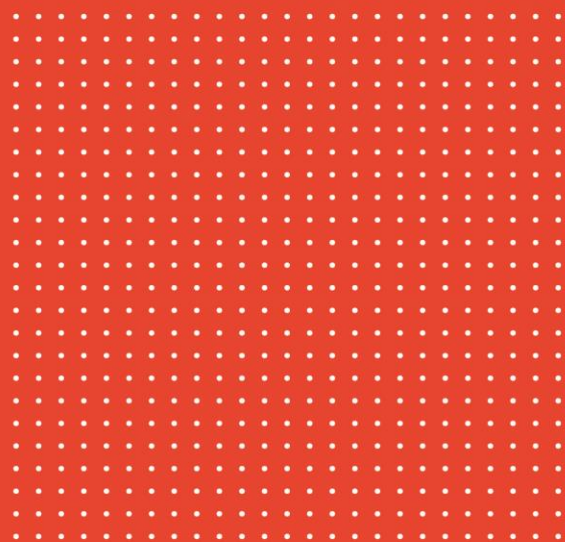


Language
Cert



Cathy Jones

Orienting the
LanguageCert
IESOL C1
Examination
to an Academic
Context



Abstract

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

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

Introduction

This paper is based on Chapter 1 of Falvey and Coniam (eds.), Volume 2 of the LanguageCert series *Certifying Quality in Assessment and Learning*.

It is often said that ‘qualifications open doors’ in the sense that candidates take high-stakes exams to access life-changing opportunities such as university admission or migration for work. In the same way that a door provides access to a new room or space, a qualification can provide access to higher education, career opportunities, experiences and communities. Assessment of all kinds can have a transformative impact on the life chances of individuals and as such there is an ethical and moral responsibility to ensure that, as powerful gateways to new learning experiences, personal growth and professional development, exams are reliable, secure and fit for purpose.

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

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

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

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

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

Background

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

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

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

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

Purpose

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

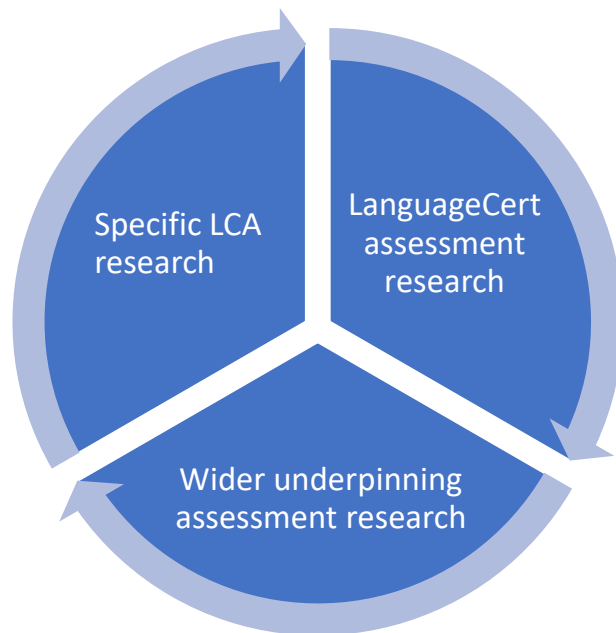
An Evidence-informed Approach

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

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

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

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



Underpinning Evidence

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

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

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

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

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

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

LanguageCert research has also been instrumental in evaluating the stability and robustness of large-scale item banking for high-stakes qualifications. LanguageCert uses a proprietary secure item bank (IB) to manage all tests with strict access protocols and workflows for process compliance. Reports can be run to interrogate the volume of materials at different stages of production in the item bank, by test and task type. Reports can also be run for more detailed information, such as the amounts of materials at certain difficulty levels by test part. Items and tasks are commissioned into the IB with the intention of re-using the material over time and across test versions. A variety of re-use parameters are in place for different types of products and different skills. LanguageCert have explored item bank stability in several research pieces, Lee et al (2023b) through live and simulated datasets and Coniam et al (2023a), in reference to the creation of multiple test forms. Both these pieces have directly supported the development of LanguageCert Academic and LanguageCert General, providing necessary evidence of the integrity and stability of the item banks.

