. This process highlights the practical value of iterative anchoring in large-scale adaptive testing programmes.

Keywords: Rasch measurement, displacement, reading and listening tests, LTE, adaptive test

Rasch Common Item Anchoring

Equating two or more test forms to an existing measurement scale requires linkages between tests. In Rasch measurement, such linkages may be achieved via either common persons or common items, although the modus operandi is the latter, namely via anchor items (Aryadoust et al., 2020). Calibration involves estimating the locations of a set of items and a sample of persons on the probabilistic Rasch scale (Linacre, 2024), yielding a *frame of reference* (FOR) specific to the items, persons, and data used (Humphry & Andrich, 2008). Rasch (1977) referred to this property within Rasch measurement as *specific objectivity*, meaning that calibrated values of items are contextualised and objective within a particular FOR, but equating across tests requires stable anchors to preserve that objectivity.

The shifting of items in a test during the process of equating refers to the test as a whole (see Lee et al., 2022). If there is a common measurement scale involved, this serves as an underlying measurement scale not unlike the metre. The common measurement scale should not then be taken as an actual scale consisting of items. The tests mapping onto a common measurement scale have a common reference metric but need to remain distinct (Goodman, 1990).

In the above context, irrespective of how a calibration may be manipulated, the FOR – the ordering of items and persons resulting from the calibration – needs to be preserved (see e.g., Coniam et al., 2022). In the current paper, anchoring via common items is described, with the anchoring process undergoing as many iterations as necessary to resolve unacceptable displacements (see Linacre, 2024: 144). The displacement statistic indicates

how much an item's difficulty estimates shifts when it is treated as an anchor compared to when it is freely estimated (Stahl and Muckle, 2007).

Displacement is important since it reports the degree of match between the calibration of anchor items and original item locations, and can reveal misfit between the anchor items and the current sample. The set of anchor items selected for the test linking / equating after each iteration are therefore chosen such that the overall calibration results are not interfered with.

A small displacement typically suggests that the anchor item is functioning consistently, supporting its use for establishing a stable scale.

In contrast, a high displacement value – conventionally over half a logit (Wright & Douglas, 1976) – indicates that an item's difficulty estimate may not be stable across samples or conditions. It may indicate issues such as possible misalignment between an item's difficulty and the ability of the sample.

Since anchor items with high displacement values may unduly disturb the original calibration, they should then be unanchored, and a further calibration linking / equating conducted. Linacre (2024) states:

> *items with displacements of more than 0.5 logits, that are also bigger than the item S.E.s, are candidates for recalibration.*

The recalibration process may involve a number of iterations until all displacement values are, ideally, near zero. A further issue relates to how many anchor items are needed in the calibration - recalibration process. Linacre (2024) suggests that:

> *The percent of anchor items is less important than the number of anchor items. We need enough anchor items to be statistically certain that a large incorrect value of one anchor item will not distort the equating.*

While the concept of displacement is not new, there has been minimal discussion in the literature of how to operationalise it in practice. To achieve a steady state whereby no items have displacement values greater than the 0.5 logit maximum recommended may well require a number of iterations. The process is referred to in this paper as *iterative anchoring*.

In light of the agendas laid out above, the current study contributes by documenting the process of iterative anchoring in a large-scale operational setting. Specifically, we report on the LANGUAGECERT Test of English (LTE) adaptive Reading and Listening tests, administered in 2023–2024. Each test involved over 400 items and a comparatively large set of expert-judged anchors (76 in Reading, 36 in Listening). In essence, the paper illustrates how iterative anchoring ensures robust calibration in large adaptive language assessments, and how apparent anchor loss is an expected outcome of a deliberate stress-testing process.

LANGUAGECERT Test of English (LTE) Adaptive Test

An analysis of two tests – a Reading test and a Listening test – is described below. These were drawn from the LANGUAGECERT Test of English (LTE) adaptive test administered in 2023-2024.

The LTE adaptive test is a level-agnostic assessment of listening and reading which reports test-taker results from CEFR levels A1 to C2. The LTE draws from a series of item banks, each comprising approximately 1,000 items that include a range of listening and reading items and testlets of between two and five items.

The Listening test component comprises four sections targeting different reading sub-skills. These include understanding detailed information, following the sequential aspects of an exchange, comprehending longer spoken texts and, at the higher CEFR levels, appreciating speaker intention, inference, and summarising.

The Reading Test component comprises six sections, each focusing on different reading sub-skills. These include reading and understanding short notices, vocabulary use in context, lexico-grammatical awareness, and comprehension of longer texts which tap into different reading sub-skills depending on the level of the test taker (from understanding factual information at lower levels to recognising writer intention at higher levels.)

The LANGUAGECERT adaptive test algorithm functions such that all test takers are presented with 58 items – normally 28 listening and 30 reading items. The first item presented is at approximately B1 level. Test takers subsequently move dynamically between item types depending upon performance and predicted ability.

After having completed the 58 items, a test taker's level is determined at a specific point on the 100-point LANGUAGECERT Global Scale and its mapped CEFR level from A1 to C2. The data for this study is drawn from a dataset of more than one hundred thousand test takers who completed the LTE adaptive test in 2023-2024, both during pre-testing and live testing modes. It should be noted in passing that the LTE and its item bank have been calibrated jointly such that test takers receive the same grade whether they take the paper-based or the computer-adaptive version of the test.

Statistical Analysis and Baseline Descriptives

In the analysis reported below, calibrations were conducted via the software Winsteps (Linacre, 2024). The data matrix was calibrated with logit values rescaled to the LANGUAGECERT Item Difficulty (LID) scale mid-point of 100 and a spacing factor of 20 (see Pike et al., 2024). Table 1 provides detail on the number of items and anchors in the two tests – Reading and Listening – analysed in the current study.

Table 1: *Items and anchors in the LTE Reading and Listening tests*

Skill	Items	Anchors
Reading	489	76
Listening	442	36

The Reading test contained slightly more items and anchor items than the Listening test. Both tests, however, included large enough sets of anchor items to allow for their removal if displacement figures so indicated, while maintaining sufficient items for calibration purposes. For ease of exposition, most of the following discussion focuses on the Reading test, although the

performance of the Listening test is also outlined. Before proceeding to an analysis of the iteration process, baseline descriptive statistics for the Reading test are provided in Table 2.

Table 2: *Baseline Descriptive Statistics*

	S.E.	Infit mean squares	Outfit mean squares
N	489	489	489
Mean	2.04	0.99	0.99
Std. Deviation	1.21	0.08	0.15
Maximum	11.8	1.33	2.05
99th percentile	6.29	1.2	1.45
75th percentile	2.24	1.04	1.07
50th percentile	1.85	0.98	0.98
25th percentile	1.35	0.94	0.91
Minimum	0.54	0.75	0.47

Infit and outfit mean square values show that very few items fell outside the 0.5–1.5 acceptable range (Lunz and Stahl, 1990). More significantly, at the 99th percentile, standard errors were only just over half a logit (6.29 LID scale points), indicating that the data used in the calibration can be considered statistically robust.

Iterative Anchoring in Practice

The general procedure for iterative anchoring is as follows. A 'free' analysis – that is, one in which anchor values are not used – is first performed to verify that the new data are correct and contain no item miskeys, data-entry errors, miscoding, etc. This initial calibration is referred to as 'Iteration 0'.

Subsequently, a series of item-anchored analyses is conducted in which item displacements indicate the extent to which anchor values have 'drifted'. If any anchor item has a displacement greater than 0.5 logits, the item is unanchored and its value allowed to float freely along with all other items in the dataset (see Linacre, 2024).

The first analysis using anchor item values is termed 'Iteration 1'. Iterations continue until none of the anchor items display displacements greater than 0.5. Once this 'steady state' has been reached, the final item-anchored analysis is used for reporting reliability indices, fit statistics and other calibration metrics.

Research Questions

Two research questions are being pursued in this paper.

1. How many iterations are required for the set of anchor items to reach a 'steady state', defined as no anchor item having a displacement greater than 0.5 logits?
2. How many, or what percentage of, anchor items remain at the final iteration once this steady state has been achieved?

Iterations with the LTE Reading Test

In the discussion below, different facets of the iterative anchoring procedure are elaborated upon. The Reading test is the major focus in the analysis and discussion below since it contained slightly more items (489) than the Listening test as well as more anchor items to which expert-defined values had been assigned (76 against 36).

To arrive at a position whereby none of the 76 anchor value items had displacement values greater than 10 LID scale points, or half a logit, a number of iterations were required, as will be outlined.

For ease of interpretation, the tables below present the results for a limited set of 12 items (the first 12 items in the Reading test, as it happens) initially defined as anchor items. For these 12 items, the respective displacement and calibrated LID scale values are provided after each iteration. Table 3 presents the results of the initial iteration (Iteration 0). Here, it will be recalled, no anchor values were used; rather, items were allowed to float freely.

Table 3: *Iterative Anchoring – Iteration 0*

Item	Displacement 0	LID 0
R001	-0.1	120
R002	-0.1	123
R003	-0.1	125
R004	-0.1	117
R005	-0.1	120
R006	0	148
R007	0	170
R008	0	160
R009	0	160
R010	0	170
R011	-0.15	119
R012	-0.17	119

Displacement values can be seen in column 2 and LID values in column 3. Because the analysis is permitted to run as many as times as is necessary to converge, displacement values at this initial stage were small. No displacement exceeded five LID scale points, i.e., a quarter of a logit.

Next, the first actual anchoring iteration (Iteration 1) used the anchor values supplied for the 76 Reading test anchor items.

Table 4 below provides detail on the same 12 Reading test anchor items shown in Table 3 above. The difference this time is that anchor values were used, as indicated in Column 2, by the letter 'A' standing for 'anchor'. Column 2 states which items were set as anchor items (all 12, as can be seen). Column 3 lists displacement values, with those above half a logit (10 LID scale points) highlighted in red. Column 4 reports LID scale values incorporating displacement values.

Table 4: *Anchoring Iteration 1*

Item	Anchor 1	Displacement values 1	LID scale values 1
R001	A	-5.44	120
R002	A	6.7	123
R003	A	-26.63	93.28
R004	A	-11.8	99.91
R005	A	8.69	120
R006	A	20.71	170.67
R007	A	-7.9	170
R008	A	19.41	181.54
R009	A	4.27	160
R010	A	-1.01	170
R011	A	22.35	136.04
R012	A	0.04	119

As can be seen, five of the 12 items (i.e., those in red) had displacement values above half a logit. These items are subsequently unanchored and allowed to float freely.

Table 5 below presents an update to Table 4. The final column in Table 5 ("→ Anchor 2") shows which items were retained as anchors for the next iteration.

Table 5: *Anchoring Iteration 1 with Anchor 2s defined*

Item	Anchor 1	Displacement values 1	LID scale values 1	→ Anchor 2
R001	A	-5.44	120	A
R002	A	6.7	123	A
R003	A	-26.63	93.28	
R004	A	-11.8	99.91	
R005	A	8.69	120	A
R006	A	20.71	170.67	
R007	A	-7.9	170	A
R008	A	19.41	181.54	
R009	A	4.27	160	A
R010	A	-1.01	170	A
R011	A	22.35	136.04	
R012	A	0.04	119	A

Examination of the whole 489-item dataset revealed that 46 anchor items had unacceptably high displacement values. Consequently, the number of anchor items was reduced from 76 to 30. While unachoring over half of the original set of anchors may appear drastic, Rasch analyses often converge differently with varying items, and different items may subsequently present as unacceptable. Table 6 shows how anchoring Iteration 1 extended into Iteration 2 using the same 12 items.

Table 6: *Anchoring in Iteration 2*

	Iteration 1					Iteration 2		
Item	**Displace 1**	**LID 1**	**→ Anchor 2**		**Displace 2**	**LID 2**	**Anchor 2**	**→ Anchor 3**
R001	-5.44	120	A		-10.86	107.89	A	
R002	6.7	123	A		1.44	123	A	A
R003	-26.63	93.28			-0.01	91.93		
R004	-11.8	99.91			-0.01	98.57		
R005	8.69	120	A		3.42	120	A	A
R006	20.71	170.67			0	170.62		
R007	-7.9	170	A		-6.04	170	A	A
R008	19.41	181.54			0	181.49		
R009	4.27	160	A		6.16	160	A	A
R010	-1.01	170	A		0.96	170	A	A
R011	22.35	136.04			-0.03	134.63		

	Iteration 1				Iteration 2		
Item	Displace 1	LID 1	→ Anchor 2	Displace 2	LID 2	Anchor 2	→ Anchor 3
R012	0.04	119	A	-5.73	119	A	A

As can be seen, the five items (in red) from Iteration 1 which were allowed to float had, by Iteration 2, settled into a steady state of acceptability. In Iteration 2, however, item R001 reported a displacement value above half a logit, meaning that it needed to be unanchored for the subsequent Iteration 3. The final column ("→ Anchor 3") shows which items were retained as anchors for the following iteration, Iteration 3.

Across the full dataset, Iteration 2 showed that of the remaining 30 anchor items, only two showed unacceptable displacement values, resulting in the number of anchor items being reduced to 28.

In Iteration 3, one anchor item again showed a displacement above 0.5 logits and was unanchored, reducing the set to 27.

It Iteration 4, no anchor items reported displacements above the 0.5-logit threshold. At this point, the iteration process was therefore terminated.

To illustrate the iteration process visually, a schematic summary is provided in the Appendix. The figure shows how, after the first major iteration, a substantial adjustment – or "shakeout" - occurs, while subsequent iterations show only minimal further change. This pattern indicates that the anchoring process stabilised quickly after the initial iteration, with later iterations confirming the attainment of a steady state.

The same iteration process was repeated with the LTE Listening test items. The results are summarised below in Table 7, which provides a comparative picture of the calibration for both the Reading and Listening items.

Table 7: *Calibration of the Reading and Listening items*

Skill	Items	Anchors	Anchors removed after Iter^{n} 1	Anchors removed after Iter^{n} 2	Anchors removed after Iter^{n} 3	Anchors removed after Iter^{n} 4	Iterations Needed to full calibration	Good anchors
Reading	489	76	46	2	1	0	4	27/76
Listening	442	36	20	4	2	0	4	10/36

As shown, both Reading and Listening tests required four iterations for the items to be acceptably calibrated; by Iteration 4 no item displayed a displacement value greater than half a logit.

Discussion

This study has reported on the process of item calibration using anchor items within the LTE adaptive Reading and Listening tests administered in 2023-2024. The investigation focused on determining the number of iterations required for the set of anchor items to reach a 'steady state' – defined as no anchor item showing a displacement value greater than 0.5 logits. A secondary aim was to establish the number or the percentage of anchor items remained once this steady state had been achieved.

The LTE adaptive Reading and Listening tests were selected for analysis because both featured large item pools – each containing over 400 items – and relatively substantial sets of anchor items (76 for the Reading and 36 for the Listening test). These large numbers provided a suitable basis for exploring how successive iterations function in practice.

For the Reading test, four iterations were required, at the end of which 27 out of the 76(35.5%) of anchor items remained. A similar figure was recorded with Listening where 10 of the 36 (27.8%) anchor items remained after four iterations. Whether such percentages are typical or low will require future research. Nevertheless, the close alignment between the two skills, together with previous evidence that LANGUAGECERT's expert-defined anchor values are reliable (Coniam et al., 2024), supports the robustness of the present calibration process.

In addressing the question of the number of anchor items necessary for effective equating, the findings indicate that quality outweighs quantity. In other words, it is not simply the case that more anchor items produce better equating; rather the acceptability of the anchors depend on smaller displacement statistics. Pibal and Cesnik (2019: 3) state:

> *…while a higher number of anchors might help obtain a better estimate of the common scale, it is not always the quantity but rather the quality of the anchors that is decisive in tying different scales together.*

The current analysis reinforces this view. Sufficient anchor items to be provided ab initio to enable several iterations, but only those demonstrating stability should remain. As long as a sufficient number of acceptable anchor items remain at the end of the series of iterations across the range of item difficulties, the resulting calibration can be regarded as robust and defensible.

Conclusion

The study demonstrates that iterative anchoring offers an effective, evidence-based approach to test equating within adaptive language assessments. By progressively identifying and removing unstable anchors, the process safeguards the integrity of the Rasch frame of reference and ensures that only stable items contribute to the common measurement scale.

The results underline two key principles. First, the iterative process is not a sign of instability but a form of systematic quality control. Second, the goal of anchoring is to align the *test as a whole* rather than to preserve specific item

parameters. The final anchor set should span the full range of item difficulties while maintaining the overall person–item distribution pattern.

In sum, iterative anchoring enables LANGUAGECERT to maintain a fair, transparent, and statistically sound calibration framework, ensuring that test scores remain directly comparable across administrations and delivery modes.

References

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Coniam, D., Lee, T., & Lampropoulou, L. (2024) The use of expert judgement and pairwise comparison in the establishment of an assessment scale. *Journal of Applied Linguistics, 37*. https://doi.org/10.26262/jal.v0i37.10392

Coniam, D., Lee, T., Milanovic, M., Pike, N., & Zhao, W. (2022). Role of expert judgement in language test validation. *Language Education & Assessment, 5*(1), 18-33. https://doi.org/10.29140/lea.v5n1.769

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. http://doi.org/10.2307/271088

Humphry, S. M., & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement, 9*(3), 249-264.

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. http://doi.org/10.46451/ijts.2022.01.12

Linacre, J. M. (2024). *Winsteps® Rasch measurement computer program User's Guide.* Version 5.8.3. Portland, Oregon: Winsteps.com.

Lunz, M., & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*(4) 425-444. https://doi.org/10.1177/0163278790013004

Pibal, F., & Cesnik, H. S. (2019). Evaluating the quantity-quality trade-off in the selection of anchor items: A vertical scaling approach. *Practical Assessment, Research, and Evaluation, 16*(1), 6. doi: https://doi.org/10.7275/nncy-ew26

Pike, N., Papargyris, Y., Dourda, C., Lee, T., & Coniam, D. (2024). *Recalibrating and extending the analysis of the LANGUAGECERT Test of English*. London, UK: LANGUAGECERT.

Rasch, G. (1977). On Specific Objectivity an Attempt at Formalizing the Request for Generality and validity of Scientific Statements. *Danish Yearbook of Philosophy, 14*(1), 58-94. https://doi.org/10.1163/24689300-01401006

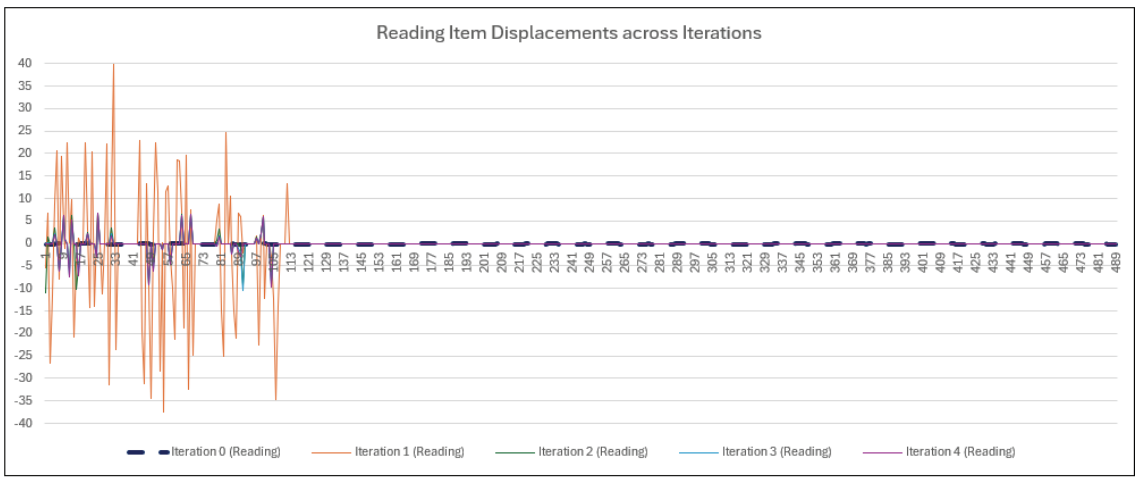
Stahl, J., & Muckle, T. (2007). Investigating drift displacement in Rasch item calibrations. *Rasch Measurement Transactions, 21*(3), 1126-1127.

Wright, B. D., & Douglas, G. A. (1976). Rasch item analysis by hand. *Research Memorandum 21*. Chicago, IL: University of Chicago.

Appendix: Schematic representation of the four iterations

The figure provides a schematic representation of the four iterations for the LTE Reading Test. The 76 anchor items were positioned among the first 113 items.

Iteration 0 is represented by dark blue markers along the horizontal zero line across all 489 items. **Iteration 1** (orange lines) shows variations extending up to +40 (two logits) and –30 (1.5 logits) LID scale points. **Iteration 2** (green lines) shows much smaller variations, remaining within ±10 LID scale points. **Iteration 3** (blue lines) also remains within ±10 LID scale points. **Iteration 4** (purple lines) similarly remains within ±10 LID scale points. After the first major iteration, a noticeable adjustment occurs, while subsequent iterations show minimal movement. This demonstrates that the process quickly converged to a stable configuration.