What is Academic English and Why is it Important?

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

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

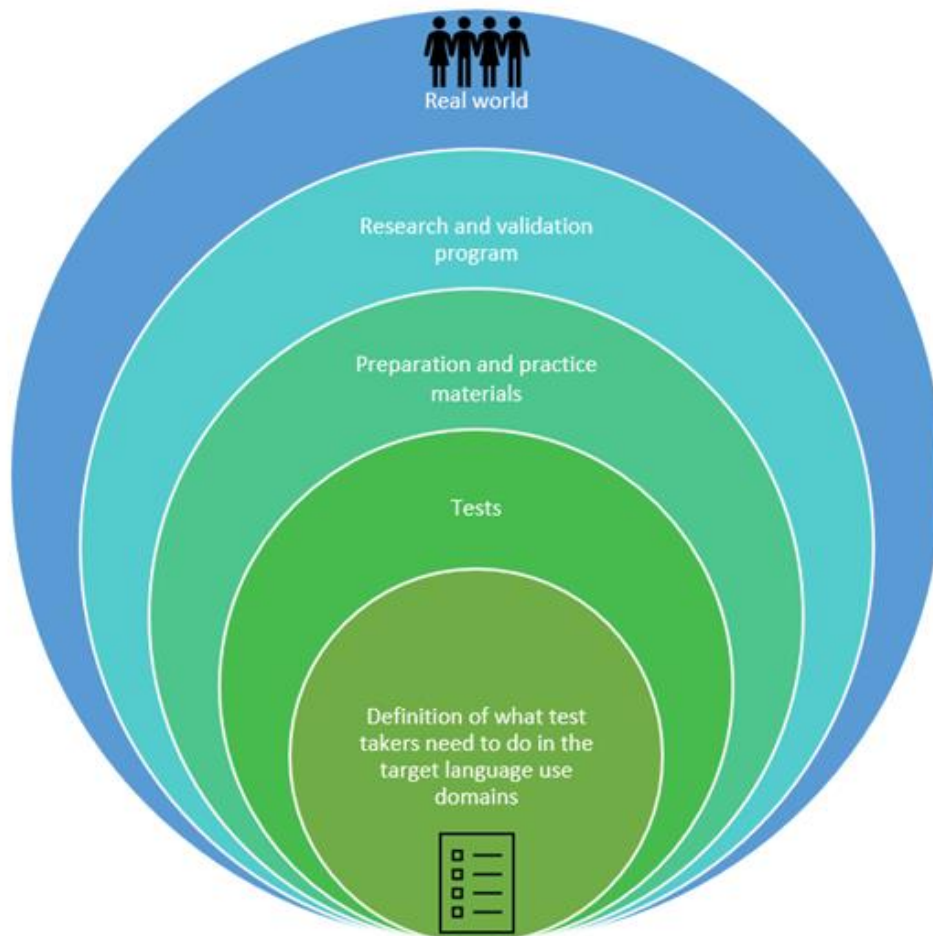
In evolving the IESOL SELT test and populating the item banks with materials appropriate for an academic context, it was essential to understand the skills, competences and cognitive processes that are specific to general academic English. Refocusing the IESOL SELT test involved more than simply recasting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

Defining the Target Language Use Domain

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

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

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



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

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

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

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

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

Washback by Design

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

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

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

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

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

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

Domain Relevance

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

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

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

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

Designing Tests that Measure Language Competence

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

The CEFR divides a learner's competences into General Competences and Communicative language competences. Communicative language competences are then further subdivided into three: Linguistic, Sociolinguistic and Pragmatic competences. These involve consideration not only of the communication, but also of the strategies used by learners, and hence the functional language skills learners demonstrate when they communicate. In a thought-provoking contribution to this area, Lampropoulou (2023), discusses a subset of Pragmatic Competence, namely Interactional Competence (IC). IC is discussed and described within the context of speaking skills where, it is proposed, IC can be assessed most usefully through the methodology of role-playing in a speaking skills task. The data were gathered from LanguageCert tests. This promising development is an example of how the LanguageCert research team constantly seeks innovative, improved and effective methods of assessing language proficiency skills.