SECTION 3: COMPARATIVE STUDIES & LEARNING CONTEXTS

Chapter 8: A Construct-Based Comparison of LANGUAGECERT Academic and IELTS Academic

Leda Lampropoulou, Michael Milanovic, Johnathan Jones and Anthony Green

Abstract

This study examines the extent of comparability between LANGUAGECERT Academic and IELTS Academic, two assessments designed to measure English language proficiency for use in academic contexts. Adopting a construct-based approach, the analysis focuses on the alignment of task design, input characteristics, and scoring frameworks across the four language skills. Detailed comparisons are conducted for Speaking, Listening, Writing, and Reading, considering the extent to which each test elicits comparable language abilities and engages similar cognitive processes.

Findings indicate a high degree of alignment in the constructs targeted by the two tests, particularly in relation to the assessment of academic language proficiency and the underlying processes required for comprehension and production. At the same time, differences are identified in task structure,

target language use domains, and scoring and reporting practices. These differences do not appear to undermine overall comparability, but may influence how performance is elicited and interpreted in specific contexts. The study contributes to the validity argument underpinning comparisons between the two tests and provides a principled basis for the interpretation of concordance outcomes.

Keywords: Concordance study, score comparability, LANGUAGECERT Academic

Introduction

The interpretation and use of English language proficiency test scores in academic and migration contexts often require comparisons across different assessments. Institutions, policymakers, and test takers are frequently faced with decisions that involve evaluating results from tests that are designed independently but intended to measure similar aspects of language ability. Such comparisons, however, are only meaningful to the extent that the tests in question demonstrate sufficient alignment in the constructs they measure and the domains of language use they target.

Concordance studies aim to establish relationships between scores on different language assessments, supporting informed decision-making regarding test interpretation and use. These studies provide empirical evidence of the extent to which scores on one test may be associated with scores on another. While such evidence is valuable, it rests on an underlying assumption that the tests being compared are sufficiently similar in terms of construct, task design, and scoring. Without this foundation, statistical relationships between scores may be difficult to interpret meaningfully.

A key distinction in this context is that between equating and concordance (Pommerich & Dorans, 2004). Equating applies when two tests measure the same construct with comparable reliability and can therefore be used interchangeably. Concordance, by contrast, is used when tests are built to different specifications and measure similar, but not identical, constructs. In such cases, linked scores are not expected to be interchangeable, but rather

to provide indicative relationships that support interpretation. This distinction highlights the importance of examining construct comparability as a prerequisite for any concordance analysis.

A substantial body of research has investigated statistical relationships between major English language tests, typically reporting moderate to strong correlations. However, less attention has been given to detailed comparisons of test content, task characteristics, and scoring frameworks. These elements are central to understanding how language ability is operationalised and assessed, and therefore to evaluating the extent to which comparisons between tests are justified.

The present paper examines the extent of comparability between LANGUAGECERT Academic (LCA) and IELTS Academic, two assessments designed to measure English language proficiency for academic purposes. Both tests aim to evaluate candidates' ability to use English in contexts relevant to higher education, while differing in their design, task types, and scoring approaches. Establishing the degree of alignment between these tests at the level of construct and content is therefore essential in supporting the interpretation of their relationship.

The study adopts a dual approach to comparability, combining qualitative and quantitative analyses. The qualitative component focuses on comparisons of constructs, task demands, and scoring criteria across the four language skills, while the quantitative component examines statistical relationships between test taker performances on the two assessments. The full concordance study, including the quantitative analyses, is reported in a technical report (Lampropoulou et al., 2024), and in a separate paper focusing on score linking.

The present paper reports the qualitative findings, addressing the following research question: To what extent do LANGUAGECERT Academic and IELTS Academic demonstrate comparability in terms of construct, task design, and scoring frameworks?

Summary of LANGUAGECERT Academic and IELTS Academic tests

LANGUAGECERT Academic

The LANGUAGECERT Academic test is an assessment option for test takers seeking to study in English-medium, higher education settings. The test takes approximately 2.5 hours to complete and consists of four parts corresponding to four language skills: Listening (40 minutes), Reading (50 minutes), Writing (50 minutes), and Speaking (14 minutes). The test is designed to assess between the B1 and C2 levels of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001). Multiple task types are included to reflect the range of skills used in academic environments, and multiple response types are used for each skill, helping to ensure that no one response type will dominate and potentially bias results (Taylor & Chan, 2015).

The test content analysis measures competences appropriate for academic study in English-medium programmes, such as reading and listening for gist, or detailed understanding of a range of written and audio sources (e.g., academic articles, lectures, podcasts, interviews, and discussions). Additional competences include writing reports, articles and essays for an academic purpose; giving presentations; reading aloud; and taking part in a discussion. Test takers receive a score for each skill (Listening, Reading, Speaking and Writing) along with an overall score and an indication of their CEFR level. In addition, the test taker receives their scores on the LANGUAGECERT Global Scale, a scale which ranges between 0–100 and is aligned to the six levels (A1–C2) of foreign language proficiency described in the CEFR. The alignment between LANGUAGECERT Academic scores on the Global Scale and CEFR levels has also been externally validated (Ecctis, 2023).

The LANGUAGECERT Academic test has been designed for computer-based administration. However, for the concordance study reported here, the administration was initially paper based, as this allowed us to apply strict security measures while operational systems were under development. For

the second stage of the study, participants were offered the computer-based version. Regardless of mode of administration, like IELTS, the Speaking component was carried out live with an interlocutor. An equivalence study examined the results obtained on paper-based and computer-based versions of LANGUAGECERT Academic, indicating that the choice of mode did not have a meaningful impact on results.

IELTS Academic

The IELTS Academic test is an established assessment of English language proficiency for individuals who plan to study in an English-speaking setting. The test takes approximately 2 hours and 45 minutes and evaluates the test taker's ability to use and understand English at an academic level. The test is designed to assess test takers between the A1 and C2 levels of the CEFR. It has four modules that assess Listening (30 minutes), Reading (60 minutes), Writing (60 minutes) and Speaking (11–14 minutes). The IELTS Academic and General Training tests are only partly distinct: they have different Reading and Writing components but have the same Listening and Speaking components. In other words, for Speaking and Listening, both the General and Academic tests share the same format and content and the test materials balance more general with more academic English features.

The Listening module requires test takers to listen to recordings of clear, intelligible speech and answer related comprehension questions. The Speaking module involves a face-to-face interview with an IELTS examiner to assess the test taker's English communication skills. The Reading module involves reading three texts and answering comprehension questions. The Writing module consists of two tasks, one involving the description and comparison of data, and the other requiring an essay in response to a prompt. The test taker's overall score on the IELTS Academic test ranges from 0 to 9, with 9 representing the highest level of proficiency. A score is awarded for each skill module as well as for overall performance across all four test modules.

IELTS is available in paper-based and computer formats. For those taking IELTS on computer, the Reading, Writing and Listening modules are completed on a computer, but the Speaking module is conducted face-to-face with an IELTS examiner. For the computer-based test, Speaking is done on the same day as the other components. With the paper version, there can be a 7-day delay between taking the Speaking test and the other modules. Test takers in the concordance study could select their preferred mode of delivery but were not allowed to sit an online proctored exam.

Reporting results and feedback to participants

Results and feedback on performance can have a direct influence on teaching and learning. Similar to the IELTS test, results for LANGUAGECERT Academic are reported both as an overall score and as a score for each of the four language skills. This profile of scores is intended to help language learners to identify areas of strength and areas for improvement. Unlike IELTS, LANGUAGECERT also offers feedback on the productive skill sections (i.e., Speaking and Writing). For Speaking, feedback is given on Task Fulfilment and Communicative Effect, Coherence, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, and Pronunciation, Intonation and Fluency. Feedback on Writing covers Task Fulfilment, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, and Organisation.

Providing more detailed information for feedback can help prompt learners to reflect on their performance (Chapelle et al., 2015) and can provide greater opportunity for learners to learn from their current performance so they may move toward their desired level of performance (Lam, 2021). Feedback can promote self-regulation (Mežek et al., 2022) and can enhance cognitive and emotional engagement (Mayordomo et al., 2022).

Content comparisons

This section reports the content comparisons made between the LANGUAGECERT and IELTS Academic tests. Sections are subdivided by language skill (i.e., Speaking, Listening, Reading, Writing).

Speaking comparison

The Speaking components of LANGUAGECERT Academic and IELTS Academic were compared in terms of task design, interactional features, and scoring frameworks, with particular attention to the extent to which they elicit comparable aspects of spoken language ability. For IELTS, there is a single Speaking module used in both the Academic and General Training versions of the test. In contrast, the LANGUAGECERT Academic Speaking component has been developed specifically for academic purposes.

The following table presents a structured comparison of the key characteristics of the two tests. The analysis considers the structure of tasks, the integration of language skills, and the criteria used to evaluate performance, in order to examine how speaking ability is operationalised in each test.

Table 1: *Speaking test comparison of LANGUAGECERT Academic and IELTS (Academic and General Training)*

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
Target Level	B1-C2 / Markscheme covers performance at A1-C2 levels	A1-C2 / Markscheme covers performance at A1-C2 levels
Skills focus	Tasks are designed to elicit speaking skills such as communicating opinions and ideas on a variety of topics (e.g., study skills, news, daily life) and giving personal opinions on contemporary issues. Test takers will also demonstrate their ability to communicate (targeting higher education settings) using a range of functional language to elicit or respond as appropriate, to show the ability to use a wide range of language functions and use of register, to read aloud and answer questions, and to prepare and deliver a presentation in response to a visual stimulus and answer subsequent questions.	Tasks are designed to elicit speaking skills such as communicating personal information, expressing and justifying opinions, explaining, suggesting, speculating, expressing preferences, comparing, summarizing, and narrating.
Skill integration	LANGUAGECERT Academic Speaking entails an integration of speaking, listening, reading, and writing modalities. An examiner (interlocutor) orally explains the tasks and asks questions, requiring the test	IELTS Academic Speaking entails an integration of speaking, listening, reading, and writing modalities.

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	taker to listen and respond appropriately. In Part 3, the test taker must read a short passage and answer questions from the examiner, and in Part 4 test takers must discuss a visual stimulus, such as a chart. In this task, test takers are able to take notes in preparation for the task, engaging writing skills.	
Task description	The exam is delivered in person at a distance by the interlocutor. Speaking tests are recorded. **Task 1**. (Initial exchange) The examiner introduces himself/herself and confirms the test taker's identity. Test takers give and spell their names and give their country of origin. The examiner then asks up to five general questions on different topics which are expected to be familiar to test takers, such as study skills or daily life. Questions are scaffolded in advance in the examiner's script sheet. After the test taker replies, the examiner responds and/or comments briefly and thanks the test taker before moving to Part 2.	The Speaking Test consists of an oral interview between the test taker and an examiner. Speaking tests are recorded. **Task 1**. (Introduction and interview) The examiner introduces himself/herself and checks the test taker's identity. Then the examiner asks the test taker general questions on some familiar topics such as home, family, work, studies, interests. To ensure consistency, questions are taken from a scripted examiner frame. This part of the test focuses on the test taker's ability to communicate opinions and information on everyday

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	Task 2. (Role play) Two situations are presented by the examiner (interlocutor) and test takers are required to respond to and initiate interactions. The examiner begins by explaining that this part of the test is a role play and that the test taker is expected to either start or respond to a situation. The examiner then selects the first role play situation from a prepared list on the examiner sheet, and the interaction continues for approximately two turns before stopping and the next situation is introduced. The examiner selects the second topic from a second list of topic options provided on the examiner sheet. Approximately two turns are given for each interaction, and if there is additional time, an additional situation may be introduced. **Task 3**. (Read aloud and discussion of passage) The examiner provides the test taker with a Task Sheet which contains a reading passage of	topics and common experiences or situations by answering a range of questions. **Task 2**. (Long turn) The examiner gives the test taker a task card which asks the test taker to talk about a particular topic, includes points to cover in their talk and instructs the test taker to explain one aspect of the topic. Test takers are given 1 minute to prepare their talk, and are given a pencil and paper to make notes. Using the points on the task card effectively, and making notes during the preparation time, will help the test taker think of appropriate things to say, structure their talk, and keep talking for 2 minutes. The examiner asks the test taker to talk for 1 to 2 minutes, stops the test taker after 2 minutes, and asks one or two questions on the same topic. Part 2 lasts 3-4 minutes, including the preparation time. This part of the test

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	approximately 100 words. The examiner allows 30 seconds of preparation time and asks the test taker to read the text out loud. The examiner further explains that afterwards, the examiner will ask the test taker some questions about the topic. One or more questions, taken from a list on the examiner sheet, is then asked dependent upon time. **Task 4**. (Presentation) The examiner explains that the test taker will now be asked to give a presentation based on a visual stimulus (e.g. a chart or graph). The test taker has 60 seconds of preparation time, and then must talk about a topic provided by the interlocutor for two minutes. Once the test taker has presented for two minutes or has finished their presentation, the examiner asks follow-up questions for the remaining time.	focuses on the test taker's ability to speak at length on a given topic (without further prompts from the examiner), using appropriate language and organising their ideas coherently. It is likely that the test taker will need to draw on their own experience to complete the long turn. **Task 3**. (Discussion) The examiner and the test taker discuss issues related to the topic in Part 2 in a more general and abstract way and – where appropriate – in greater depth. This part of the test focuses on the test taker's ability to express and justify opinions and to analyse, discuss and speculate about issues.
Timing	14 minutes approximately (Part 1: 3 minutes; Part 2: 2 minutes; Part 3: 4 minutes; Part 4: 5 minutes)	11-14 minutes (Part 1: 4-5 mins; Part 2: 3-4 mins; Part 3: 4-5 minutes)

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
Scoring and weighting	Examiners award a raw score of up to 48 points: Task Fulfilment and Communicative Effect (8 points x 2), Coherence (8 points), Accuracy and Range of Grammar (8 points), Accuracy and Range of Vocabulary (8 points), and Pronunciation, Intonation and Fluency (8 points). The exam is delivered in person at a distance by the interlocutor. All tests are recorded. The interlocutor awards the marks for Task Fulfilment and Communicative Effect in real time. A second examiner listens to the recording and awards marks for the other criteria. Discrepancies are checked by a chief examiner. The criterion for Task Fulfilment and Communicative Effect is double weighted. **Task Fulfilment and Communicative Effect** A measure of the ability to manage the tasks adequately for the level and to communicate successfully with flexibility and naturalness.	Examiners award a band score for each of four criterion areas: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy and Pronunciation. The four criteria are equally weighted. Scores are reported in whole and half bands. Detailed performance descriptors have been developed which describe spoken performance at the nine *IELTS* bands. **Fluency and Coherence** The ability to talk with normal levels of continuity, rate and effort and to link ideas and language together to form coherent, connected speech. The key indicators of fluency are speech rate and speech continuity. The key indicators of coherence are logical sequencing of sentences, clear marking of stages in a discussion, narration or argument, and the use of cohesive devices (e.g. connectors, pronouns

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	Coherence A measure of the ability to provide coherent responses, particularly over extended speech, and the linking of ideas and contributions. **Accuracy and Range of Vocabulary** A measure of the ability to vary and demonstrate control of lexis and register as appropriate to the task. **Accuracy and Range of Grammar** A measure of the ability to vary and demonstrate control of grammatical structures as appropriate to the task. **Pronunciation, Intonation and Fluency** A measure of the ability to produce the sounds of English in order to be understood with appropriate stress and intonation and maintain the flow of speech. **Rating**	and conjunctions) within and between sentences. **Lexical Resource** This criterion refers to the range of vocabulary the test taker can use and the precision with which meanings and attitudes can be expressed. The key indicators are the variety of words used, the adequacy and appropriacy of the words used and the ability to circumlocute with or without noticeable hesitation. **Grammatical Range and Accuracy** This criterion refers to the range and the accurate and appropriate use of the test taker's grammatical resource. The key indicators of grammatical range are the length and complexity of the spoken sentences, the appropriate use of subordinate clauses, and the range of sentence structures, especially

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	For Speaking, LANGUAGECERT uses two raters. The interlocutor awards marks on 'Task Fulfilment and Communicative Effect' as they are communicating directly with the test taker. The other criteria are marked remotely by a second examiner. Collecting marks from two different examiners allows for information to be gathered on performance from two perspectives. The interlocutor/assessor is directly involved in the interaction with the test taker and is therefore in a strong position to judge task completion and communicative effect (also helping prevent prepared answers); however, the interlocutor must concurrently focus on test administration. Having a second examiner who can listen to a recording asynchronously permits greater time and focus to evaluate the more 'analytical' criteria of coherence (i.e., grammar, vocabulary, and pronunciation, intonation and fluency). Enabling examiners to focus on their individual	to move elements around for information focus. The key indicators of grammatical accuracy are the number of grammatical errors in a given amount of speech and the communicative effect of error. **Pronunciation** This criterion refers to the ability to produce comprehensible speech to fulfil the Speaking test requirements. The key indicators will be the amount of strain caused to the listener, the amount of the speech which is unintelligible and the noticeability of L1 influence. **Rating** For Speaking, IELTS uses a single rater. The Speaking is recorded, however, and if there is need for remarking, a second, separate rater will be used for rating.

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	assessment tasks helps promote fair assessment. Incongruent scores—where a test taker's performance on the task fulfilment criterion is markedly different from performance on the other criteria—are flagged to be reviewed by a third (and deciding) examiner.	
Cognitive processing: Levels of speaking	Conceptualisation	Conceptualisation
	Grammatical encoding	Grammatical encoding
	Phono-morphological encoding	Phono-morphological encoding
	Phonetic encoding	Phonetic encoding
	Self-monitoring	Self-monitoring
	Reciprocal, i.e. direct (face-to-face)	Reciprocal, i.e. direct (face-to-face)
	Planning time allowed	Planning time allowed
Discourse mode	Descriptive, biographical, expository, argumentative	Descriptive, biographical, expository, argumentative
Nature of information	Mix of concrete and abstract	Mix of concrete and abstract
Presentation	Both verbal and non-verbal (e.g. graphs)	Verbal and textual (e.g. a cue card)

Summary of Speaking - Key similarities and distinguishing features

Skill integration

The Speaking components of both tests require an integration of skill modalities, incorporating listening, reading, and writing with speaking. Integrating speaking with reading and listening provides a functional assessment, simulating real-world language use where language skills and processes are simultaneously engaged (Butler et al., 2000), thereby providing greater generalisability to the target language use domain. For both LANGUAGECERT Academic and IELTS Academic, listening is required as the test taker must listen to the interlocutor, interpret aural instructions and questions, and vocalize responses accordingly. In LANGUAGECERT Academic, test takers read and discuss a passage for Task 3, and for Task 4, test takers review a visual stimulus, such as a graph, and present it to the interlocutor. Reading is used in the IELTS exam where participants read instructions for the long monologue task. Note taking is permitted for both tests, meaning the Speaking sections for each can engage all language skills.

Tasks

LANGUAGECERT Task 1 and IELTS Task 1

The initial tasks are similar for LANGUAGECERT and IELTS, discussing concrete, familiar topics; however, less time (by 1-2 minutes) is spent in introductory exchanges in LANGUAGECERT compared to IELTS. Because test takers have more time in the initial IELTS Speaking task, more time is spent on concrete, common, and personally familiar topics. Topic familiarity and concreteness are associated with ease of understanding, theoretically making this part of the speaking section more accessible to lower-level language users. Conversely, in LANGUAGECERT, more focus is given to sections with less personal familiarity, potentially making it more challenging.

LANGUAGECERT Task 2

The two minutes less in LANGUAGECERT's initial task is compensated for in the role play. Role play is a unique element to the LANGUAGECERT Academic intended to engage pragmatic interaction and interactional competence (Lampropoulou, 2022).

LANGUAGECERT Task 3 and IELTS Task 2

The third task in LANGUAGECERT and the second task in IELTS require a combination of reading and speaking. In IELTS, test takers are presented with a card which contains brief notes on a given topic. Test takers are given 1 minute to prepare and are encouraged to take notes, further engaging a multimodal skillset with writing. This contrasts with LANGUAGECERT where test takers are given a passage to read aloud and subsequently discuss.

LANGUAGECERT Task 4 and IELTS Task 3

The final task for both tests is a longer response of approximately 5 minutes. For IELTS, participants build from responses in the previous section, and, with the help of the examiner, are expected to articulate more abstract notions. The final task for LANGUAGECERT Academic is an independent response intended to reflect the participant's ability to present academic information. To aid preparation and organisation, LANGUAGECERT encourages note-making at this stage. Participants are given 60 seconds to view the question prompt (e.g., a graph) and plan their response. After giving a presentation of up to 2 minutes, participants are asked a follow-up question or questions as time permits.

Rating and scoring

Task fulfilment (and rubric misalignment)

LANGUAGECERT Academic and IELTS Academic both explicitly measure coherence, vocabulary, grammatical range, and pronunciation. There is a difference, however, in how the two tests approach task fulfilment. Task fulfilment assesses a test taker's ability to meet the specific requirements of a task, such as providing a relevant answer to an interlocutor's question. The

ability to complete tasks successfully is an important aspect of language proficiency and indicates the test taker's ability to use language for real-world purposes. LANGUAGECERT explicitly includes task fulfilment as an assessment criterion. IELTS Academic may implicitly include task fulfilment as part of the test; however, task fulfilment is not explicitly included in the publicly reported grading criteria. Consequently, there appears to be no direct mechanism for markers to address or report in this regard.

Weighting

In LANGUAGECERT, test takers are awarded a mark from 0-8 for each of five criteria (Task Fulfilment and Communicative Effect, Coherence, Accuracy and Range of Vocabulary, Accuracy and Range of Grammar, and Pronunciation, Intonation, and Fluency). The criterion of Task Fulfilment and Communicative Effect is double-weighted, therefore the maximum raw marks a test taker can be awarded is 48. This necessarily has an influence on analytic scoring. IELTS is evenly weighted across descriptors.

Holistic vs analytic scoring

Both Speaking tests are examined by human raters. IELTS is marked analytically, but scores are reported on a single holistic scale with scores ranging from 0 to 9, in half or full bands. Holistic reporting does not permit detailed feedback on specific areas for improvement and no feedback is provided to test takers other than this single score. LANGUAGECERT Speaking is assessed and reported analytically, with sub-scores for task fulfilment (16 points) along with coherence, vocabulary, grammatical range, and pronunciation (8 points each). LANGUAGECERT provides feedback to the test taker on each of these analytical components. In addition to indicating a more refined explication of performance which may highlight areas of strength and areas for improvement, an analytic approach to scoring promotes transparency.

Number of raters

LANGUAGECERT Speaking includes two raters for Speaking assessment opposed to a single rater used in IELTS Speaking. Analytic scoring increases systematicity and transparency of Speaking ratings; however, rating remains

subjective. Raters may have varying standards and can vary in their own consistency. While a second rater can be used to enhance the reliability and validity of the marking (Bejar, 1985), LANGUAGECERT does not implement two raters for each criterion. Instead, LANGUAGECERT splits rating duties between the interlocutor and a second rater. The interlocutor rates "Task Fulfilment and Communicative Effect" while the second rater rates the remaining criteria remotely. Splitting the rating in this way permits the interlocutor to rate in real-time (i.e., is the test taker directly responding to the questions being asked) without the cognitive burden of balancing test administration with analytic elements of scoring. The remaining criteria being scored remotely enables raters more time for analytic scoring. With IELTS, raters must balance test administration with real-time rating, which can add load and make the examination more challenging for raters (Isaacs et al., 2015). For IELTS, the Speaking section is recorded, and a second rater could be used in cases where remarking is necessary.

In conclusion, the Speaking sections of both LANGUAGECERT Academic and IELTS Academic tests demand an integration of multiple language skills in order to operationalise the assessment of speaking skills. There are, however, some differentiating factors. IELTS Academic has a shared format for both its Academic and General Training variants, while LANGUAGECERT Academic exclusively targets the general academic English language domain. Another difference lies in the marking arrangements. The division of rating responsibilities in LANGUAGECERT Academic enables the interlocutor to focus more on the interaction, potentially leading to a more natural and less pressured conversation for the test taker. In contrast, the IELTS Speaking section requires the examiner to juggle both the administration of the test and the evaluation across all criteria simultaneously. LANGUAGECERT Academic's explicit inclusion of task fulfilment as a criterion, with double weighting, clearly communicates the importance placed on achieving communication. Given these observations, a strong alignment in performance would be expected between the speaking scores of the two tests, with some variation in performance at specific points on the proficiency scale.

Listening comparison

The Listening components of LANGUAGECERT Academic and IELTS Academic were compared in terms of task design, input characteristics, and processing demands, with particular attention to the extent to which they elicit comparable aspects of listening ability. A key distinction between the two tests is that the IELTS Listening module is shared across both the Academic and General Training versions, whereas the LANGUAGECERT Academic Listening component is designed specifically for academic contexts. The following table presents the Listening skill comparison of the two tests. The analysis considers features such as task structure, item types, input formats, and the cognitive processes required to comprehend spoken language, in order to examine how listening ability is operationalised in each test.

Table 2: *Listening test comparison of LANGUAGECERT Academic and IELTS (Academic and General Training)*

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
Target level	B1-C2 / Results are reported on a scale which covers CEFR levels A1-C2, with most items targeting B2 and C1 levels	A1-C2 / Results are reported on a scale which covers CEFR levels A1-C2
Skills focus	Test assesses test taker's ability to: Section 1: identify meaning, purpose and function and understand speaker relationship/context. Conversation completion further tests global comprehension and pragmatic knowledge. Section 2: understand meaning, intention, viewpoint,	Test assesses test taker's ability to: Understand details- e.g., listen for names, numbers, and locations and complete a form Section 1: understand concrete, factual information (fill-in form; label a map or diagram)

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	argumentation and speaker relationship Section 3: extract key information from a monologue, synthesise and summarise ideas Section 4: follow a discussion between three speakers	Section 2: understand concrete, factual information (complete text; label diagram) Section 3: understand more abstract information (e.g., opinions, arguments, attitudes, inference). Complete comprehension questions or fill in the blank. Section 4: understand more abstract information (e.g., opinions, arguments, attitudes, inference) Complete the summary/fill in the blank.
Skill integration	LANGUAGECERT Academic Listening contains an integration of skills which include reading an answer sheet to identify information required (and preparation time is given to preview), listening to an audio recording, and writing responses. Read forms and texts, predict and identify missing information to listen for, listen and process, then write word(s), label diagrams, or select correct option from MCQ.	The IELTS Listening test contains an integration of skills which include reading an answer sheet to identify information required (and preparation time is given to preview), listening to an audio recording, and writing responses. Read forms and texts, predict and identify missing information to listen for, listen and process, then

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	For gap fill, test takers must select exact words heard, which is a challenge as test takers may provide synonyms based on short-term memory and lexical access.	write word(s), label diagrams, or select correct option from MCQ. For gap fill, test takers must select exact words heard, which is a challenge as test takers may provide synonyms based on short-term memory and lexical access.
Number of items	30	40
Structure and description	4 sections: MCQ (3 tasks) and cloze listening (1 task) **Task 1**. Short conversation completion or continuation. Seven conversations with one 3-option MCQ each - select the option which completes the conversation. **Task 2**. 5 conversations, 2 MCQ each. Ten total 3-option MCQ across 5 conversations. **Task 3**. Cloze listening task featuring a monologue (e.g., a lecture, podcast, narrative, presentation, etc.). There are seven information gaps to fill with up to three words.	4 sections: mix of MCQ, cloze listening, and labelling **Task 1**. Dialogue (concrete, cloze listening), a conversation between two people set in an everyday social context **Task 2**. Monologue, set in an everyday social context, e.g., a speech about local facilities **Task 3**. Dialogue (academic and training), a conversation between up to four people set in an educational or training context, e.g., a university

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	Task 4. Academic discussion (e.g., podcast) or lecture between up to three people. Six 3-option MCQ.	tutor and a student discussing an assignment **Task 4**. Lecture, a monologue on an academic subject, e.g., a university lecture.
Timing	40 minutes (with double play) of listening. Question preview time is given for test takers to prepare for the listening. This facilitates contextual understanding as well as focusing attention towards required information.	40 minutes: 30 minutes listening, (paper-based gets 10 minutes to transfer answers from question paper to answer sheet; computer-based gets 2 minutes for answer review). Question preview time is given for test takers to prepare for the listening. This facilitates contextual understanding as well as focusing attention towards required information.
Weighting	Each item is worth one point. Correct answers receive one point while incorrect answers receive zero points. Tasks have different weights (i.e., each input recording is accompanied by a different number of items). Tasks 1 and 3 are worth 7 points each, Task 2 has five dialogues worth 2 points	Each item is worth one point. Correct answers receive one point while incorrect answers receive zero points. Tasks are equally weighted at 10 points each (i.e., each input recording is accompanied by 10 items).

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
	each, totalling 10 points, and Task 4 is worth 6 points. Having fewer points on the more difficult task (Task 4) helps mitigate the effect of its difficulty on the overall skill score.	
Item density	Given 20 minutes of original recording (omitting double play) and 30 items, one item can be expected approximately every 40 seconds. This is more dense than IELTS. However, the item density is mitigated by double playing the audio, which permits more processing time to confirm responses.	Given 30 minutes of audio recordings and 40 items, one item can be expected approximately every 45 seconds.
Presentation	Listening passages are played twice.	Listening passages are played once.
Cognitive processing: **Targets**	Factual information Interpretive information related to context In short tasks, overall understanding of the passage (e.g., Task 1, dialogue completion or continuation)	Factual information Interpretive information related to context
Cognitive processing: **Levels of listening**	Word recognition	Word recognition
	Lexical access	Lexical access
	Syntactic parsing	Syntactic parsing

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
targeted by items	Identifying the speaker's point	Identifying the speaker's point
	Inference	Inference
	Making referential links	Making referential links
	Inferring the speaker's attitude	Inferring the speaker's attitude
	Integrating meaning to understand key points or meaning in a conversation.	Integrating meaning to understand key points or meaning in a conversation.
Domain	A range of audio sources including academic articles, lectures, podcasts, interviews, discussions.	Recording 1: a conversation between two people set in an **everyday social context**. Recording 2: a monologue set in an **everyday social context**, e.g., a speech about local facilities. Recording 3: a conversation between up to four people set in an **educational or training context**, e.g., a university tutor and a student discussing an assignment. Recording 4: a monologue on an **academic subject**, e.g., a university lecture

Test	LANGUAGECERT Academic	IELTS (Academic and General Training)
Interaction	Monologic and dialogic	Monologic and dialogic
Discourse mode	Expository, analytical, discursive	Historical/biographical, expository, argumentative
Nature of information	Concrete and abstract, in accordance with an introductory lecture topic	Concrete and abstract, in accordance with an introductory lecture topic
Text length	20 minutes (played twice)	30 minutes

Summary of Listening - Key similarities and distinguishing features

Target levels and timing

Target levels are the first distinguishing feature of the tests. The IELTS Listening test targets CEFR levels A1-C2, which is a broader range than LANGUAGECERT Academic, which targets B1-C2. The IELTS Listening test is correspondingly longer than LANGUAGECERT in terms of number of items (40 items for IELTS compared to 30 items for LANGUAGECERT) and duration (30 minutes of listening for IELTS compared to 20 minutes, single played, for LANGUAGECERT Academic). The IELTS Listening test is shared between the Academic and General Training variants.

Weighting

LANGUAGECERT Academic and IELTS have 30 and 40 items, respectively, with each item weighted equally (one correct response is worth one point). Items are presented in the same order that the relevant information occurs in the audio recordings, offering a degree of scaffolding for the test taker and

lessening the cognitive load of the listening task (i.e., less short-term memory is needed). Regarding the weighting of tasks, there is a difference between the tests. Whereas IELTS has equal weighting across its four tasks (10 points each), LANGUAGECERT tasks have different weights, with slightly more weight given to Task 2 (five dialogues worth 2 points each) compared to the others, and slightly less on Task 4 (6 MCQ based on a discussion). As Task 4 is the most challenging, given the nature of the content, this ensures lower-level test takers are not unduly penalised for a task which is beyond their present level.

Item density

LANGUAGECERT has a slightly higher item density, but permits double play, meaning test takers hear information twice. Addressing the lack of double play, IELTS repeats or spells out important information (e.g., a long name or number). Double play also allows time for adjustment as a variety of accents appear across the Listening test in LANGUAGECERT Academic.

Task structure and description

For task structure and description, both LANGUAGECERT and IELTS make use of monologues and dialogues, though there are more dialogues in LANGUAGECERT (12 across the first two tasks opposed to two in IELTS). Dialogues are more extended in IELTS, which tends to make processing more challenging. LANGUAGECERT and IELTS both employ MCQ and cloze listening activities, though in different proportions. In LANGUAGECERT Listening, there are three MCQ tasks and one cloze listening/gap-fill activity. One of the MCQ tasks is unique (there is just one question for each recording) in that the options continue a short conversation, testing global comprehension and pragmatic knowledge. Contrastively, IELTS balances MCQ and cloze listening/gap fill across tasks. Similarly to LANGUAGECERT, one of the MCQ tasks is unique; however, for IELTS, the unique task employs information matching (e.g., who says what information, or match descriptions to a list of nouns) opposed to conversation continuation. Theoretically, LANGUAGECERT

Academic should have an easier structure to follow and less cognitive processing load compared to IELTS.

Nature of information

The LANGUAGECERT Listening test is all set in the academic domain whereas IELTS Listening is shared between the Academic and General Training variants and is, thus, more general English in nature. Descriptively, the nature of information appears more challenging for LANGUAGECERT Academic, with greater emphasis on abstract information compared to IELTS, which is generally more concrete, particularly in Tasks 1 and 2. This may reflect the use of a single Listening test covering both the Academic and General Training tests, as more concrete information may be needed for the General Training than for the Academic test taking population. LANGUAGECERT Academic Listening is less closely aligned with IELTS, as it contains less general and familiar topic areas compared with the dual-purpose nature of the Listening test in IELTS Academic. This can be found in the use of lectures as opposed to informational talks in Task 3, or in the use of an academic rather than a topical discussion in Task 4. LANGUAGECERT Listening is, therefore, more closely tied to the academic domain.

Cognitive processing

Higher levels of comprehension, which involve interpreting meaning in context and constructing an understanding of a conversation or lecture, are crucial in academic and professional settings. Both tests assess these higher levels to some degree. LANGUAGECERT and IELTS Listening tasks provide textual support with visual aids, such as a torn piece of paper with writing on it or a notepad with a pencil; however, IELTS additionally offers diagrammatic aids to fill in (e.g., map labelling). This offers additional context for learners, and labelling tasks may benefit or hinder test takers depending on their aptitude for spatial awareness. Some individuals may be disproportionately affected by a given item or task type and having multiple task types helps to mitigate this risk. However, MCQ is a very commonly employed and widely known item type,

and LANGUAGECERT's predominant use of MCQ may help to facilitate understanding of the task, permitting participants to focus more on the content than on how to respond to the task.

Comparison limitations: listening difficulty is impacted by numerous factors which have not been described here as we do not have access to the IELTS test specifications. For instance, speech rate, accent, and lexical and grammatical complexity in the audio recordings can all impact a listener's ability to accurately comprehend speech. However, assuming certain industry standards are followed (e.g., speech rate of 150 words/minute; targeting a percentage of B2-level vocabulary for academic texts), LANGUAGECERT and IELTS may be expected to be similar in these respects.

In conclusion, the comparison between LANGUAGECERT Academic and IELTS Academic Listening tests reveals several significant similarities, which suggest that the listening scores will be positively correlated. Both tests assess listening skills in English, including the ability to understand main ideas and specific factual information, recognise the opinions, attitudes, and purpose of speakers, and follow the development of an argument. Both tests are aligned with the Common European Framework of Reference for Languages (CEFR), even though they target slightly different ranges within the framework. Both tests include academic content and aim to assess the test taker's ability to understand spoken English in academic settings. At the same time, there are certain factors that can be expected to affect the strength of the correlation. IELTS Academic offers a broader range of CEFR levels and a shared format suitable for both Academic and General Training candidates. LANGUAGECERT Academic focuses more on the academic context with a slightly higher item density and the feature of double play. These patterns suggest that performance is likely to align across the two tests, and that candidates who score well on one test are likely to score well on the other.

Writing comparison

The Writing components of LANGUAGECERT Academic and IELTS Academic were compared in terms of task design, response requirements, and scoring frameworks, with particular attention to the extent to which they elicit comparable aspects of written language ability. Both tests assess candidates' ability to produce extended written discourse in response to academic-style prompts, although they differ in the way tasks are framed and supported. The following table summarises the key characteristics of the two tests. The analysis considers task types, input formats, expected response length, and the criteria used to evaluate performance, in order to examine how writing ability is operationalised in each test.

Table 3: *Writing test comparison of LANGUAGECERT Academic and IELTS Academic*

Test	LANGUAGECERT Academic	IELTS Academic
Target Level	B1-C2 / Markscheme covers performance at A1-C2 levels	A1-C2 / Markscheme covers performance at A1-C2 levels
Response format and Genre	Task 1: Continuous writing; Information transfer from visual or textual input (e.g., a chart, graph, or table) Task 2: A longer piece of continuous writing; Essay	Task 1: Continuous writing; Information transfer from multiple non-verbal inputs Task 2: A longer piece of continuous writing; Essay
Task description	The LANGUAGECERT Academic Writing test is designed to assess a wide range of writing skills, including how well test takers: • write a response appropriately • organise ideas • use a range of vocabulary and grammar accurately	The IELTS Academic Writing test is designed to assess a wide range of writing skills, including how well test takers: • write a response appropriately • organise ideas • use a range of vocabulary and grammar accurately

Test	LANGUAGECERT Academic	IELTS Academic
	Task 1. Test takers explain a visual input (e.g., a graph) in 150-200 words with the intended reader specified. Test takers might describe and explain data, express a stance, opinion, justification, or argument in accordance with the provided prompt. Task 2. Test takers explain opposing views of a given topic and express their own opinion in 250 words. Test takers read two opposing opinions, discuss both views, and write their personal perspective.	In Task 1 test takers are presented with a graph, table, chart or diagram. Test takers are asked to describe, summarise or explain the information in their own words. This might involve describing and explaining data, describing the stages of a process or how something works, or describing an object or event. In Task 2 test takers are asked to write an essay in response to a point of view, argument or problem.
Domain	Academic, social	Academic, social
Purpose	Task 1: • To demonstrate the ability to understand and synthesise visual or textual inputs • To show the ability to write a report, argument or article using a written, graphic or visual input with the intended reader specified expressing stance, opinion, justification, argumentation. Task 2: • To write a formal piece of	Task 1: • To transfer information from multiple inputs • To collate different pieces of information in order to describe, summarise or explain the information. Task 2: • To write a persuasive essay • To defend or attack a particular argument or opinion, compare or contrast aspects of an argument, and give reasons

Test	LANGUAGECERT Academic	IELTS Academic
	writing for a specified reader which may compare and contrast, persuade, argue, hypothesise, evaluate, analyse, or present solutions.	for the argument.
Timing	50 minutes. No explicit instruction is provided to divide time, but an expected word count is indicated for Task 1 (150-200 words) and Task 2 (250 words).	60 minutes. Test takers should spend 20 minutes on Task 1, and 40 minutes on Task 2. Test takers need to manage their own time.
Text length of expected response	Task 1: 150-200 words Task 2: 250 words	Task 1: at least 150 words Task 2: at least 250 words
Weighting	Task 1: 40% Task 2: 60%	Task 1: 33.3% Task 2: 66.6%
Skills assessed	In both tasks, test takers are assessed on their ability to write a response which is appropriate in terms of content, the organisation of ideas, and the accuracy and range of vocabulary and grammar. Task fulfilment is an explicit part of the marking criteria.	In both tasks, test takers are assessed on their ability to write a response which is appropriate in terms of content, the organisation of ideas, and the accuracy and range of vocabulary and grammar. Task fulfilment is an explicit part of the marking criteria.

Test	LANGUAGECERT Academic	IELTS Academic
Cognitive processing	**Task 1:** • Macro-planning: goal setting and task representation • Transforming ideas from verbal and non-verbal inputs • Organising ideas • Translating • Micro-planning • Monitoring and revision • Comprehending non-graphic task instructions • Comprehending (and interpreting) the components of graphs • Re-presenting or re-producing the non-graphic and graphic information as continuous discourse in written form in English as a foreign language **Task 2:** • Macro-planning: goal setting and task representation • Generating ideas • Organising ideas • Translating • Micro-planning • Monitoring and revision	**Task 1:** • Macro-planning: goal setting and task representation • Transforming ideas from non-verbal inputs • Organising ideas • Translating • Micro-planning • Monitoring and revision • Comprehending non-graphic task instructions • Comprehending (and interpreting) the components of graphs • Re-presenting or re-producing the non-graphic and graphic information as continuous discourse in written form in English as a foreign language **Task 2:** • Macro-planning: goal setting and task representation • Generating ideas • Organising ideas • Translating • Micro-planning • Monitoring and revision

Test	LANGUAGECERT Academic	IELTS Academic
Discourse mode (rhetorical task)	Descriptive, expository, argumentative/persuasive	Descriptive, expository, argumentative/persuasive
Scoring approach	Analytic	Analytic (reported holistically)
Scoring and marking	Test takers are assessed on their performance on each task by trained examiners according to four criteria: Task Achievement, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, and Organisation (Coherence). Scores are reported both holistically and analytically. A holistic score is provided using the Global Scale (out of 100), and additional feedback is provided based on performance on individual marking criteria (i.e., task achievement, accuracy and range of grammar, accuracy and range of vocabulary, and organisation (coherence). Marking Marking is double blind between two examiners. If a large discrepancy exists between the two examiners, the	Test takers are assessed on their performance on each task by certificated IELTS examiners according to the four criteria of the IELTS Writing Test Band Descriptors (task achievement/response, coherence and cohesion, lexical resource, grammatical range and accuracy). Scores are reported holistically in whole and half bands. Between two and four examiners mark IELTS Writing assessments. Marking Marking is done between two and four examiners for accuracy and fairness.

Test	LANGUAGECERT Academic	IELTS Academic
	writing response is passed to a third senior examiner whose marks are final.	

Summary of Writing - Key similarities and distinguishing features

Timing

Less time is devoted to the Writing section for LANGUAGECERT Academic (50 minutes) compared to IELTS Academic (60 minutes). The expected word counts are similar across tasks, however, and test takers taking the LANGUAGECERT exam will likely experience more time pressure compared to IELTS test takers.