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

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

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

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

Testing the Domain Across the Four Skills

This section outlines domain relevance across the skills.

Developing Domain Relevance in the Listening Tests

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

Range of Accents

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

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

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

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

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

Developing Domain Relevance in the Reading Tests

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

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

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

Developing Domain Relevance in the Writing Tests

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

Writing test quality was the focus of a study conducted by Coniam et al (2023c) in which many facet Rasch analysis was used to explore consistency in marking and linkage and calibration to the CEFR. The study found the two extended writing tasks writing tests from which LCA and LCG are derived, robust and fit for purpose. Indeed, the two extended writing task formats are well established in tests of English for academic purposes, including TOEFL and IELTS (Cumming, 2013), as they reflect a range of expository and descriptive task types encountered in academic contexts across disciplines.

Developing Domain Relevance in the Speaking Tests

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

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

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

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

Developing Domain Relevance in the Marking Criteria

Domain-specific mark schemes are employed for LanguageCert Academic.

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

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

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

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

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

In high stakes exams such as LCA, it is essential for examiners to make informed and reliable expert judgements. The role of expert judgement in language test validation was examined in a LanguageCert study (Coniam et al. 2023 b), that established the how examiner familiarity with items, standards and scales affects the accuracy of their judgement. Examiner training and standardisation documentation has been produced for LCA with these key findings in mind.

Reliability and Scoring

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

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

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

The LCA test is a multi-level assessment, unlike the level-specific IESOL tests. Level-based tests however, can also typically measure across multiple levels. For instance, LanguageCert research (Lee et al., 2023a), has shown that the its IESOL SELT level-based tests assess at their target CEFR levels, but also contain an appropriate number of items to allow assessment across levels. Specifically, the IESOL SELT C1 examination has items which assess above and below C1. Likewise, at the B2 level, there are items in the IESOL SELT B2 examination which assess both above and below B2. This feature is extremely useful for stakeholders who have to make decisions about candidates based on their results.

These findings are contained in a study by Lee et al. (2023 a) on aligning LanguageCert SELT examinations to the LanguageCert Item Difficulty scale in which the alignment of LanguageCert IESOL SELTs is explored in relation to the two objectively marked components of Listening and Reading. The use of externally referenced anchoring demonstrated the robustness of the four CEFR test levels B1–C2. For example, in the case of LanguageCert IESOL SELT C1 test, most accurate measurement was observed across two CEFR levels (B2 and C1) and reasonable measurement at the lower end of C2 and upper end of B1.

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

New materials target specific levels as defined in the Item Writer Guides (IWGs). The materials are created by experienced LanguageCert writers and reviewers. Used in combination with calibrated anchor items, LanguageCert are confident that both tests assess across the stated ability range effectively. This is reinforced by ongoing research to locate all LanguageCert assessment products on its underpinning measurement scale, and aligning all LanguageCert products to the CEFR through which equivalence with other qualifications can be drawn.