Tasks

Task 1

The initial tasks are similar for LANGUAGECERT Academic and IELTS Academic, where test takers are prompted to discuss a visual input (e.g., graph, table, diagram) for a minimum of 150 words. The writing is formal in register. LANGUAGECERT Task 1 differs from IELTS Task 1 in how it supports responses. LANGUAGECERT provides a response scaffold through the prompt. For instance, a chart is provided with the instruction to write a report which describes the main trends, give reasons for the trends, and predict likely changes over a given period of time. Contrastively, IELTS Academic instructions are more open to interpretation on the part of the test taker: for example, asking test takers to summarise a chart and "make comparisons where relevant".

Task 2

The second task for both tests is a longer response, argumentative essay, with a recommended minimum of 250 words. As with Task 1, the writing is formal in register. For LANGUAGECERT, test takers are asked to read two opposing opinions, summarise them, and offer their own perspective on the topic. For IELTS, one opinion is offered, and the test takers must explain the extent to which they agree or disagree with the given perspective. The opposing view (e.g., "children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents" is inferred (IELTS online test preparation). Because only one perspective is offered to an argumentative topic, the prompts may more strongly elicit a personal or emotional reaction than LANGUAGECERT, which may influence results.

Rating and scoring

Task fulfilment (and rubric alignment)

The Writing section rubrics for LANGUAGECERT and IELTS are more strongly aligned than in the Speaking section, with a direct correspondence in each rubric component. In the Writing section markers score coherence, grammatical and vocabulary accuracy and range, but distinct for LANGUAGECERT Academic compared to IELTS Academic, task fulfilment is an explicit criterion.

Weighting

For both tests, greater scoring weight is given to Task 2, the extended writing task, than Task 1. For LANGUAGECERT, Task 1 is given a weight of 40% of the total marks for Writing, and Task 2 is given a weight of 60% of total marks. For IELTS, slightly more weight is given to Task 2, with approximately 2/3rds of the weight allocated to Task 2 compared to 1/3rd for Task 1. Given the weighting, test takers in IELTS are instructed to divide their time accordingly across tasks, with 20 minutes devoted to Task 1 and 40 minutes to Task 2. On the test form, LANGUAGECERT does not explicitly advise test takers on which component

should receive more time, though it does state the suggested number of words that should be provided for each task.

Holistic vs analytic scoring and reporting

Both Writing tests are examined by human markers. IELTS Writing is assessed analytically and reported holistically on a scale of 0–9, whereas LANGUAGECERT Writing is assessed and reported analytically, with sub-scores (worth a maximum of 8 points each) for Task Achievement, Accuracy and Range of Grammar, Accuracy and Range of Vocabulary, and Organisation (Coherence). LANGUAGECERT provides feedback on these analytical components, whereas IELTS simply reports a band score. In addition to indicating a more refined explication of performance which may highlight areas of strength and areas for improvement, an analytic approach to scoring promotes transparency in scoring.

Number of markers

Both LANGUAGECERT and IELTS attempt to ensure reliability and validity of marking through multiple markers. LANGUAGECERT and IELTS employ a minimum of two markers for the Writing tasks. LANGUAGECERT will use a third (who holds the final decision) if discrepancies exist, while IELTS can use up to four.

In conclusion, the Writing sections of LANGUAGECERT Academic and IELTS Academic share key similarities in their aims to evaluate candidates' ability to articulate ideas in written English, emphasizing formal register, critical thinking, and analytical engagement with prompts. Both tests underscore the importance of the extended writing task through higher weighting, reflecting the value placed on argumentation and coherence in academic contexts. Despite a slight difference in timing, the fundamental objectives and evaluation criteria are aligned, as rating scales for both tests are heavily derived from the relevant CEFR scales. Given these similarities, a strong alignment in performance would be expected between scores in the Writing sections of LANGUAGECERT Academic and IELTS Academic.

Reading comparison

The Reading components of LANGUAGECERT Academic and IELTS Academic were compared in terms of text characteristics, task design, and processing demands, with particular attention to the extent to which they elicit comparable aspects of reading ability. Both tests aim to assess candidates' ability to comprehend academic texts and extract relevant information, although they differ in the structure of tasks and the nature of input texts. The following table presents a structured comparison of the key characteristics of the two tests. The analysis considers features such as text length and complexity, task types, item distribution, and the cognitive processes required for comprehension, in order to examine how reading ability is operationalised in each test.

Table 4: *Reading test comparison of LANGUAGECERT Academic and IELTS Academic Task Features*

Test	LANGUAGECERT Academic	IELTS Academic
Target Level	B1-C2 / Results are reported on a scale which covers CEFR levels A1-C2, with most items targeting B2 and C1 levels	A1-C2 / Results are reported on a scale which covers CEFR levels A1-C2
Skills focus	To test students' ability to comprehend academic texts and to extract important information from those texts. The test is designed to assess a wide range of reading skills, including how well test takers: • read for the general sense of a passage • read for the main ideas • read for detail • understand vocabulary used in academic texts, identify	To test students' ability to comprehend academic texts and to extract important information from those texts. The test is designed to assess a wide range of reading skills, including how well test takers: • read for the general sense of a passage • read for the main ideas • read for detail • understand inferences and

Test	LANGUAGECERT Academic	IELTS Academic
	synonyms and use vocabulary in context • understand lexico-grammatical features in academic texts • understand inferences and implied meaning • understand how meaning is built up in discourse and show awareness of text organisation and discourse features • understand long complex texts, including discourse, opinion, purpose, argumentation, exemplification, comparison and contrast, cause and effect and locate specific information	implied meaning • recognise a writer's opinions, attitudes and purpose • follow the development of an argument
Task description	There are 30 questions. A variety of question types is used, chosen from the following types: multiple choice, gap fill, identifying information, identifying writer's views/claims, matching information to source texts, sentence and text completion, synonym identification. **Task 1a.** (4-option synonyms) Six sentences written in an	There are 40 questions across three parts (three passages). A variety of question types is used, chosen from the following types: multiple choice, identifying information, identifying writer's views/claims, matching information, matching headings, matching features, matching sentence endings, sentence completion,

Test	LANGUAGECERT Academic	IELTS Academic
	academic style with one word highlighted. Test takers choose a synonym for highlighted word from a list of four words. **Task 1b.** (3-option cloze) An authentic academic text that may include academic ideas, arguments and opinions with five words removed. Test takers choose the correct word from a choice of three to full each gap **Task 2.** (Gapped sentences) An academic text with 6 sentences removed. Test takers choose from 8 sentences to complete the text. **Task 3.** (Short text comprehension with multiple matching) Test takers match seven questions to four texts. Texts may be reviews, reports, articles, journals, opinion pieces, etc. with a linked theme, but with a different purpose. **Task 4.** (Long text comprehension with 4-option MCQ) An extended text (e.g., narrative, descriptive, explanatory, expository,	summary completion, note completion, table completion, flow-chart completion, diagram label completion, short-answer questions. Sometimes one-word answers are required, sometimes a short phrase, and sometimes simply a letter, number or symbol. Mainly receptive, some limited writing involved in short answer questions, but only brief answers are required; no more than a given number of words. Test takers lose marks for incorrect spelling and grammar.

Test	LANGUAGECERT Academic	IELTS Academic
	biographical, instructive) with 6 MCQs.	
Number of items	30	40
Timing	50 minutes to answer 30 questions on 7 passages (across 4 tasks.)	60 minutes to answer a total of 40 questions on 3 passages. Individual tasks are not timed.
Weighting	All items equally weighted. Each correct answer receives one mark. Tasks have different weights (i.e., each text is accompanied by a different number of items). Scores out of 30 are converted to the Global Scale out of 100 for reporting purposes.	All items equally weighted. Each correct answer receives one mark. Scores out of 40 are converted to the IELTS Academic 9-band scale. Scores are reported in whole and half bands
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics) **Expeditious reading: global** (skim for gist/search for key ideas/detail) **Careful reading: local** (understanding sentence) **Careful reading: global** (comprehend main idea(s)/overall text(s))	**Expeditious reading: local** (scan/search for specifics) **Expeditious reading: global** (skim for gist/search for key ideas/detail) **Careful reading: local** (understanding sentence) **Careful reading: global** (comprehend main idea(s)/overall text(s))
Cognitive processing	Word recognition	Word recognition
	Lexical access	Lexical access
	Syntactic parsing	Syntactic parsing

Test	LANGUAGECERT Academic	IELTS Academic
Levels of reading	Establishing propositional meaning (cl./sent. level)	Establishing propositional meaning (cl./sent. level)
	Inferencing	Inferencing
	Building a mental model	Building a mental model
	Creating a text level representation (disc. structure)	Creating a text level representation (disc. structure)
	Creating an intertextual representation (multi-text)	Creating an intertextual representation (multi-text)

Table 5: *Reading test comparison of LANGUAGECERT Academic and IELTS Academic: Features of the Input Text*

Test	LANGUAGECERT Academic	IELTS Academic
Word count	There are 7 passages to read, plus questions, across 4 sections. Given provided texts, test takers must read approximately 2750-2800 words, including source text and questions. Long and short texts are provided to elicit different skills. Longer texts permit an assessment of skimming and scanning for information, and discriminate between important and secondary details. Shorter texts are used to assess vocabulary (synonyms) and lexico-grammatical knowledge.	Three different passages to read, each with accompanying questions. Officially test takers have to read 2,150 - 2,750 words in total. There are three sections to the IELTS Academic Reading test, and each contains one long text. Green, Ünaldi and Weir and (2010) analysed 42 texts making up 14 IELTS reading tests. The passages in their study contained 854 words on average (maximum 1063 words, minimum 589 words).
Average sentence length	16.32 words per sentence on average across all sentences.	21.89 words per sentence on average across all sentences.
Domain	Academic	Academic
Discourse mode	Narrative, descriptive, explanatory, expository, biographical, instructive	Historical/biographical, expository, argumentative
Nature of information	Mix of concrete and abstract	Mostly concrete
Presentation	Verbal (textual)	Both Verbal (textual) and Non-verbal (i.e., graphs)

Test	LANGUAGECERT Academic	IELTS Academic
Lexical Level; further criteria	The cumulative coverage reaches 95.6% at the K3 level. 75.2 K1, 12.3 K2, 8.1 K3. Lexical Density 0.58	The cumulative coverage reaches 92% at the K3 level. 76.4 K1, 11.36 K2, 3.26 K3. Lexical Density 0.57
Readability	Flesch-Kincaid Grade Level 10.27	Flesch-Kincaid Grade Level 12.64
Topic	A range of topics which may reasonably reflect introductory coursework and texts at the university level are employed. Topics represented in sample material include controversial technology in sport, animal domestication, language preservation, and wildlife and deep-sea mining.	A broad range of subject areas were represented among the 42 IELTS texts examined by Green et al. (2010) with the categories of *Social studies* (10 or 11 texts), *Engineering & technology* (6 or 7) and *Business & administrative studies* (4 or 5) emerging as the most popular topic areas for the test.
Text genre	The seven reading texts are appropriate to the general academic genre, including vocabulary and topic content. Longer texts mimic introductory texts (e.g., textbook excerpts, academic reports) which cover a range of concrete and abstract information. Given the nature of the test and the lack of assumed technical knowledge, text excerpts are	Three reading texts with a variety of question types. The kinds of text used in IELTS introduce academic topics to a general audience, often in the form of articles sourced from newspapers or magazines presenting research findings to a general audience. These include self-contained reports on developments in science and technology and

Test	LANGUAGECERT Academic	IELTS Academic
	potentially less difficult than course content which is both extended and domain specific.	overviews of academic debates. While IELTS passages are at a level of difficulty appropriate to university study, they are not as challenging as some of the texts encountered in the more linguistically demanding areas such as the law textbook analysed by Green et al. (2010).

Summary of Reading - Key similarities and distinguishing features

Target levels and timing

Targeted proficiency levels are the first distinguishing feature of the tests as they impact test design. The IELTS Academic Reading test targets CEFR levels A1-C2, a broader range than LANGUAGECERT Academic, which targets B1-C2. The IELTS Academic Reading test is correspondingly longer than LANGUAGECERT Academic in terms of number of items (40 items for IELTS Academic compared to 30 items for LANGUAGECERT Academic) and duration (60 minutes of Reading for IELTS Academic compared to 50 minutes for LANGUAGECERT Academic).

Weighting

Reading components for LANGUAGECERT Academic and IELTS Academic have 30 and 40 items, respectively, with each item weighted equally (one correct response is worth one point). Items are presented in the same order that the relevant information occurs in the passages, offering a degree of scaffolding for the test taker and lessening the cognitive load of the reading task (i.e., less short-term memory is needed). Regarding the weighting of tasks, there is a difference between the tests. Whereas IELTS Academic has equal weighting across its four tasks (10 points each), LANGUAGECERT tasks have different weights across tasks. Tasks 1A and 1B combine for 11 points, Task 2 is worth 6 points, Task 3 is worth 7 points, and Task 4 is worth 6 points. As Task 4 is the most challenging, given its length and the nature of the content, this ensures lower-level test takers are not unduly penalised for a task which is beyond their present level.

Task description

LANGUAGECERT has more reading passages compared to IELTS, but fewer questions. Longer texts are associated with greater levels of difficulty; however, an increased number of passages may also increase difficulty. LANGUAGECERT includes components which more directly assesses elements of vocabulary and grammar (Task 1). The structure of LANGUAGECERT is more predictable than IELTS, as tasks and item types are largely uniform in the Reading test. Contrastively, IELTS Reading can have different numbers of items per passage and there are at least 14 question types which may be given (as shown in the Comparison Table). LANGUAGECERT's test, therefore, may permit greater scaffolding and targeted preparation. However, it does so at the cost of response type coverage. Further, offering different item types can help prevent the performance of individual test takers being unduly influenced by specific item types, whether due to relative strength or weakness.

Input texts

Input texts are similar in genre as both tests target the academic domain. However, the texts in IELTS Academic were longer and somewhat more difficult overall compared with LANGUAGECERT Academic, with a readability score of 12.64 compared to 10.27 (readability is associated with grade level according to the US school system), and fewer words at the K3 level or below (92% for IELTS Academic compared to 96% for LANGUAGECERT Academic).

Cognitive processing

Though IELTS Academic offers more diversity in option types, both of the tests appear to engage similar levels of processing, including word recognition, lexical access, syntactic parsing, establishing propositional meaning, inferencing, building a mental model, creating a text level representation of discourse structure, and creating an intertextual representation across texts.

In conclusion, the Reading sections of IELTS Academic and LANGUAGECERT Academic are designed to evaluate candidates' proficiency in handling

academic texts, albeit with distinct approaches in terms of target proficiency levels, test duration, and item distribution. In comparing the two tests, several aspects suggest that the LANGUAGECERT test could present certain challenges. While IELTS Academic targets a wider range of CEFR levels (A1-C2), LANGUAGECERT Academic focuses on B1-C2 levels, concentrating its content on higher proficiency levels. Despite having fewer items, the weighting of tasks in LANGUAGECERT Academic is distinct, with varying points allocated to different sections. Moreover, the LANGUAGECERT Academic test incorporates more reading passages than IELTS Academic. The inclusion of more passages could elevate the difficulty by requiring test takers to adjust to different texts more frequently, thereby taxing their cognitive resources more heavily.

Conclusion to Content Comparison Section

Compared in detail in the sections above, the constructs and their assessment across the two examinations – LANGUAGECERT Academic and IELTS Academic – show high degrees of similarity. Both tests are four-skill assessments intended to measure language competence in order to predict the ability to operate successfully in an academic setting.

At the same time, a number of differences are evident. LANGUAGECERT Academic is consistently situated within the academic domain across all four skills, whereas IELTS Academic focuses entirely on the academic domain in the Reading and Writing tests, but is more General English in nature in the Listening and Speaking tests.

The objectively marked components of the IELTS Academic tests have more items in order to cover the full range of CEFR levels (A1-C2) whereas the LANGUAGECERT Academic tests for Reading and Listening focus exclusively on the B1-C2 levels, primarily targeting B2 and C1 as these are the levels typically required for entrance onto English-language medium courses of study.

There is a wider range of task types for the receptive tests of IELTS Academic while the LANGUAGECERT Academic examination has a fully consistent test

format, with each sub-skill test having the same task formats and number of items in each test version. The Reading and Listening sub-skills tested show a very high degree of similarity.

The marking criteria for the tests of Writing and Speaking show a high degree of similarity. One exception is the 'Task Fulfilment and Communicative Effect' criterion for LANGUAGECERT Academic Speaking. This is not explicitly measured as a separate criterion in IELTS. In terms of the scoring of tasks, the tests assess a very similar range of levels. LANGUAGECERT Academic has a two-examiner model for the Speaking test meaning marks are given on a test taker's performance from two different perspectives, whereas IELTS Academic Speaking has, as a standard, a one-examiner model. The marking of both Writing tests features at least two examiners who award marks independently.

The outcome of the content comparison indicates that the two tests address comparable domains of language use and share numerous similarities. Notable differences do emerge at the skill level, as highlighted in the conclusions for each skill component. Among others, these differences include such features as content, question density, and specificity of focus. Differences in task design and scoring are not unexpected given the independent development of the tests, and do not appear to undermine their overall comparability. Rather, they highlight areas where variation in performance may arise and should be taken into account in the interpretation of results.

References

Bejar, I. I. (1985). Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language. ETS Research Report Series, 1985(1), i-57.

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

Chapelle, CA, Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. Language testing, 32(3), 385-405.

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

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. Language Testing, 27(2), 191-211. https://doi.org/10.1177/0265532209349471

Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B. M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. IELTS research reports online series, 4.

Lam, R. (2021). Using eportfolios to synergise assessment of, for, as learning in EFL writing. European Journal of Applied Linguistics and TEFL, 10(1), 101-120.

Lampropoulou, L. (2022). Interactional competence and the role role play plays. International Journal of TESOL Studies, 4(1), 32-47. doi.org/10.46451/ijts.2022.01.04. · Mar 30, 2022

Lampropoulou, L., Milanovic, M., Jones, J., Green, A. (2024), *LANGUAGECERT Academic Concordance Report.* LANGUAGECERT.

Mayordomo, R. M., Espasa, A., Guasch, T., & Martínez-Melo, M. (2022). Perception of online feedback and its impact on cognitive and emotional engagement with feedback. Education and Information Technologies, 27(6), 7947-7971.

Mežek, Š., McGrath, L., Negretti, R., & Berggren, J. (2022). Scaffolding L2 academic reading and self-regulation through task and feedback. TESOL Quarterly, 56(1), 41-67.

Pommerich, M. & Dorans, N. J. (Eds.) (2004). Concordance [Special issue]. *Applied Psychological Measurement, 28*(4), 216–289.

Taylor, L., & Chan, S. (2015). IELTS Equivalence Research Project (GMC 133)

Chapter 9: A Score-Based Concordance Analysis of LANGUAGECERT Academic and IELTS Academic

Leda Lampropoulou, Michael Milanovic, Jonathan Jones and Anthony Green

Abstract

This study examines the relationship between performance on LANGUAGECERT Academic and IELTS Academic to support score interpretation across the two assessments. A single-group, counterbalanced design was used, with 1,008 test takers completing both tests. Correlation analyses showed a strong relationship between overall scores (r = 0.87), with consistent associations across individual skills. Equipercentile linking was applied to establish correspondences between score scales, producing concordance tables that provide an interpretable mapping between the two tests. Results are most stable in the mid-range of proficiency, with greater variability observed at the extremes due to smaller sample sizes and distributional differences. Population invariance analyses indicate that score relationships are broadly consistent across key subgroups. Overall, the findings support the use of concordance relationships to interpret LANGUAGECERT Academic

scores alongside IELTS Academic performance, while acknowledging that concordance reflects statistical alignment rather than direct equivalence.

Keywords: construct comparability, concordance studies, contents comparison, LANGUAGECERT Academic

Introduction

The interpretation of scores across different English language proficiency tests is a common requirement in academic admissions, migration, and professional contexts. In such settings, stakeholders are often required to make decisions based on results obtained from different assessments, raising questions about how scores can be compared across tests. Concordance studies aim to address this need by establishing empirical relationships between scores on different assessments, thereby supporting score interpretation and use.

Concordance differs from equating in that it applies to tests that are developed independently and may measure similar, but not identical, constructs (Pommerich & Dorans, 2004). As a result, concordance relationships are not intended to support score interchangeability, but rather to provide indicative mappings that assist users in interpreting performance across assessments. The validity of such mappings depends not only on statistical relationships between scores, but also on the degree of construct alignment between the tests being compared.

A detailed comparison of LANGUAGECERT Academic and IELTS Academic, including both qualitative and quantitative analyses, is reported in a technical report (Lampropoulou et al., 2024). That report demonstrates substantial alignment in the domains and abilities assessed by the two tests. Building on this foundation, the present study focuses specifically on the quantitative relationship between performances on the two assessments, using a large-scale concordance dataset in which test takers completed both tests.

The study adopts a single-group, counterbalanced design and applies equipercentile linking to establish relationships between scores on LANGUAGECERT Academic and IELTS Academic. In addition to overall score concordance, relationships between individual skill components are examined, alongside analyses of population invariance and the stability of linking across subgroups.