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

Measurement Scale

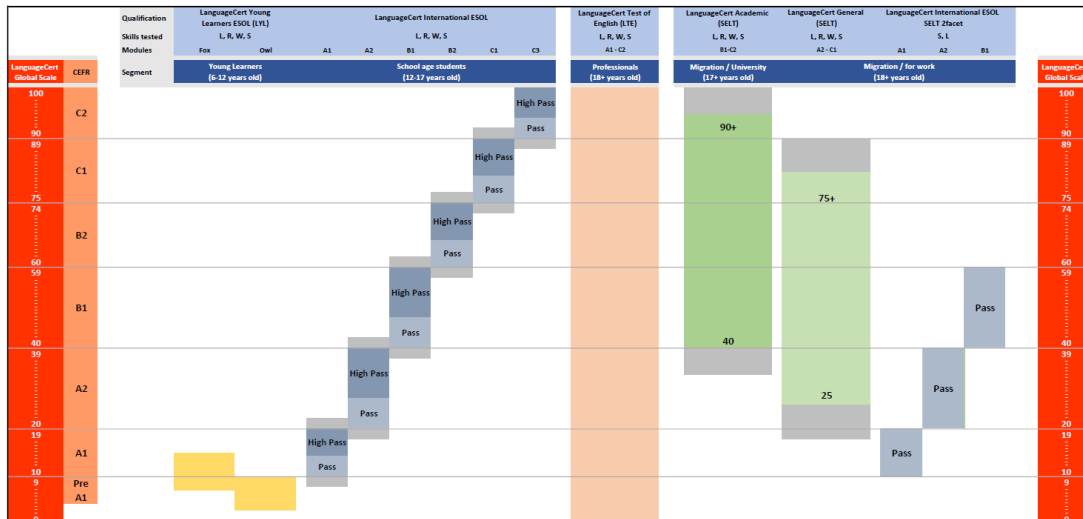
The Global Scale is used to measure each test taker's performance. The Global Scale reports scores on a 0 to 100 scale. These levels of attainment can relate to overall performance in one examination, performance by skill or both these parameters. The Global Scale corresponds directly to LanguageCert's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

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

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

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

Figure 3: The LanguageCert Global Scale



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

The Global Scale allows ease of interpretation for test users and a finely tuned results service across all language skills. As shown, performance can be separated in each skill and overall, so that a test taker is not only described as having 'B2 ability', but a more precise level of detail is provided on test taker's performance. The Test Report shows an overall score, the overall CEFR level of attainment reached, and the score for each of the skills using both the Global scale and the CEFR level of attainment.

The Global Scale, launched with the LanguageCert Test of English (LTE), measures from pre-A1 to high C2 (i.e., across the full 0–100 range). The LTE has been successfully administered to tens of thousands of test takers worldwide, and the Global Scale has received good customer feedback in terms of its simplicity, clarity, and ease of use.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Currently, test taker output in the Speaking test is marked by two human examiners; by the interlocutor immediately after the test and by a second examiner who awards marks subsequently by accessing the video recording. The first criteria 'Task Fulfilment and Communicative Effect' is marked by the interlocutor and provides more of a 'general impression' score, while the second examiner marks the other criteria. The interlocutor general impression mark is then double-weighted. If there is a significant difference in marks awarded, then the recording goes to a third (more senior) examiner whose marks are final.

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

Methodology

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

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

Bias

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

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

For gender – typically a key variable in the exploration of DIF – there was a very low incidence of 3% DIF. An examination of Reading or Listening items indicated that there was no significant DIF in either skill. With the findings confirming that the LanguageCert tests analysed exhibit low levels of gender bias, a methodology is in place for the ongoing monitoring of DIF on all LanguageCert exams. Native language and age showed moderate-to-large DIF. This, however, is likely to be due to these two categories being diverse with only very few entries from small sub-test populations.