The analysis addresses the following question:

How do scores on LANGUAGECERT Academic relate to performance on IELTS Academic?

Background and Conceptual Framework

Concordance studies play a crucial role in language assessment by investigating the comparability and alignment of different language proficiency tests. These studies aim to establish empirical evidence of the alignment between test scores from different language assessments, enabling informed decision-making regarding test interpretation and usage. They are crucial for several reasons, including, among others, establishing score mapping, evaluating test comparability, informing test interpretation, as well as enhancing test development. Results from concordance studies can inform policymakers and educational institutions in setting standards for language proficiency and selecting appropriate tests for different educational purposes.

Over the past two decades, there has been sustained interest in linking scores across different English language tests. Linking methods may be viewed on a cline (Pommerich & Dorans, 2004). In equating, tests measure the same construct with equal reliability. Analysis is conducted through equating raw scores from both tests. Scores are reported on the same scale and can be used interchangeably. Concordancing establishes a relationship between scores on tests that are built to different specifications. They measure similar but not identical constructs. Scores from two tests linked through concordancing are not expected to be interchangeable.

A substantial body of research has examined concordance between IELTS Academic and other widely used English language tests. These studies have generally reported moderate to strong correlations, typically in the range of *r* = 0.70 to 0.78 (Hatch & Lazaraton, 1991). Table 1 summarises recent English language test concordance studies involving IELTS Academic, illustrating the consistency of these findings across different test pairs. It could therefore be expected that overall correlations between the LANGUAGECERT Academic and IELTS Academic tests would also be in this range, given that both are English language tests intend to cover similar target language use domains.

Table 1. *English language tests concordance studies with IELTS Academic*

Test pair	Authors	Sample size	Overall correlation r
IELTS Academic and PTE-Academic	Clesham & Hughes, 2020	573	0.74
IELTS Academic and PTE-Academic	Elliot et al., 2021	523	0.70
TOEFL iBT and IELTS Academic	ETS, 2010	1,153	0.73
Duolingo English Test and IELTS Academic	LaFlair & Settles, 2019	991	0.78

These findings suggest that, despite differences in design and task characteristics, tests targeting similar domains of academic English proficiency tend to show strong relationships in performance. Within this context, the present study examines the relationship between LANGUAGECERT Academic and IELTS Academic.

The full concordance study, including both qualitative and quantitative analyses, is reported in a technical report (Lampropoulou et al., 2024). That report demonstrates substantial alignment in the constructs and domains assessed by the two tests. Building on this foundation, the present paper focuses specifically on the statistical relationship between performances on the two assessments.

The study was conducted over a two-year period (2022–2024) and involved more than 1,000 test takers, each of whom completed both the LANGUAGECERT Academic and IELTS Academic tests. A single-group, counterbalanced design was adopted to control for order effects, with test administrations scheduled within a

defined time window to minimise the impact of language development between sittings.

Two complementary approaches were used in the broader study. The first involved qualitative comparisons of test content, task characteristics, and scoring frameworks. The second, which is the focus of this paper, involved statistical analyses of test taker performance, including correlation analysis and equipercentile linking, in order to establish relationships between scores on the two tests.

It should be noted that concordance findings are inherently dependent on the performance of the specific sample of the population assessed, and although a large sample size and a robust methodology can allow generalisation, score mappings should be interpreted as indicative rather than exact equivalences (Knoch, 2021). As such, concordance tables are intended to support, rather than replace, informed judgment in the interpretation and use of test scores, and test score users are advised to consult concordance tables in combination with additional validation documents and score interpretation frameworks.

Test Context

Both LANGUAGECERT Academic and IELTS Academic are four-skill assessments of English language proficiency designed for use in academic contexts, reporting both overall and skill-specific performance. LANGUAGECERT Academic reports scores on a 0–100 scale aligned to the Common European Framework of Reference for Languages (CEFR), whereas IELTS Academic reports scores on a 9-band scale. While the tests differ in their task design and scoring frameworks, prior analyses have demonstrated substantial alignment in the constructs they target (Lampropoulou et al., 2024). These similarities provide the basis for examining statistical relationships between scores on the two assessments.

Statistical Analysis Framework

The aim of the concordance study was to identify how performance on the LANGUAGECERT Academic (LCA) exam relates to performance on IELTS Academic. IELTS Academic was selected as a comparison for the present study as it is a similar measure of English language proficiency targeting general academic English as a language use domain, it has been mapped to several different frameworks, is generally accepted for admissions and immigration purposes, and has publicly available performance data, published annually, which could readily be used for comparison.

The analysis focuses on the statistical relationship between scores on the two tests, including overall and skill-level performance. Correlation analyses are used to examine the strength of association between scores, and equipercentile linking is applied to establish correspondences between score scales. In addition, the stability of these relationships is examined through analyses of population invariance across key subgroups.

The results and conclusions of the concordance study are reliant on the sample they are based on, and it is therefore instructive to first identify the sample as it compares to the larger population of potential test takers. The following section describes the test taker population and evaluates their representativeness in relation to the broader population of the IELTS Academic test and the wider population of individuals needing to prove their English competency for higher education admission and immigration purposes to English-speaking countries, and to Australia in particular.

Test taker demographics

Sample overview

The concordance dataset comprises 1,008 test takers who completed both LANGUAGECERT Academic and IELTS Academic. A single-group design was used, with all participants contributing scores to both assessments, providing

a direct basis for examining relationships between performances. Participants were required to complete both tests within a specified time window and to provide official IELTS test scores, ensuring the accuracy and comparability of the dataset.

The sample population was carefully selected to closely mirror the population of interest for the study, i.e., individuals with profiles similar to IELTS Academic test takers, who wish to use their test score to pursue migration for academic or educational purposes in higher learning institutions where English is the medium of instruction, such as universities in Australia. Participants represented a wide range of nationalities and first language backgrounds, with proficiency levels primarily spanning the B1 to C1 range of the CEFR. Gender split was relatively equal.

Demographic profile

The mean age of participants was 24.64 years (SD = 6.31), meaning test takers were between 18 and 31 years old, consistent with typical age ranges for undergraduate and postgraduate study (House, 2010). Although publicly available IELTS statistics do not report age distributions, comparison with migration data suggests that the sample aligns with the expected age profile of candidates seeking study or migration opportunities in English-speaking contexts.

In terms of gender, the sample comprised 602 (59.72%) female participants and 404 (40.08%) male participants, with a small number of undisclosed responses. This distribution is broadly consistent with IELTS Academic test populations. Mean IELTS scores in the study closely align with published IELTS statistics, indicating comparable performance levels across genders. Gender split and performance for both tests are shown in the tables below, demonstrating a close match to the total scores means of the published data.

Table 2. *IELTS Academic mean total scores by gender for the study sample compared to published IELTS Academic data (2022).*

Gender	Test takers	% of sample	IELTS 2022 data - % of sample	LCA total score mean	IELTS total score mean	IELTS 2022 data - total score mean
Female	602	59.72	51.78	64.50	6.33	6.28
Male	404	40.08	48.22	62.29	6.16	6.22
Undisclos	2	0.20	N/A	66.50	6.75	N/A
Total	1008	100	100	63.62	6.26	N/A
Female *n* = 602, male *n* = 404, undisclosed *n* = 2, total *n* = 1008.						

Table 3. *IELTS Academic mean skill component scores by gender for the study sample compared to published IELTS data (2022).*

Gender	Concordance study	IELTS 2022
	Listening	
Female	6.40	6.51
Male	6.40	6.52
	Reading	
Female	6.50	6.26
Male	6.20	6.20
	Writing	
Female	5.90	5.95
Male	5.90	5.86
	Speaking	
Female	6.00	6.17
Male	6.00	6.06

Regarding nationality, the concordance study aimed to be as representative of the test population as possible. The largest non-native English speaking populations, who may require an English language test for migration purposes, arrived from India, China, and the Philippines (Australian Bureau of Statistics, 2024). Half (51%) of international student arrivals originated from China and India, with additional

significant numbers coming from Nepal, Colombia, the Philippines, Thailand and Pakistan. Concordance test taker nationalities are presented in table 4.

Table 4. *Test taker Nationality*

Nationality	Test takers	% of sample
Chinese	471	46.73
Indian	260	25.79
Iraqi	89	8.83
Spanish	38	3.77
Greek	30	2.98
Thai	22	2.18
Turkish	18	1.79
Other	80	7.93
Total	1008	100

To further evaluate representativeness, mean scores for overall performance and individual skills were compared with published IELTS Academic statistics. The results are presented in Tables 5 and 6.

Table 5. *IELTS Academic mean total scores by nationality for the study sample compared to published IELTS data (2022).*

Nationality	Concordance study	IELTS Academic 2022
	Overall	
All (n=1008)	6.3	6.3
China (n=471)	6.2	6.1
India (n=260)	6.3	6.2
Colombia (n=11)	6.7	6.6
Thailand (n=27)	6.0	6.1
Pakistan (n=10)	5.8	6.2

Table *6.* *IELTS Academic mean skill component scores by nationality for the study sample compared to the published data (2022).*

Nationality	Concordance study	IELTS 2022	Concordance study	IELTS 2022
	Listening		Reading	
All (n=1008)	6.4	6.5	6.4	6.2
China (n=471)	6.4	6.1	6.7	6.4
India (n=260)	6.6	6.6	6.1	6.0
Colombia (n=11)	6.9	6.7	6.7	6.9
Thailand (n=27)	6.1	6.3	6.0	6.1
Pakistan (n=10)	6.1	6.5	5.4	6.1
	Writing		Speaking	
All (n=1008)	5.9	5.9	6.0	6.1
China (n=471)	5.9	5.8	5.7	5.6
India (n=260)	6.1	5.9	6.1	6.1
Colombia (n=11)	5.9	6.0	6.7	6.6
Thailand (n=27)	5.6	5.7	5.7	5.8
Pakistan (n=10)	5.7	5.9	5.5	6.2

The study population closely resembles the test population of interest in terms of performance. For both overall scores and subskills, test taker performance is within 0.1-0.7-point difference. The maximum difference is less than an IELTS Academic band (0.7) and this is only in two subskills and a very small sample size (i.e., Pakistan, n=10).

First language backgrounds were similarly diverse, with the largest groups corresponding to major global test-taking populations, as shown in Table 7.

Table 7. *Test taker first language background*

First language (mother tongue)	Test takers	% of sample
Chinese	476	47.22
Punjabi	96	9.52
Kurdish	95	9.42
Malayalam	75	7.44
Spanish	51	5.06
Tamil	41	4.07
Greek	30	2.98
Hindi	22	2.18
Thai	22	2.18
Other	100	9.93
Total	1008	100

This distribution further supports the representativeness of the sample in relation to international English language testing populations.

Self-reported proficiency

Participants reported their English language proficiency with reference to CEFR levels. As it could not be expected that all participants would be familiar with the CEFR levels, and/or able to self-estimate accurately, they were asked to take an online placement test which LANGUAGECERT has available on its website. The distribution of self-reported proficiency spanned all levels, with the majority of test takers falling within the B1–C2 range (A1 = 49, A2 = 21, B1 = 84, B2 = 206, C1 = 204, C2 = 66). The bulk of proficiency levels being at B1-C2 reflects the target range of the LANGUAGECERT Academic test.

Sequence and Exam interval

All participants took both tests (LANGUAGECERT Academic and IELTS Academic), and a counterbalanced design was used to control for potential order effects. Counterbalancing occurred such that test takers were split into two administration groups: one group completed IELTS Academic first followed by LANGUAGECERT Academic, while the other group completed the tests in the reverse order. The final distribution was 459 participants taking IELTS Academic first and 549 taking LANGUAGECERT Academic first.

An independent samples t-test was conducted to examine the effect of exam sequence on performance. For LANGUAGECERT Academic, mean scores were slightly lower for those taking LANGUAGECERT Academic first (M = 63.17, SD =12.85) compared to those who took the IELTS Academic first (M = 64.12, SD = 11.97); however, this difference was not found to be statistically significant, t(994.311) = 1.216, p = .224. Similarly, for IELTS Academic, mean scores were marginally lower for those taking IELTS Academic first (M = 6.25, SD = .85) compared to those who took the LANGUAGECERT Academic first (M = 6.26, SD = .91). This again was not found to be statistically significant, t(1008) = .265, p = .791. For this dataset, sequence, or taking one exam before the other, did not have a significant effect on performance.

Restricting the time between administrations helps minimise the possibility of test takers' language proficiency improving from one test to another, potentially biasing results. This study was designed so that the interval between test takers attempting both tests was as small as possible, and that it should not be greater than three months (Dorans et al., 1997).

The targeted three-month window was largely maintained, with the average time between the two sittings being 38 days, and the majority (n = 868, or 86%) of the test takers taking both tests within 90 days. Despite best efforts to adhere to this timeframe, however, the study period of 2022-2024 coincided with the enforcement of COVID-19 pandemic-related regulations in certain countries, notably affecting the schedule. Specifically, the Chinese government imposed severe restrictions in various cities, including Beijing and Shanghai, from November 2022 to February 2023, disrupting exam schedules due to

prohibitions on access to testing facilities for candidates and staff. Additionally, this led to difficulties for participants in securing available IELTS Academic slots following a nearly three-month suspension of exams. As a result, a minority of participants (n = 99, or 10%) experienced slight delays, but managed to undertake the second test between 91 and 120 days. A further 4% of participants (n = 41) encountered more significant delays, completing the second examination between 121 days and 203 days of the first.

We retained test takers who took the tests more than 90 days apart because:

- statistical analyses showed that the difference in test taker performance on the first test and second test taken was not statistically significant regardless of whether tests were taken within 90 days or beyond 90 days,
- the average interval between administrations of 38 days is relatively short compared to other large scale language testing concordance reports (e.g., Clesham and Hughes, 2020),
- the additional 140 test takers provided valuable data and helped achieve the targeted 1,000 participants for data analysis and
- any sensitivity to duration between test administrations would be mitigated through the study's counterbalanced design.

Familiarity

As LANGUAGECERT Academic was, at the time of the concordance study, a relatively new language test, differences in test familiarity between the two tests were anticipated. Familiarity (and subsequently motivation) may have an effect on test taker performance, and therefore reasonable attempts to control and document it were made. It was controlled to the extent possible by making test information materials available to all test takers. Test takers were provided with familiarisation packages for both tests covering test format and task types, practice papers and a prerecorded webinar. They were also advised and reminded to access the materials prior to sitting the exam.

Familiarity was measured prior to test administration using a four-point Likert scale (not familiar, a little familiar, familiar, very familiar). Responses indicated that participants were generally more familiar with IELTS Academic than with LANGUAGECERT Academic. However, analysis of mean scores showed minimal differences in performance associated with familiarity levels. For LANGUAGECERT Academic, mean scores were consistent across familiarity groups. For IELTS Academic, participants reporting higher familiarity showed slightly higher mean scores; however, these differences were not statistically significant. These findings suggest that test familiarity did not have a meaningful impact on performance in the concordance dataset. The relationship between test familiarity and performance on LANGUAGECERT Academic and IELTS Academic is summarised in Tables 8 and 9 respectively.

Table 8. *Test taker familiarity with the LANGUAGECERT Academic test at exam registration*

LC Academic	n	Mean	SD
No response	383	62.82	11.74
Not /A little familiar	405	64.19	12.38
Very / Familiar	220	63.88	13.75

Table 9. *Test taker familiarity with the IELTS Academic test at exam registration*

IELTS Academic	n	Mean	SD
No response	377	6.13	0.86
Not /A little familiar	127	6.03	0.94
Very / Familiar	504	6.41	0.86

Reasons for taking the test

Using the same questionnaire, test takers were asked about their primary reasons for taking each test. Reasons were provided by 659 test takers concerning the LANGUAGECERT test while 622 test takers gave reasons for taking the IELTS Academic exam. Using the certificate for “a higher education extended course (more than three months)” was the main reason for taking

both tests (26% for LCA and 38% for IELTS) followed by "other educational purposes" and "personal reasons".

Score distribution and Descriptive statistics

Descriptive statistics for LANGUAGECERT Academic and IELTS Academic scores are presented in Table 10. The distributions for both tests approximate normality, with minimal skewness and kurtosis, indicating that scores are well distributed across the proficiency range. Descriptive statistics were also calculated for individual skill components, as shown in Table 11.

Table 2. *Distribution statistics for LANGUAGECERT Academic and IELTS Academic*

	Mean	SD	Skew	Kurtosis	Min	Max
LANGUAGECERT	63.62	12.46	-.106	-.077	22	96
IELTS Academic	6.26	0.88	-.059	-.144	3	8.5
Sample size = 1008						

Table 3. *Subskill descriptive statistics comparison between LANGUAGECERT Academic and IELTS Academic*

	Overall	Listening	Reading	Writing	Speaking
LANGUAGECERT Academic	63.62 (12.46)	60.48 (17.18)	63.07 (17.25)	61.96 (11.07)	68.43 (12.76)
IELTS Academic	6.26 (0.88)	6.42 (1.28)	6.38 (1.31)	5.93 (0.64)	6.01 (0.88)
Sample size = 1008. *Indices include mean followed by standard deviation in brackets.*					

Test taker performance analysis

Correlations

Correlations were calculated to examine the relationship between LANGUAGECERT Academic and IELTS Academic. Taylor and Chan (2015) reported correlations between several tests of English for Academic Purposes. They found overall correlations between selected tests (i.e. CAE, CPE, TOEFL iBT, OET and PTE-A) and IELTS Academic ranged from 0.73 to 0.87. A strong, positive correlation between tests (i.e. above 0.7) suggests that performing well on one test would translate to performing well on the other, while performing poorly on one test generally corresponds with performing poorly on the other.

Because both tests in this study measure English language proficiency, a strong overall correlation can be expected. The correlation between LANGUAGECERT Academic and IELTS Academic was $r = 0.87$, indicating a strong positive relationship between performances on the two tests. This value is at the upper end of correlations reported in comparable studies, suggesting a high degree of alignment in how the two assessments rank test takers according to their language proficiency. This relationship is illustrated in Figure 1.

Figure 1. *Overall score relationship between LANGUAGECERT Academic and IELTS Academic*

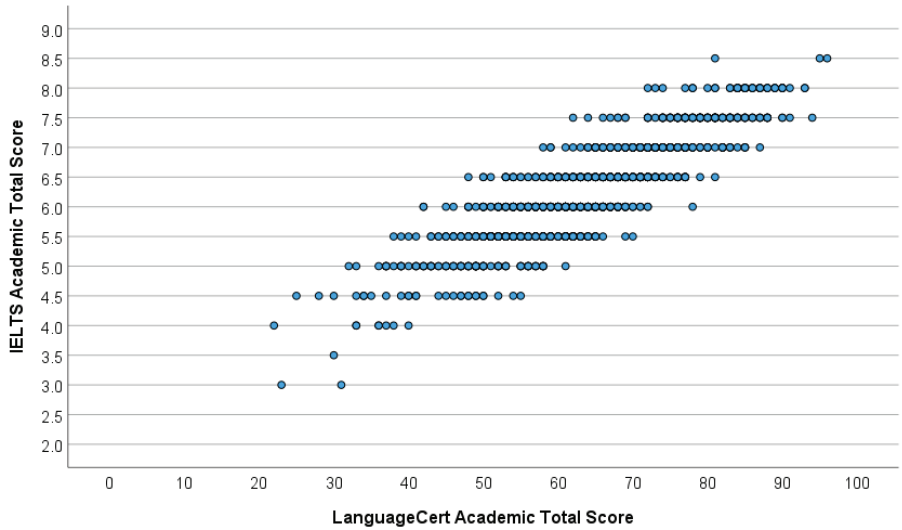


Figure 1 shows the linear relationship between tests, where a high score on LANGUAGECERT Academic is associated with a high score in IELTS Academic, while a low score on LANGUAGECERT Academic is associated with a low score in IELTS Academic.

Correlations between scores for component skills were also examined, with results shown in Table 12 alongside Overall performance.

Table 12. *Correlations between LANGUAGECERT Academic and IELTS Academic*

Overall *r*	Reading *r*	Writing *r*	Listening *r*	Speaking *r*
.87	.76	.71	.71	.71
Note: *r* = correlation. All correlations were statistically significant at the p < .001 level. Sample size = 1008.				

The strong correlation for Overall performance (r=.87) indicates the two tests measure similar underlying abilities (Knoch, 2021). The correlations for individual subscales (r > .7) suggest that LANGUAGECERT exams perform in accordance with the benchmark language test designed for similar purposes. This meets or exceeds correlations between alternative tests and IELTS Academic as reported in the studies in table 1 in this report.

Comparing test scores

To support score interpretation across the two assessments, it is necessary to establish how scores on LANGUAGECERT Academic correspond to scores on IELTS Academic. This is achieved through statistical linking, which provides a basis for interpreting performance across different scoring scales.

Two common linking methods were considered: linear regression and equipercentile ranking. With linear regression, it is not necessary to have data for each point in a given test score range and it can be calculated with smaller samples, making it desirable for studies with relatively few participants. Consequently, linear regression was implemented in the earlier stages of the concordance project when the sample was less robust. The use of linear regression for linking, however, is limited because it is unidirectional and produces different results depending on which test is treated as the predictor. As a result, it is less suitable for generating concordance tables intended for practical interpretation by score users.

Equipercentile linking addresses this limitation by establishing score correspondences based on percentile ranks, allowing for bidirectional interpretation of scores across tests. This approach is widely used in concordance studies and is particularly appropriate when sufficient sample sizes are available.

Given the size of the dataset (n = 1008), equipercentile linking was applied using a single-group, counterbalanced design. Analyses were conducted using the statistical software R (R Core Team, 2021) and the *equate* package (Albano, 2016). Pre-smoothing techniques were considered; however, as results were consistent with and without smoothing, unsmoothed data were retained, as it

was determined that the small changes made were insufficient to justify altering the data with smoothing.

The resulting concordance between LANGUAGECERT Academic and IELTS Academic scores is presented in Table 13.

Table 13. *Overall alignment table for LANGUAGECERT Academic and IELTS Academic performance.*

See recommendations for interpretation and use of linkage results for test users below.

IELTS Academic	LANGUAGECERT Academic	n-size of study sample	SE
4.5	38-45	27	1.56
5.0	46-53	74	0.79
5.5	54-60	185	0.43
6.0	61-66	225	0.40
6.5	67-72	189	0.46
7.0	73-80	148	0.47
7.5	81-87	104	0.59
8.0	88-94	42	0.78
8.5	95+	3	4.84
9.0	n/a	0	n/a

SE= Standard deviation of LCA scores at each IELTS half band level, divided by the square root of the sample size at that level.

Skill comparison

Following the analysis of overall scores, concordance relationships were examined for each of the four language skills. The same equipercentile linking approach was applied to skill-level scores using the full dataset and without smoothing.

Compared to overall performance, relationships at the skill level show greater variability. This is expected, as individual skill scores are subject to greater measurement error and reflect differences in task design, scoring approaches, and construct emphasis across the two tests.

The resulting concordance relationships for Listening, Reading, Writing, and Speaking are presented in Tables 14–17.

Table 14. *Listening skill alignment table for LANGUAGECERT Academic and IELTS Academic performance.*

See recommendations for interpretation and use of linkage results for test users below.

IELTS Academic Listening	LANGUAGECERT Academic Listening	n-size of study sample at this level	Standard Error
4.5	35-40	40	1.48
5.0	41-48	108	1.13
5.5	49-56	147	1.03
6.0	57-61	152	1.03
6.5	62-66	135	0.99
7.0	67-72	109	1.30
7.5	73-79	100	1.28
8.0	80-88	78	1.31
8.5	89-94	77	1.20
9.0	95-100	25	1.51

Table 15. *Reading skill alignment table for LANGUAGECERT Academic and IELTS Academic performance.*

See recommendations for interpretation and use of linkage results for test users below.

IELTS Academic Reading	LANGUAGECERT Academic Reading	n-size of study sample at this level	Standard Error
4.5	36-43	59	1.51
5.0	44-53	125	1.04
5.5	54-59	161	0.91
6.0	60-64	126	1.03
6.5	65-70	133	0.98
7.0	71-76	98	1.10
7.5	77-82	93	1.08
8.0	83-88	68	1.35
8.5	89-96	82	1.11
9.0	97-100	34	1.58

Table *16.* *Writing skill alignment table for LANGUAGECERT Academic and IELTS Academic performance.*

See recommendations for interpretation and use of linkage results for test users below.

IELTS Academic Writing	LANGUAGECERT Academic Writing	n-size of study sample at this level	Standard Error
4.5	33-44	18	3.34
5.0	45-55	91	1.01
5.5	56-63	221	0.56
6.0	64-70	389	0.36
6.5	71-77	199	0.48
7.0	78-83	65	0.71
7.5	84-88	11	1.06
8.0	89-92	2	9.00
8.5	93+	1	n/a
9.0	n/a	0	n/a

Table 17. *Speaking skill alignment table for LANGUAGECERT Academic and IELTS Academic performance.*

See recommendations for interpretation and use of linkage results for test users below.

IELTS Academic Speaking	LANGUAGECERT Academic Speaking	n-size of study sample at this level	Standard Error
4.5	44-53	34	1.79
5.0	54-61	107	0.92
5.5	62-69	211	0.71
6.0	70-75	275	0.49
6.5	76-81	171	0.60
7.0	82-86	109	0.70
7.5	87-88	45	1.16
8.0	89-92	19	1.52
8.5	93-98	8	1.22
9.0	99-100	3	3.38

Across the subjectively assessed skills (Speaking and Writing), concordance patterns are relatively consistent, with score relationships closely aligned across most of the proficiency range. This is in line with the similarity of scoring criteria used in both tests, which are grounded in CEFR-based descriptors. It is only at the lowest end of the proficiency continuum that the writing test results diverge more significantly, which may in part be attributable to the paucity of data at the lowest levels. Narrower standard deviations observed in IELTS Speaking and Writing indicate a tendency for scores to cluster more closely around the mean, whereas LANGUAGECERT Academic shows a wider spread of performance. These differences in score dispersion may influence the mapping between tests, particularly at the extremes of the proficiency scale.

Greater divergence is observed in the objectively marked skills (Listening and Reading), where LANGUAGECERT Academic appears to be more demanding. In particular, IELTS Academic shows a higher proportion of test takers achieving scores in the upper bands (8.0–9.0), whereas corresponding scores on LANGUAGECERT Academic are less frequent at these levels. This pattern

suggests differences in score distributions between the two tests, which may be related to variations in test design, input characteristics, and difficulty. This was expected to some extent in Listening, given its more general focus in IELTS Academic compared to the more academic orientation of LANGUAGECERT Academic.

On close inspection, this divergence is particularly evident in the distribution of high-performing test takers in IELTS Academic Listening and Reading, a pattern not observed in the Speaking and Writing components. The IELTS Academic Reading test has a more academic orientation, similar to LANGUAGECERT Academic; however, a similar concentration of high scores is observed in Reading as in Listening. While the underlying causes of this pattern are not entirely clear, it affects the relationship between the two tests at the upper end of the scale.

Despite these variations, the overall pattern of results indicates consistent alignment between the two assessments across all skills. Differences are most evident at the lower and upper ends of the scale, where smaller sample sizes and distributional effects contribute to increased variability in score correspondence. These differences were not considered to compromise the equipercentile scaling described in this section.

Recommendations for interpretation and use of linkage results for test users

Score users, for example institutions who use certain test scores for decisions about test takers, are advised that score comparisons across tests, while based on empirical research, are estimates only and should be treated with caution for the following reasons:

- Tests differ, sometimes significantly, in the ways information about English language ability is elicited and assessed. Score comparisons are only meaningful to the extent that the tests are measuring the same ability or skill.
- Tests often differ in the length of the reporting scales used (for example one test may report on a 6-point scale and another on a 100-point scale). As a result, a one-to-one mapping of scores from one test to another is rarely possible.
- The choice of concordance study methodology may produce variations in results.
- The populations of test takers may differ (e.g., with respect to ages, nationalities, language backgrounds of test takers) from the population used in the research that generated the score equivalences.
- The sample sizes used for comparing scores from different tests are generally small across all levels/ranges, especially at extreme ends of the scale.
- Score concordance results are generally more robust for proficiency levels with larger numbers of test takers.
- Large Standard Errors show that score equivalences are particularly imprecise at certain points on the ability scale.

Because the score comparisons presented in the score comparison table are indicative only, score users are advised not to rely solely on published score equivalences in making their decisions. They should weigh evidence from additional sources where feasible.

Population invariance

Population invariance—where similar linking results should be found regardless of the subpopulation that the linking function is applied to—has been posited as a fundamental condition of linking and equating studies (Dorans & Holland, 2000). It was therefore relevant to not only identify the concordance of scores between LANGUAGECERT Academic and IELTS Academic, but to investigate how results with subpopulations corresponded with these findings. Though subdividing the study to investigate subgroup performance necessarily reduces the sample size and can complicate subsequent analyses (Brennan, 2008), preliminary calculations were conducted to establish processes going forward. Two measures were taken to explore invariance, one at the test level, and one at the item level. Invariance at the test level will be outlined here by using the previously described equipercentile method on key subgroups.

An initial population invariance was explored by looking at the largest subgroups within the study, namely, gender (female and male) and nationality (Chinese and non-Chinese). Results are summarised in the tables 18 to 19 below.

Table 18. *Population Invariance - Overall*

IELTS Academic Score	Female	Male	Chinese	All Other Nationalities
4.5	37-45	40-46	37-47	39-45
5	46-53	47-53	48-54	46-52
5.5	54-60	54-59	55-60	53-59
6	61-67	60-65	61-67	60-66
6.5	68-72	66-71	68-72	67-71
7	73-80	72-79	73-81	72-78
7.5	81-86	80-87	82-89	79-86
8	87+	88+	90+	87+

Table 19. *Population Invariance – Speaking*

IELTS Academic Score	Female	Male	Chinese	All Other Nationalities
4.5	43-52	46-53	44-54	43-49
5	53-62	54-60	55-62	50-60
5.5	63-70	61-69	63-71	61-67
6	71-75	70-76	72-78	68-74
6.5	76-81	77-81	79-85	75-79
7	82-86	82-86	86-88	80-85
7.5	87-88	87	89-93	86-87
8	89+	88+	94+	88+

Table 20. *Population Invariance – Listening*

IELTS Academic Score	Female	Male	Chinese	All Other Nationalities
4.5	36-41	33-39	38-43	27-37
5	42-49	40-46	44-52	38-44
5.5	50-57	47-54	53-59	45-53
6	58-62	55-59	60-63	54-59
6.5	63-68	60-64	64-69	60-64
7	69-73	65-69	70-73	65-71
7.5	74-81	70-75	74-80	72-77
8	82+	76+	81+	78+

Table 21. *Population Invariance – Reading*

IELTS Academic Score	Female	Male	Chinese	All Other Nationalities
4.5	37-44	35-43	34-41	36-44
5	45-53	44-53	42-51	45-54
5.5	54-59	54-59	52-58	55-60
6	60-63	60-66	59-62	61-66
6.5	64-69	67-72	63-68	67-72
7	70-75	73-77	69-74	73-78
7.5	76-81	78-82	75-80	79-84
8	82+	83+	81+	85+

Table 42. *Population Invariance – Writing*

IELTS Academic Score	Female	Male	Chinese	All Other Nationalities
4.5	35-44	31-44	36-44	32-44
5	45-56	45-55	45-56	45-55
5.5	57-64	56-62	57-63	56-63
6	65-71	63-69	64-71	64-69
6.5	72-77	70-77	72-78	70-75
7	78-84	78-82	79-85	76-81
7.5	85-89	83-87	86+	82-88
8	90+	88+	n/a	89+

Across subgroups, the concordance relationships show a high degree of overlap, indicating that score mappings are broadly consistent across gender and nationality. Some variation is observed, particularly at the upper and lower ends of the proficiency scale, where smaller sample sizes contribute to increased variability. Though more data is needed, the strong reliability shown in both tests further indicates support for invariance within the sample. As Dorans and Holland (2000, pp 300-301) state, "whenever the reliabilities of X and Y (the two tests, LANGUAGECERT Academic and IELTS Academic, in this case) are both high, "near population invariance" is expected to hold for a wide range of subpopulations." LANGUAGECERT Academic was found to have strong reliability across language skills, supporting potential invariance.

Concordance results discussion

The qualitative comparisons between LANGUAGECERT Academic and IELTS Academic demonstrate that the tests are very similar in construct and content, with LANGUAGECERT Academic being more focused on the academic domain in terms of the Listening and Speaking tests. Statistically, there are robust correlations between LANGUAGECERT Academic and IELTS Academic overall

(.87) and by skill (all >.70) suggesting that the two tests are measuring language abilities in very similar ways.

Speaking

In terms of the separate language skills that results are reported on, for LANGUAGECERT Academic Speaking – each main score with a full descriptor on the rating scale (8, 6, 4, 2 and 1) is tied to a CEFR level. That is, 8 is designed to reflect C2 level CEFR Can-Do descriptors, 6 is tied to C1, 4 to B2 and so on. An analysis of the marking criteria for IELTS Academic and LANGUAGECERT Academic shows that they are similar in nature and descriptions of performance, and so, assuming reliable interpretation of the mark scheme, similar results would be expected. This is, indeed, the case with a slightly wider spread of scores for LANGUAGECERT Academic. Therefore, from a qualitative and quantitative perspective the two Speaking tests align very well. It is noticeable that the concordance study candidature achieved their highest mean scores on LANGUAGECERT Academic on the Speaking component. This may be because, even though the LANGUAGECERT test is less well-known, the Speaking test is the one component where a test taker is actively engaged in an interaction with a human interlocutor and so performs to the best of their abilities. This is likely to have some effect on the overall concordance despite best efforts to minimise any lack of motivation through effective design and incentivisation. It is also worth noting the paucity of IELTS Speaking performances at levels 8 and above (only 2.8% of the test takers), which contrasts with the percentage of test takers achieving a total score of 8 and above overall (30.2%).

Writing

In terms of the Writing tests, for LANGUAGECERT Academic each score with a full descriptor on the markscheme for LANGUAGECERT Academic (8, 6, 4, 2 and 1) is tied to a CEFR level. That is, 8 reflects CEFR C2 level Can-Do descriptors, 6 is tied to C1, 4 to B2 and so on. An analysis of the marking criteria for IELTS Academic and LANGUAGECERT Academic shows that they are closely aligned (having the same roots in the CEFR Can-do descriptors). Assuming consistent interpretation of the rating scales, similar results would be expected. This is, indeed, the case with a slightly wider spread for LANGUAGECERT Academic scores. Therefore, from both the qualitative and quantitative perspective the two Writing tests appear to align well. A closer examination of the Writing section shows that IELTS Academic scores are slightly compressed compared to LANGUAGECERT Academic scores. Only 3 out of the 1008 test takers in the study achieved an IELTS Academic writing score of 8 or above.

Figure 2. *IELTS Academic Writing scores distribution*

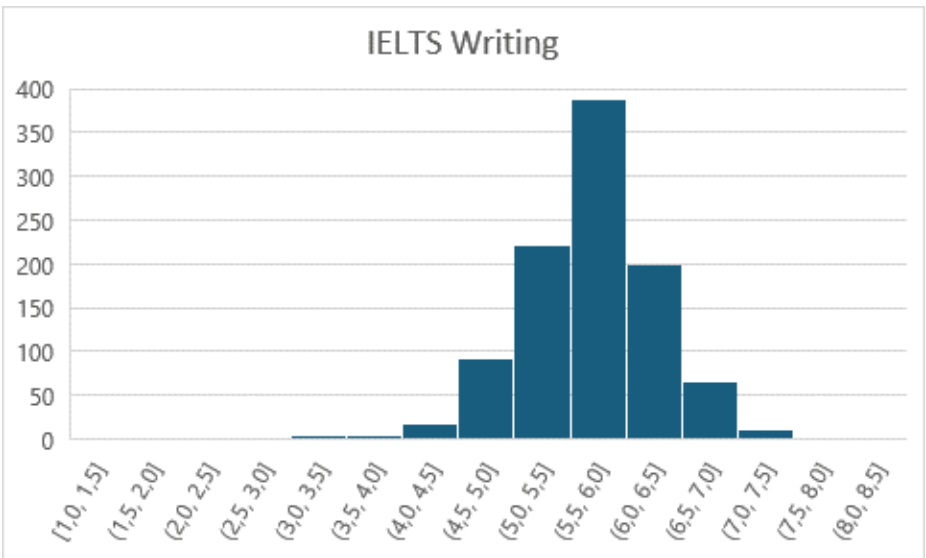
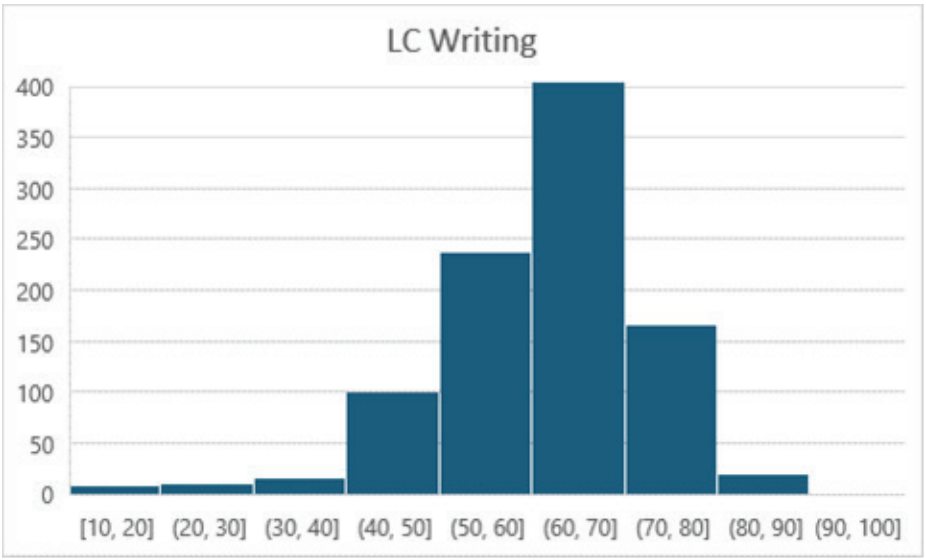


Figure 3. *LANGUAGECERT Academic Writing scores distribution*



Listening

From the concordance study, the receptive skills tests show more difference in terms of test taker results across the two examinations. There are certain features of the comparison between Listening scores in LANGUAGECERT Academic and IELTS Academic that should be noted. In the data used for this study, the LANGUAGECERT Academic Listening results show a classic normal distribution, whereas the IELTS Academic distribution has a normal distribution with a slight positive skew illustrating a larger proportion of test takers scoring very highly (bands 8.0-9.0) compared to overall IELTS Academic scores.

Figure 4. *IELTS Academic Listening scores distribution*

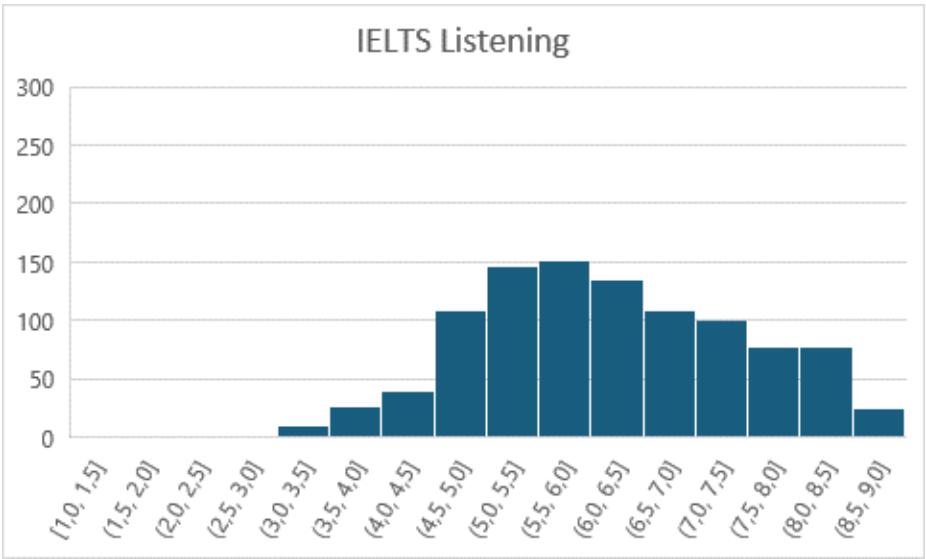
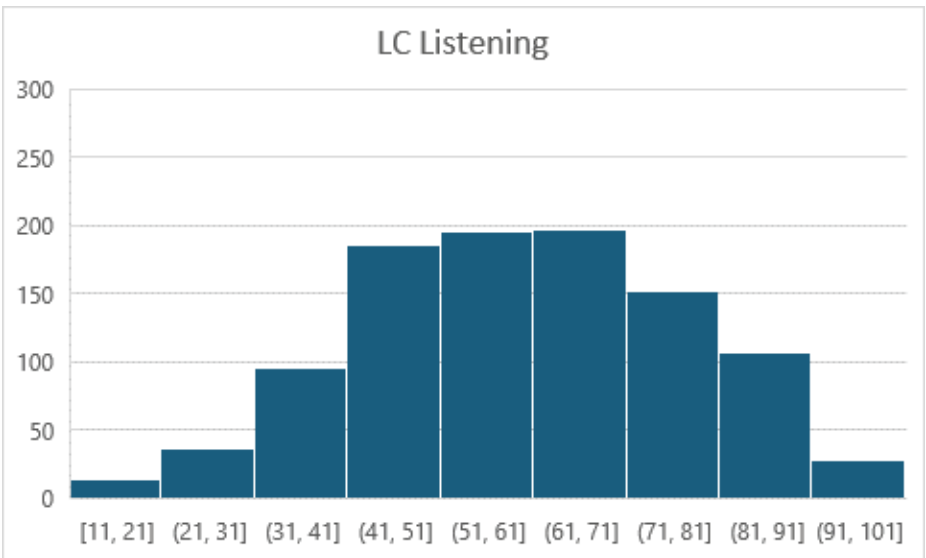


Figure 5. *LANGUAGECERT Academic Listening scores distribution*



The LANGUAGECERT Academic Listening items used in the concordance study have good item statistics, with test reliability figures above .80, and item facilities almost all in the target range (.30 to .85) and high levels of discrimination. These statistics indicate tests that have performed well. However, in addition to the differences in score distributions, there was some evidence in the concordance study that the LANGUAGECERT Academic Listening test, perhaps due to its more academic nature, was slightly more difficult for test takers than the IELTS Academic Listening test.

Reading

An analysis of the LANGUAGECERT Academic Reading tests presents a similar picture to Listening, although with less marked differences. The distribution for IELTS Academic Reading again shows a slightly positively skewed normal distribution showing a larger proportion of candidates scoring very highly (bands 8.0 to 9.0) and higher than their overall IELTS Academic scores. The LANGUAGECERT Academic Reading tests administered in the concordance study also produced strong reliability figures (around .85), item facilities were almost all in the target range (.30 to .85) and with high levels of discrimination. These statistics again show tests that performed well. However, there was some evidence in the concordance study that the LANGUAGECERT Academic Reading test was slightly more difficult for test takers than the corresponding IELTS Academic Reading test.

Conclusion

Concluding this analysis, it is worth pointing out that alignment in English language testing is not a singular, unidimensional process; rather, it is a continuous, complex and intricate procedure that demands sustained attention and evolves over time. Continuous validation involves regularly assessing the test against established criteria to verify its accuracy and effectiveness, whereas alignment requires consistent adjustments to ensure that test content aligns with current language standards and accurately reflects the skills and competencies being measured. These processes are essential for maintaining the validity and fairness of English language tests over time, acknowledging the evolving nature of language and the diversity of the test taker population. LANGUAGECERT advises test score users to treat all score comparison tables as indicative only, and not to rely solely on published score equivalences in making their decisions. They should weigh evidence from additional sources where feasible. LANGUAGECERT's concordancing studies programme will continue beyond the point of conclusion of this study with the view to a) confirming the healthy performance of our test within the different contexts and b) determining concordance with comparable English language tests, such as TOEFL iBT, PTE Academic and others as appropriate in the context and for the purpose the test is used.

References

Albano, A. D. (2016). equate: An R Package for Observed-Score Linking and Equating. *Journal of Statistical Software*, *74*(8), 1–36. https://doi.org/10.18637/jss.v074.i08

Australian Government, Department of Home Affairs. (2023). *Migration strategy*. Retrieved from https://immi.homeaffairs.gov.au/programs-subsite/migration-strategy/Documents/migration-strategy.pdf

Clesham, R., & Hughes, S. R. (2020). concordance report PTE Academic and IELTS Academic. London: Pearson.

Dorans, N.J., Lyu, C.F., Pommerich, M., & Houston, W.M. (1997). Concordance between ACT Assessment and recentered SAT I sum scores. College & University, 73, 24-34.

Elliot, M., Blackhurst, A., O'Sullivan, B., Clark, T., Dunlea, J., & Saville, N. (2021). Aligning IELTS and PTE-Academic: A measurement study. In N. Saville, B. O'Sullivan & T. Clark (Eds.), *IELTS Partnership Research Papers: Studies in Test Comparability Series*, No. 2, (pp. 42–64). IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia

ETS. (2010). *Linking TOEFL iBT scores to IELTS scores – a research report*. https://www.ets.org/pdfs/toefl/linking-toefl-ibt-scores-to-ielts-scores.pdf

Hatch, E. M., & Lazaraton, A. (1991). The research manual: Design and statistics for applied linguistics. New York, NY: Newbury House Publishers.

House, G. (2010). Postgraduate Education in the UK. HEPI Analytical Report 1. *Higher Education Policy Institute*.

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

LaFlair G. T., & Settles B. (2019). *Duolingo English Test: Technical manual* (Duolingo Research Report). https://s3.amazonaws.com/duolingo-papers/other/Duolingo%20English%20Test%20-%20Technical%20Manual%202019.pdf

Lampropoulou, L., Milanovic, M., Jones, J., & Green, A. (2024), LANGUAGECERT Academic Concordance Report. LANGUAGECERT.

Pommerich, M. & Dorans, N. J. (Eds.) (2004). *Concordance* [Special issue]. Applied Psychological Measurement, 28(4), 216–289.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Taylor, L., & Chan, S. (2015). *IELTS Equivalence Research Project* (GMC 133)

Chapter 10: Mobile-based English language learning through bite-sized formative CEFR-aligned assessment

Michael Milanovic and Catherine Jones

Abstract

This paper reports on a large-scale pilot implementation of CATs Core, a mobile English language learning and assessment system covering CEFR levels A0 to B1. Conducted in operational educational contexts, the study focuses on usability, engagement, and indicative learning outcomes in real-world settings, rather than on controlled experimental validation. The analyses are therefore descriptive and exploratory, providing a foundation for subsequent, more rigorous validation research. CATs Core forms part of the wider CATs system, which comprises CATs Primary (A0 to A1) for young learners aged 4 to 10, CATs Core (A1 to B1) for school students and young adults, and CATs Higher (B2 to C1) for students in higher education and workplace professionals.

The CATs system uses frequent formative testing of small, achievable, CEFR-aligned learning 'steps', recognising the diversity of learner proficiency, pace and learning style in classrooms. While functional as a standalone system, CATs is designed to supplement conventional classroom instruction, supporting students requiring additional support while stretching those ready for greater challenge.

The pilot involved 2,773 learners across five higher and further education institutions enrolled in CEFR-aligned programmes at levels A1 to B1. Learner engagement in the pilot was strong with a 63% completion rate. Learners demonstrated measurable gains during their use of CATs averaging 48.1 points on the CATs CEFR-aligned scale over 12 weeks, equivalent to almost one-third of a CEFR level. Learner satisfaction was high: 78% wished to continue using CATs and 81% reported perceived improvement in English proficiency. Although gains cannot be attributed exclusively to CATs, the combination of sustained engagement, measured improvement and positive learner perceptions provides encouraging evidence for the approach.

Keywords: Mobile-assisted language learning (MALL), formative assessment, CEFR, Retrieval practice (testing effect), Mastery learning

Introduction

The challenge: disappointing outcomes in formal language education

English language education attracts substantial investment worldwide, yet completion and success rates remain persistently low. The European Survey on Language Competences (ESLC)[1], which studied 50,000 participants from 15 European education systems, illustrates the scale of the problem. Even in Sweden, with its strong English language education tradition, only 57% of students reached CEFR level B2 by age 16 (European Commission, 2012). Across participating countries, outcomes ranged from 2% to 57%, pointing to systemic problems in structure and delivery of English language teaching.

Low success rates are evident across educational contexts, including secondary schools, universities, adult education programmes, and commercial language schools. Many learners give up before achieving their goals, while those who persist often fall short of proficiency targets despite months, or even years of instruction. When designing the CATs system, we therefore asked a fundamental question: how can success rates in CEFR-aligned language education be improved?

Language learning is inherently challenging and time-intensive. Estimates suggest that progression through a single CEFR level typically requires 100 to 200 hours of study, meaning the A1 to B1 journey may require 500 to 600 hours of sustained effort. During this time, learners may study for months without receiving clear evidence of progress and this can create motivational problems. Traditional testing tends to occur infrequently, perhaps once per term and again in a final examination. Limited feedback means learners discover gaps in understanding only after weeks of instruction and this is often too late for remediation. Conventional approaches also tend to separate instruction from testing, with testing positioned as a retrospective measure of learning rather than as an integral part of the learning process itself.

This study reports on a large-scale pilot implementation of the CATs system in operational educational contexts. The study focuses on usability, learner engagement, and indicative learning outcomes in authentic educational settings, rather than on controlled experimental validation. As such, the analyses presented are descriptive and exploratory in nature, providing a foundation for subsequent, more controlled validation research.

Assessment-oriented learning: the theoretical foundation

The CATs system treats assessment as a primary driver of learning, rather than as a purely evaluative tool. This approach draws on research showing that well-designed testing enhances learning outcomes.

The testing effect (also known as retrieval practice) demonstrates that actively retrieving information from memory strengthens learning more effectively than additional study alone (Roediger & Karpicke, 2006). In language learning, retrieving vocabulary, grammatical structures, or communicative strategies through test tasks can support stronger memory and more automatic access than passive review. CATs embeds frequent testing throughout the learning process, transforming every evaluation into a learning event.

Retrieval practice is most effective when it is challenging yet achievable. In other words, learners must work to recall information but ultimately succeed. This principle of "desirable difficulty" (Bjork & Bjork, 2011) shapes how CATs targets appropriately challenging levels. If tasks are too easy, they provide little benefit because recall is automatic. If they are too difficult, they risk frustrating learners without successful retrieval.

Formative assessment research further supports this approach. Black & Wiliam's (1998) meta-analysis of more than 250 studies demonstrated that assessment designed to provide feedback for improvement, rather than grades alone, produces significant learning gains, with effect sizes from 0.4 to 0.7, representing improvement from the 50th to the 65th to75th percentile. Immediate, specific feedback helps learners identify misunderstandings, correct errors and adjust strategies. Frequent, low-stakes testing informs ongoing learning and can lead to improved performance in high-stakes exams that document achievement retrospectively. CATs operationalises these principles by providing multiple formative tests at each sub-level.

Mastery learning frameworks, developed through Bloom's (1968) research and later refined by Guskey (2010), show that when instruction is followed by formative testing and opportunities for additional corrective practice, nearly

all learners can achieve high levels of mastery. Differences in outcomes reflect differences in time and opportunity to learn rather than innate ability. Language learning exemplifies this. We all learn at least one language, proving capacity exists. By breaking learning into small units, testing mastery of each, and providing support where needed we can improve success significantly. CATs applies these principles by subdividing CEFR levels into smaller units, each requiring demonstrated mastery before progression. Learners must complete certification tests at each sub-level with more than 70% mastery before advancing, so progression reflects demonstrated competence rather than time spent.

Testing, when frequent, formative, and aligned with clear objectives, drives learning rather than merely measuring it. CATs combines these principles in a mobile-delivered system that makes frequent, high-quality testing practical and scalable.

The CATs innovation: small chunks, frequent testing, demonstrated mastery

What distinguishes CATs is not mobile delivery alone, as many systems offer that. The innovation lies in how CEFR-aligned learning is structured and validated.

CATs Core divides the three CEFR levels from A1 to B1 into nine manageable sub-levels (A1.1 to B1.3). Rather than working toward a single distant goal such as A2 (100 to 150 hours), learners complete three smaller, clearly defined steps: A2.1, A2.2, and A2.3, each representing approximately 40 to 50 hours of learning. This structure transforms overwhelming goals into achievable milestones, enabling learners to see tangible progress within weeks rather than months or years. Regular certification at each sub-level provides frequent validation and sustains motivation, particularly during the foundational A1 to B1 stages where, as the ESLC shows, attrition is highest.

More fundamentally, CATs positions testing as the central organising principle of learning, rather than an occasional interruption to instruction. Each sub-level includes regular formative practice after every two instructional units for

feedback on mastery, two complete practice tests simulating certification conditions, and certification tests validating achievement before advancement. Multiple evaluations per sub-level across 40 to 50 hours mean learners engage with testing as continuous learning, not occasional obligation.

The experience differs from traditional instruction-then-test models. Conventional approaches position tests as summative evaluations determining whether learning occurred. CATs embeds testing throughout, continuously informing and reinforcing learning. When practice reveals struggles with the past tense or listening comprehension of numbers, learners can immediately focus additional practice. Continuous feedback helps learners identify gaps before they compound. Every test measures current competence while strengthening it through measurement.

Mobile delivery makes this test-rich approach practical for supporting classroom instruction. Automated assessment of all four language skills (listening, reading, writing and speaking) provides immediate feedback that would be impractical with human raters. Learners get immediate results, enabling multiple cycles of testing, feedback, and practice in a single session.

Research objectives: validating the approach

This pilot study evaluates whether CATs pedagogy produces the engagement and learning outcomes predicted by its theoretical foundations. Two primary objectives guided this research.

The first objective addressed practical viability: whether frequent testing of small CEFR-aligned language points produces measurable proficiency gains while maintaining learner engagement? Does the approach work in practice? Will learners accept frequent testing, or will they find it burdensome despite theoretical benefits? Does it drive outcomes and sustain motivation over meaningful timeframes? We examined whether learners completing sub-levels showed proficiency gains, found the test-rich approach motivating rather than stressful, and whether the granular structure reduced dropout relative to typical online learning.

Second, the study aimed to establish a foundation for future validation research examining whether proficiency gains demonstrated within CATs transfer to external measures, particularly international exams referencing CEFR frameworks.

Three research questions guided the analysis:

1. How do learners react to frequent testing (e.g. motivation, perceived usefulness and stress), and how manageable is the system for coordinators, as evidenced through questionnaire data and focus group feedback?

2. Does frequent formative testing of small CEFR-aligned language points produce measurable proficiency gains, as evidenced through system-generated assessment data?

3. Can a test-rich mobile approach work be implemented effectively across diverse institutional contexts, as evidenced through system usage data and coordinator feedback?

Success on these dimensions – learner experience, learning outcomes and institutional feasibility – would suggest this approach has the potential to address persistent challenges in CEFR-aligned language education, with implications for instructional design.

CATs System: design and implementation

Implementing learning theory through system design

CATs operationalises three complementary learning frameworks: retrieval practice, formative assessment, and mastery learning.

A fundamental principle of the CATs system is that learners completing frequent CATs tests engage in retrieval practice that strengthens competencies better than reviewing grammar explanations or reading examples alone. This aims to leverage the testing effect, where active recall creates stronger memory traces than passive study. The "desirable difficulty" principle (Bjork & Bjork, 2011) guides task design to target appropriately challenging levels. When tasks are too easy, they provide little benefit because recall is automatic. When they are too difficult, they can be a cause of frustration.

Black and Wiliam's (1998) work on formative assessment explains why CATs design works. Their meta-analysis of more than 250 studies found interventions emphasising formative testing improved performance from the 50th to the 65th to the 75th percentile. Key characteristics align with how CATs operates – tests occur frequently, not occasionally; results arrive immediately, not after delays; learners can act on feedback through additional practice.

Bloom's mastery learning principle operates through the sub-level structure. Learners must complete certification tests at each sub-level, demonstrating more than 70% mastery before advancing. Progression depends on demonstrated competence, not time spent. A learner's B1.2 certificate indicates actual B1.2 competence, unlike traditional course completion certificates that may or may not reflect proficiency. So, progressively through any given level, the aim is to firm up the foundation so that addressing the next rung on the ladder comes from a position of competence.

All three frameworks (i.e., retrieval practice, formative assessment, and mastery learning) point to the same conclusion: testing, when frequent, formative, appropriately challenging, and connected to clear mastery criteria, becomes a powerful learning mechanism. CATs builds an entire system around this idea, making testing the continuous backbone of learning rather than an occasional event.

System architecture

The CATs system comprises four integrated components that work together to deliver learning at scale.

Component 1 – Granular CEFR progression

CATs subdivides levels A1 to B1 into nine sub-levels (A1.1, A1.2, A1.3, A2.1, A2.2, A2.3, B1.1, B1.2, B1.3), addressing motivation and persistence challenges inherent in lengthy language programmes. Full CEFR levels represent daunting goals requiring 100 to 200 hours over months or years. Sub-levels transform these into achievable milestones reachable in 6 to 8 weeks with each corresponding to roughly 40 to 50 hours of content and concluding with certification validating mastery. This structure provides regular validation through certificates documenting specific achievements, rather than extended periods without tangible progress markers. The psychological impact is significant: learners see concrete advancement every few weeks rather than waiting months or years.

Component 2 – Preparation materials

Each sub-level provides 30 topic-based units addressing real-world communication situations. Unlike traditional approaches where content is simply delivered for absorption, CATs materials prepare learners explicitly for assessment. After every two units (approximately 3 to 4 hours of learning), practice activities test recently covered material, providing immediate feedback and identifying areas requiring focus. Supplementary modules (Vocabulary Extra, Grammar Extra, Video Extra) provide targeted practice for specific competencies evaluated in certification tests. These resources align

with test requirements, enabling learners to focus on the skills certification will evaluate. Two complete practice tests per level familiarise learners with formats, reduce anxietyand allow self-evaluation of readiness.

Component 3 – Multi-layered testing framework

CATs embeds testing throughout the learning experience at multiple layers, each serving distinct purposes:

Diagnostic placement occurs at programme entry, using adaptive algorithms to determine each learner's proficiency within 20 to 30 minutes, ensuring learners begin at appropriate levels to maximize efficiency and maintain engagement. Placement that is too high can cause frustration; placement that is too low wastes time on mastered material.

Formative practice occurs after every two instructional units, providing regular feedback about mastery. These low-stakes assessment opportunities strengthen learning through retrieval practice, identify gaps, validate understanding before learners build further on that foundation, and maintain progress awareness. The regular rhythm (every 3 to 4 hours) prevents accumulating misunderstandings.

Practice tests (two per sub-level) simulate certification conditions across all four skills under realistic time constraints. These serve preparatory and evaluative functions: familiarising learners with test formats and timings, identifying areas requiring additional preparation, and building confidence through successful performance.

Certification tests validate mastery at completion of each sub-level, requiring 70%+ performance across all four skills before advancement. These evaluations gate progression, ensuring it reflects demonstrated competence. A learner achieving B1.3 certification has achieved B1.3 competence as CEFR defines it.

Component 4 – Technical infrastructure

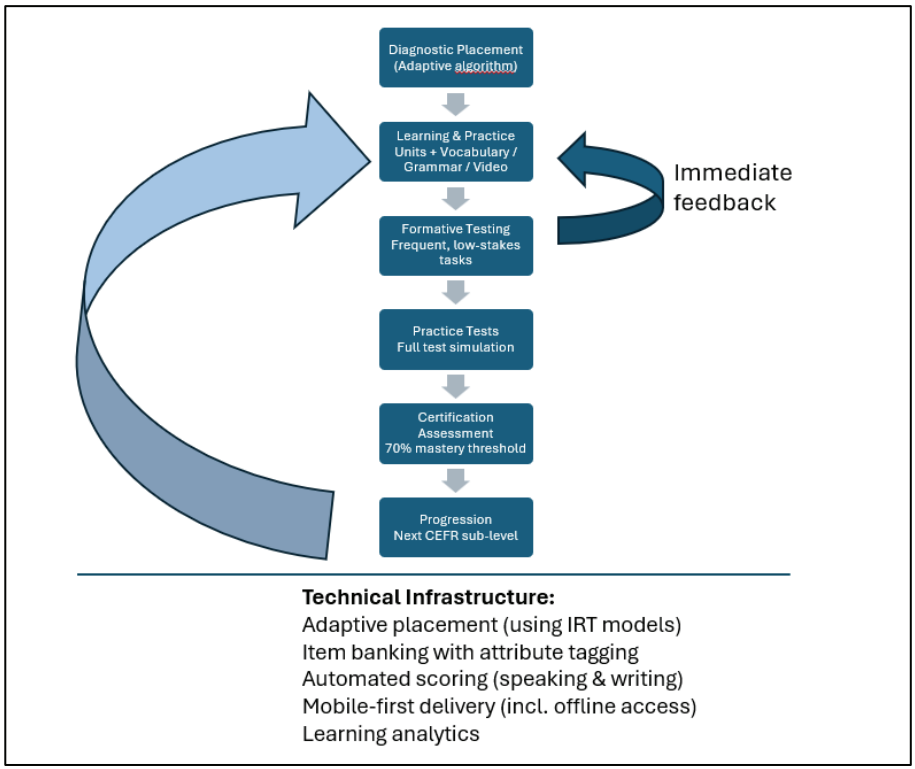
CATs uses automated scoring across all four skills, including speaking, via speech recognition and writing via algorithmic evaluation. This provides immediate, detailed feedback, making frequent, high-quality testing practical

and affordable. Learners receive results immediately rather than waiting days, enabling multiple cycles of testing, feedback, and practice within single sessions.

Sophisticated item banking systems with detailed attribute tagging ensure tests target specific competencies at appropriate difficulty levels. Adaptive algorithms adjust difficulty in response to performance patterns, maintaining optimal challenge throughout progression.

Learning analytics capture detailed information about performance patterns, enabling both individual feedback and system-level refinement. Learners receive specific guidance about strengths and development areas. Coordinators can identify learners struggling with particular competencies and provide targeted support. System developers can identify items or materials requiring refinement. This data infrastructure supports continuous improvement in both outcomes and system quality. Figure 1 provides an overview of the CATs system architecture and learning cycle, illustrating how adaptive placement, structured learning and multi-layered assessment are integrated within a continuous feedback loop. In this model, assessment functions not as a terminal event but as an ongoing mechanism of learning, with system infrastructure enabling this process at scale.

Figure 1. Overview of the CATs system architecture and learning cycle



CATs pilot study methodology

Research design

The pilot study employed a mixed-methods design combining quantitative analysis of learner performance data with qualitative evidence from questionnaires and focus groups. The research design enabled evaluation of both learning outcomes and implementation experiences.

Participants and context

A total of 2,773 learners from five higher and further education institutions participated in the pilot study. Institutions included two community colleges and three polytechnic institutions, with cohort sizes ranging from approximately 100 to more than 1,100 learners. This diverse institutional representation enabled analysis across differing organisational contexts and baseline proficiency profiles.

Representatives from each institution attended remote training sessions covering system administration, learner registration and progress monitoring. Two central pilot coordinators, based within participating institutions, facilitated regular collaboration meetings and provided ongoing implementation support.

The CATs intervention

The pilot deployed CATs Core (A1.1 to B1.3) as described in Section 2, including preparation materials, practice activities, practice tests and certificated assessments. Each sub-level provided 40 to 50 hours of structured content organized into 30 topic-based units with access to supplementary modules (Vocabulary Extra, Grammar Extra, Video Extra).

The system's technical infrastructure included adaptive placement algorithms using item response theory (IRT), automated speech recognition for speaking assessment, and algorithmic scoring for writing assessment.

Implementation process

The pilot followed four phases designed to minimize institutional burden while maintaining consistent procedures across sites and ensuring data quality:

Phase 1: Institutional setup and training

Representatives from participating institutions attended remote training sessions covering system administration and learner support procedures. Each institution designated coordinators responsible for code distribution and progress monitoring.

Phase 2: Learner registration and placement

Learners registered using institutional codes, completed demographic questionnaires, and undertook adaptive placement testing. Placement results determined appropriate starting levels.

Phase 3: Learning engagement

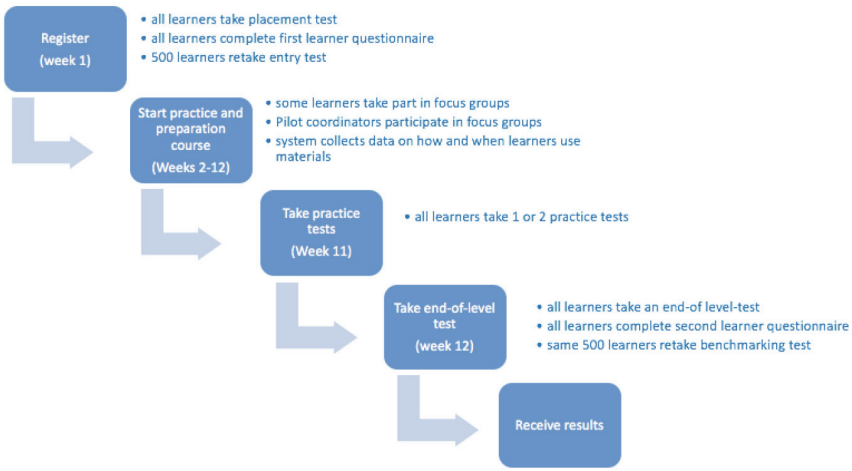
Learners accessed preparation and practice materials via a mobile phone or tablet over a 12-week period. Web-based delivery was not available during this pilot. The platform recorded usage data including time spent, completion rates, and performance on practice activities.

Phase 4: Assessment and certification

Learners who completed preparation and practice tests undertook final certification assessments. All assessments were automatically scored with results linked to learner profiles for analysis.

Figure 2 provides an overview of the pilot implementation process, illustrating how learning, assessment and data collection were integrated across the 12-week study period to support both learner progression and evaluation of the system.

Figure 2. Overview of the CATs pilot implementation process



Data collection

Four principal sources of data were used:

Learner questionnaires

Learners completed two online questionnaires, collecting demographic information, learning background, and evaluation of system effectiveness and satisfaction.

Learners completed the first learner questionnaire at the point of registration. This short questionnaire asked for some basic personal details (e.g. gender, date of birth, department and subject, national examination result for English) and information about learners' English-language context (e.g. reasons for learning English, years learning English, time spent learning English per week, access to English-language materials and media). The first questionnaire asked students to rate on a scale of 1 to 5 how difficult they find different

aspects of English language learning. This question was later repeated in the second questionnaire.

The second learner questionnaire was completed at the end of the pilot, after the end-of-level final test. This longer questionnaire included further questions about the learner (e.g. level of schooling, English proficiency of parents) and questions about the user experience learning with CATs (e.g. how effective, relevant and engaging learners found different aspects of the materials and the system).

Focus groups

Separate focus group sessions were conducted with learners and institutional coordinators to gather qualitative feedback on user experience, system effectiveness, and implementation challenges.

Learner focus groups took place eight weeks into the pilot. These explored the extent to which CATs supported English-language learning and the specific ways in which learners think this happens. Learner focus groups collected qualitative feedback on the user experience, the functionality of the system, the usability of the materials, the language activities learners find most helpful, and individual learning styles.

Pilot coordinator focus groups were designed to explore the experience of managing CATs from the institution administrator's point of view and collect qualitative feedback on running and delivering CATs. Two focus groups took place with pilot coordinators from each institution. The first took place eight weeks into the pilot and the second at the end of the pilot process. The focus groups were facilitated by the CATs Team and conducted remotely.

System-generated data

The CATs platform automatically captures comprehensive data on learner activity, including time spent, completion rates, assessment scores, and detailed interaction patterns.

During the pilot the following information was captured and stored for all learning and assessment materials:

- User System Fields (Unique CATs User Identifier, System Identifier, Device Identifier)
- Question Data (Task ID, Question ID, Answer given, Score, Maximum Possible Score)
- Task Data (Task ID, Timestamp, Score, Maximum Possible Score, Number of times this task has been opened, Open duration)
- Task Progress (Unit ID, Task ID, Maximum Score, Latest Score, Latest Score Date, Highest Score, Highest Score Date)
- Unit Progress (Unit ID, Unit Score, Unit Maximum Score, Last Attempt Date).

Assessment scores

Pre- and post-intervention assessment scores provided quantitative measures of learning effectiveness, with all assessments calibrated to CEFR levels using IRT scaling methodology.

The CATs platform captures extensive granular data on learner interactions and performance. While aggregated metrics were available for the present analysis, more detailed exploitation of item-level and interaction-level data was beyond the scope of this study and represents an important direction for future research.

Data analysis

The data analysed in this study were generated through system use and learner feedback mechanisms within an operational pilot. As such, they were not originally collected within a controlled experimental design. The analysis therefore focuses on descriptive patterns in engagement, performance and user experience, rather than on hypothesis testing or causal inference.

Quantitative analysis focused on learning gain measures, with proficiency improvement calculated as the difference between adaptive placement test scores and end-of-level assessment scores on the 0–900 CATs scale. The CATs scale is constructed using Rasch measurement methodology, with scale values representing positions on a unidimensional proficiency continuum. A difference of 50 points on this scale corresponds to progression through one CATs sub-level, which represents one-third of a full CEFR level.

Qualitative data from focus groups and questionnaires were analysed thematically to identify key factors influencing implementation success and learner satisfaction.

Results

Participation and engagement

1,738 of 2,773 registered learners (62.7%) completed end-of-level assessments, comparing favourably with typical online learning completion rates. Learners completed an average of 31% of available materials overall (48% excluding non-engagers). Critically, learners completing the first 10% of materials showed significantly higher continuation rates, indicating the importance of initial engagement for sustained preparation effort.

Baseline proficiency assessment

Placement test results revealed proficiency distributions consistent with the target population for CEFR levels A1 to B1. Learners from polytechnic institutions typically demonstrated proficiency around CEFR levels A2/B1, while those from community colleges showed slightly lower baseline levels. The placement testing successfully differentiated learners across the target proficiency range, with appropriate score distributions across the nine CATs sub-levels. Standard deviations indicated substantial variation in ability within institutions, supporting the value of individualized placement rather than institution-wide approaches.

Learning outcomes: proficiency improvements

The primary outcome measure, change in proficiency between placement and end-of-level assessments, demonstrated substantial learning gains.

Overall learning gains

Mean gain (CATs scale)	48.1
Equivalent CEFR progress	~0.33 level

Learners achieved an average improvement of 48.1 points on the CATs measurement scale over 12 weeks. Since the CATs scale is calibrated such that 50 points represents one CATs sub-level (equivalent to one-third of a CEFR level), this represents progress of approximately 0.96 CATs levels, or nearly one-third of a full CEFR level. While gain scores provide a useful indicative measure of progress within the CATs scale, they should be interpreted with caution. More robust approaches to measuring learning impact, including controlled comparisons and external validation against independent assessments, are required to draw definitive conclusions about effectiveness.

Distribution of improvements

Analysis of learner distribution across proficiency levels showed clear evidence of upward progression. Between placement and end-of-level testing, substantially fewer learners remained at A2 level, with greater proportions progressing to B1 or B2 levels.

Figure 3 illustrates the distribution of learners across CEFR-aligned sub-levels at placement and end-of-level assessment. The observed shift towards higher levels is consistent with the gain scores reported, although, as noted, these changes cannot be attributed solely to CATs.

Figure 3. Distribution of learner proficiency levels at placement and end-of-level assessment

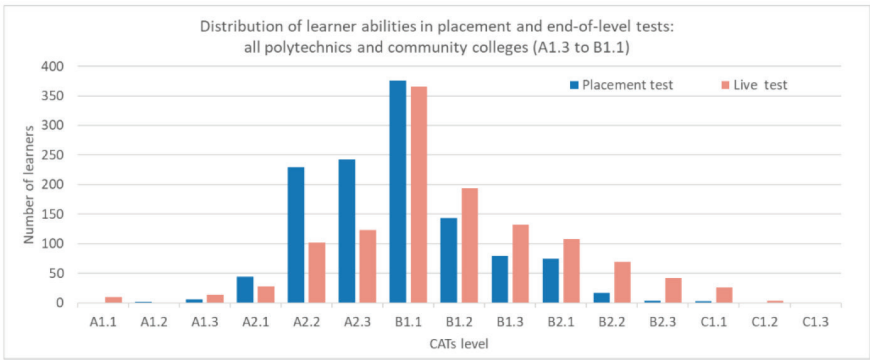
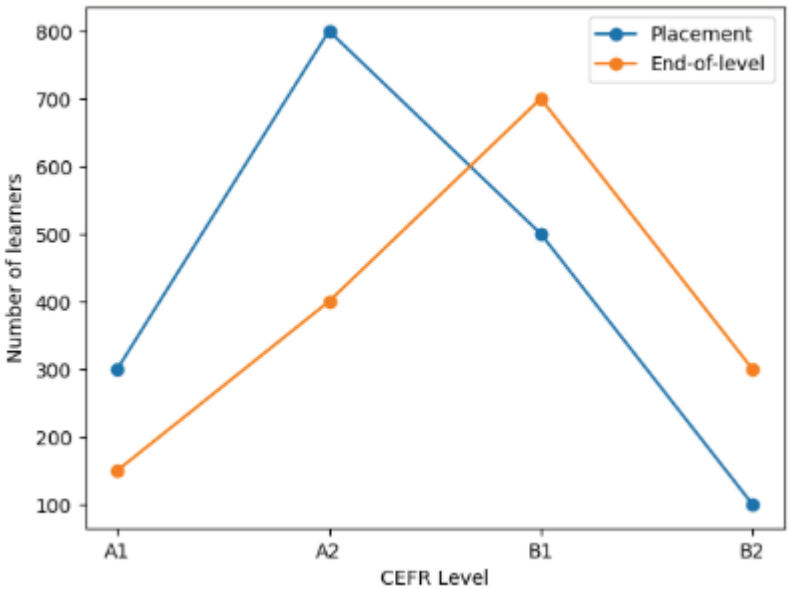


Figure 4 presents an aggregated view of learner distribution across CEFR levels at placement and end-of-level assessment. The observed shift towards higher levels is consistent with the gain scores reported.

Figure 4. Distribution of learner proficiency: placement vs end-of-level



Institutional variation

The greatest gains were observed among learners at community colleges, with one institution demonstrating an average improvement of 83 points, equivalent to nearly two-thirds of a CEFR level. This pattern suggests that mobile-accessible preparation may be particularly beneficial for learners who face greater barriers in traditional preparation contexts.

Assessment performance quality

End-of-level test scores reinforced the proficiency gains observed through the measurement scale. Average raw scores ranged from 38 and 40 out of 50 possible points, with most learners achieving scores above 70%. These results exceeded typical mastery thresholds and were observed consistently across all tested levels, indicating that learners developed genuine competence within the framework assessed by CATs.

Learner experience and satisfaction

Despite relatively low response rates to end-of-pilot questionnaires, the available data and supporting qualitative feedback provide useful insights into learner perceptions of the system. Available satisfaction data indicated highly positive learner experiences: 78% of respondents expressed desire to continue using the CATs system, 76% reported enjoying use of the application and 81% believed their English proficiency had improved through system use.

Focus group discussions revealed that learners particularly valued immediate feedback on practice activities, flexible access enabling study at convenient times and locations, clear progress indicators showing advancement toward proficiency goals and variety in learning activities maintaining engagement over extended periods. These satisfaction indicators suggest that preparation approaches like CATs can sustain the extended engagement necessary for effective language development.

Taken together, these findings suggest that learners particularly value immediacy of feedback, flexibility of access and clear indicators of progress, all of which are central to the CATs design.

Institutional implementation feasibility

Focus groups with institutional coordinators provided insights critical for understanding scalability potential:

Technical implementation

Coordinators reported successful system implementation across diverse institutional contexts with minimal technical difficulties. The registration process functioned smoothly even with large student cohorts, and internet bandwidth concerns proved largely unfounded. This technical reliability supports institutional confidence in the system.

Administrative efficiency

After initial training, coordinators found they could manage the system with minimal ongoing support from technical staff. This finding has significant

implications for scalability, as it suggests that large-scale implementation would not require extensive specialised infrastructure.

Learner support requirements

While most learners engaged successfully with materials, coordinators identified that students with initially low motivation required additional encouragement. They suggested that automated progress notifications and social media integration could enhance engagement.

Monitoring capabilities

Coordinators found the learner monitoring system effective for tracking both individual and group progress. The ability to identify learners requiring additional support proved valuable for pastoral care.

System reliability and technical performance

The CATs platform demonstrated high reliability throughout the pilot period, with no major technical failures or system crashes reported. This stability proved crucial for maintaining learner engagement and institutional confidence. The mobile-first design proved successful in accommodating diverse learning contexts, with successful engagement across various device types and demonstrated feasibility of offline content access in contexts with limited connectivity.

Discussion

CATs effectiveness for English language development

The average improvement of 48.1 points represents meaningful progress approaching one-third of a CEFR level within 12 weeks, comparing favourably with published research on similar interventions. This suggests that well-designed mobile systems can produce measurable outcomes within reasonable periods. Gains occurred consistently across institutional contexts and proficiency levels, with particularly strong results among community college learners, suggesting mobile-based learning may especially benefit learners facing traditional access barriers.

Theoretical implications for CEFR-aligned examination preparation

CATs and major CEFR-aligned examinations share construct alignment through common CEFR descriptors and emphasis on communicative competence. This theoretical foundation suggests competencies developed through CATs should transfer to examination performance. However, this requires empirical validation examining:

Assessment format correspondence

While CATs and major examinations share CEFR alignment, specific task types and response formats may differ. Research should examine whether competencies developed through CATs automated assessment tasks transfer to different assessment formats.

Proficiency range considerations

The current pilot examined effectiveness within A1 and B1 levels. Transfer relationships at higher proficiency levels (B2 and C2) remain empirically

unexamined and may differ given the increased complexity of linguistic and communicative demands.

Preparation duration adequacy

While 12 weeks proved sufficient for measurable gains on CATs assessments, optimal preparation periods for specific examination targets require investigation, particularly considering that progression through full CEFR levels typically requires substantially longer engagement.

Mobile delivery advantages for language learning access

The successful mobile-first implementation addresses several key challenges in language learning access. The elimination of geographic and scheduling barriers through mobile access significantly expands potential learner reach, particularly benefiting learners in remote areas or those with complex work/study schedules.

The individualized pacing enabled by mobile delivery contrasts favourably with fixed-schedule classroom instruction. Learners could adjust their learning intensity based on their circumstances, enabling both intensive short-term engagement and extended skill development approaches.

The high learner satisfaction rates suggest that mobile-based learning can sustain the extended engagement necessary for meaningful language development. The immediate feedback and progress tracking features appeared particularly valuable for maintaining motivation over extended periods.

Scalability for institutional implementation

The minimal institutional implementation burden reported by coordinators indicates strong scalability potential. The ability to manage large cohorts with minimal specialized staff suggests that CATs-based learning could be efficiently delivered across diverse institutional contexts.

The system reliability and technical performance demonstrated throughout the pilot support confidence in large-scale delivery. The absence of major technical issues and successful accommodation of varying connectivity contexts indicate robust platform capabilities.

The positive coordinator feedback regarding monitoring and support capabilities suggests that institutional implementation of CATs provides adequate oversight and learner support mechanisms for responsible program delivery.

Study limitations and future research directions

Several limitations affect the interpretation and generalizability of these findings:

Absence of external validation

A key limitation of this study is the absence of external validation measures. All learning gains were assessed using the CATs system itself and, while these assessments are CEFR-aligned and psychometrically informed, the findings cannot yet be directly linked to performance on independent CEFR-aligned examinations. Establishing such relationships represents a critical next step for evaluating the broader validity of the approach.

Limited proficiency range

The pilot focused on learners within A1 to B1 CEFR levels. Conclusions about CATs effectiveness for higher proficiency levels cannot be drawn from these data. The linguistic and communicative demands at B2 to C2 levels differ substantially, and effectiveness at these levels requires separate investigation.

Timeframe constraints

The 12-week study period, while sufficient to demonstrate learning gains, does not address longer-term retention or sustained proficiency development. Longitudinal research examining retention and continued progression would provide a more complete understanding of learning trajectories.

Completion bias

The 62.7% completion rate, while reasonable for online learning contexts, means that findings represent learners who sustained engagement with the system. The characteristics and outcomes of non-completing learners remain less well understood, which may limit the generalisability of the results.

Cultural and linguistic context

All participants were drawn from specific institutional and geographic contexts. Effectiveness may vary across different cultural settings, first language backgrounds and educational traditions, and further research is needed to explore these dimensions.

Concurrent learning effects

Learners continued their regular classroom studies while engaging in the CATs pilot. As a result, it is not possible to isolate the specific contribution of CATs to the observed improvements and causal inferences regarding its impact cannot be drawn.

Data scope and analytical depth

While the CATs platform captures extensive granular data on learner interaction and performance, the present study draws primarily on aggregated metrics generated during the pilot. More detailed analyses of item-level performance, learning behaviours and interaction patterns would enable a deeper understanding of how specific features of the system contribute to learning.

Future research directions

Taken together, these limitations indicate that the current findings should be interpreted as indicative rather than definitive. This study establishes an operational and empirical baseline for future research, which should focus on controlled designs, external validation against independent assessments and deeper analysis of learner interaction data to more fully evaluate the effectiveness of the CATs approach.

Conclusions

This study provides indicative evidence that the CATs mobile-based learning system effectively supports English language development within the A1 to B1 CEFR proficiency range. Learners achieved substantial proficiency improvements over a 12-week period, with learning gains averaging 48.1 points on a CEFR-aligned scale, representing approximately one-third of a CEFR level. High learner satisfaction combined with minimal institutional implementation burden suggests strong potential for scalable delivery.

Implications for CEFR-aligned examination contexts

The shared CEFR framework foundation between CATs and major international English language examinations, including those offered by LanguageCert creates theoretical conditions for transfer of competencies developed through CATs to examination performance. The emphasis on communicative competence within authentic contexts in both systems supports this theoretical alignment.

However, the critical question of whether proficiency gains demonstrated through CATs assessments transfer to performance on external CEFR-aligned examinations remains empirically unresolved. While theoretical construct alignment and successful learning outcomes create favourable conditions for such transfer, definitive conclusions require direct validation through examination performance studies.

Contributions and future directions

This study contributes to mobile-assisted language learning literature by providing empirical evidence from a large-scale pilot implementation demonstrating both learning effectiveness and implementation feasibility. The findings establish CATs as a viable mobile-based learning system with demonstrated impact on language proficiency development.

For the broader field of English language learning and assessment, these results demonstrate that mobile-based approaches can produce meaningful learning outcomes while addressing practical barriers of access, cost, and scheduling flexibility. The combination of CEFR alignment, comprehensive skill development, and flexible delivery creates a learning framework that addresses both pedagogical effectiveness and practical implementation challenges.

This study establishes an operational and empirical baseline for future validation work. The next phase of research will focus on linking CATs engagement and performance to outcomes on external CEFR-aligned assessments, enabling a more robust evaluation of its effectiveness as a preparation tool. Longitudinal studies tracking learners from CATs participation through to examination outcomes would provide more definitive evidence of its value as examination preparation and would complete the empirical foundation for mobile-based approaches.

Until such validation studies are completed, claims regarding CATs effectiveness as preparation for specific examinations must remain qualified. The current evidence supports CATs as an effective language learning system with strong theoretical alignment to examination constructs, establishing a foundation that warrants, and indeed requires, further investigation through direct examination performance research.

References

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74. https://doi.org/10.1080/0969595980050102

Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), Psychology and the real world: Essays illustrating fundamental contributions to society (pp. 56–64). Worth Publishers.

Bloom, B. S. (1968). Learning for mastery. Evaluation Comment, 1(2), 1–12.

European Commission. (2012). First European survey on language competences: Final report. Publications Office of the European Union.

Guskey, T. R. (2010). Lessons of mastery learning. Educational Leadership, 68(2), 52–57.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

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.

CERTIFYING QUALITY IN ASSESSMENT AND LEARNING
Research and Validation at LANGUAGECERT Volume 4

ISBN: 978-9925-39-382-4