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

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

Ongoing Development, Monitoring and Evaluation

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

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

Conclusion

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

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

References

- Bachman, L. F., & Palmer, A.S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.
- Cheng, L. & Sultana, N. (2022). Washback: Looking backward and forward. In Fulcher, G. & Harding, L. (Eds.). *Routledge Handbook of Language Testing*.
- Coniam, D., & Lampropoulou, L. (2020). A review of LanguageCert IESOL Listening and Reading test reliabilities 2018-2020. London, UK: LanguageCert.
- Coniam, D., & Lee, T. (2021). Potential bias in LanguageCert IESOL items: A Differential Item Functioning analysis. London, UK: LanguageCert.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021a). Validating the LanguageCert Test of English scale: The paper-based tests. London, UK: LanguageCert.
- Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021b). Validating the LanguageCert Test of English scale: The adaptive test. London, UK: LanguageCert.
- Coniam, D., Lee, A., & Milanovic, M. (2023a). Exploring item bank stability in the creation of multiple test forms. In Falvey, P & Coniam D. (eds.). *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK.
- Coniam, D., Lee, A., Milanovic, M., Pike, N., & Wen Zhao. (2023b). The role of expert judgement in language test validation. In Falvey, P & Coniam D. (eds.). *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK
- Coniam, D., Stoukou, I., Lee, A., & Milanovic, M (2023c). LanguageCert SELT Writing Test quality. In Falvey, P & Coniam D. (eds.). *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK
- Cumming, A. (2013) *The cognitive validity of the IELTS Academic Writing task*. IELTS collected papers. Cambridge: Cambridge University Press.
- Falvey, P., & Coniam, D. (eds.) (2022). *Certifying quality in assessment: Research and validation at LanguageCert*, Vol. 1. London, UK: LanguageCert.
- Falvey, P., & Coniam, D. (eds.) (2023). *Certifying quality in assessment: Research and validation at LanguageCert*, Vol. 2. London, UK: LanguageCert.
- Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education: 25. Studies in Language Testing*, Series Number 25 Cambridge: Cambridge University Press.
- Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing*, 28, 39-52.
- Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing*, 16(2), 163-188.
- Lampropoulou, L. (2023). Interactional competence and the role roleplay plays: The LanguageCert perspective. In Falvey, P. and Coniam C. *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK.

- Lee, A., Papargyris, Y., Milanovic, M., Pike, N., & Coniam, D. (2023a). Aligning LanguageCert SELT tests to the LanguageCert Item Difficulty scale. In Falvey, P. and Coniam C. *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK.
- Lee, A., Coniam, D., & Milanovic, M. (2023b). Exploring item bank stability through live and simulated datasets. In Falvey, P. and Coniam C. *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert*, Vol. 2. LanguageCert: London: UK.
- Milanovic, M., Lee, A., Coniam, D., Papargyris, Y. (2023a). Externally-referenced anchoring of LanguageCert SELT tests. In Falvey, P & Coniam D. (eds.). *Certifying Quality in Assessment and Learning: Research and Validation at LanguageCert Vol 2*. LanguageCert: London: UK.
- Milanovic, M., Pike, N., Lee, T., & Coniam, D. (2023b). *The LanguageCert Global Scale*. London, UK: LanguageCert.
- Okada, Y. (2010). Role play in oral proficiency interviews: Interactive footing and interactional competencies. *Journal of Pragmatics*, 42, 1647–1668.
- Turner, J. (2004). Language as academic purpose. *Journal of English for Academic Purposes*, 3(2), 95-109.
- Turner, J. (2012). Providing a space for the socio-political dynamics of EAP. *Journal of English for Academic Purposes*, 11(1), 17.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndmills, UK: Palgrave Macmillan.
- Xi, X., & Norris, J. M. (Eds.). (2021). *Assessing academic English for higher education admissions*. Routledge.
- Zwick, R., Thayer, D.T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, 36(1), 1-28.

Appendix 1: Reference List of Specific Skill-based Studies

Listening

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

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

Reading

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

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

Writing

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

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

Speaking

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

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


Appendix 2: Sample Candidate Test Report

Language
Cert

LanguageCert Academic (Listening, Reading, Writing, Speaking)

Test Report

Candidate Information

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

Test Centre Information

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

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

Listening		Writing	
Reading		Speaking	
Total Score			
CEFR Level			



Marios Molfetas
LanguageCert
Responsible Officer

THIS IS NOT A CERTIFICATE

LanguageCert is a business name of PeopleCert Qualifications Ltd, UK company number 09420926.
LanguageCert reserves the right to amend the information given before issuing certificates to successful candidates.
info@languagecert.org

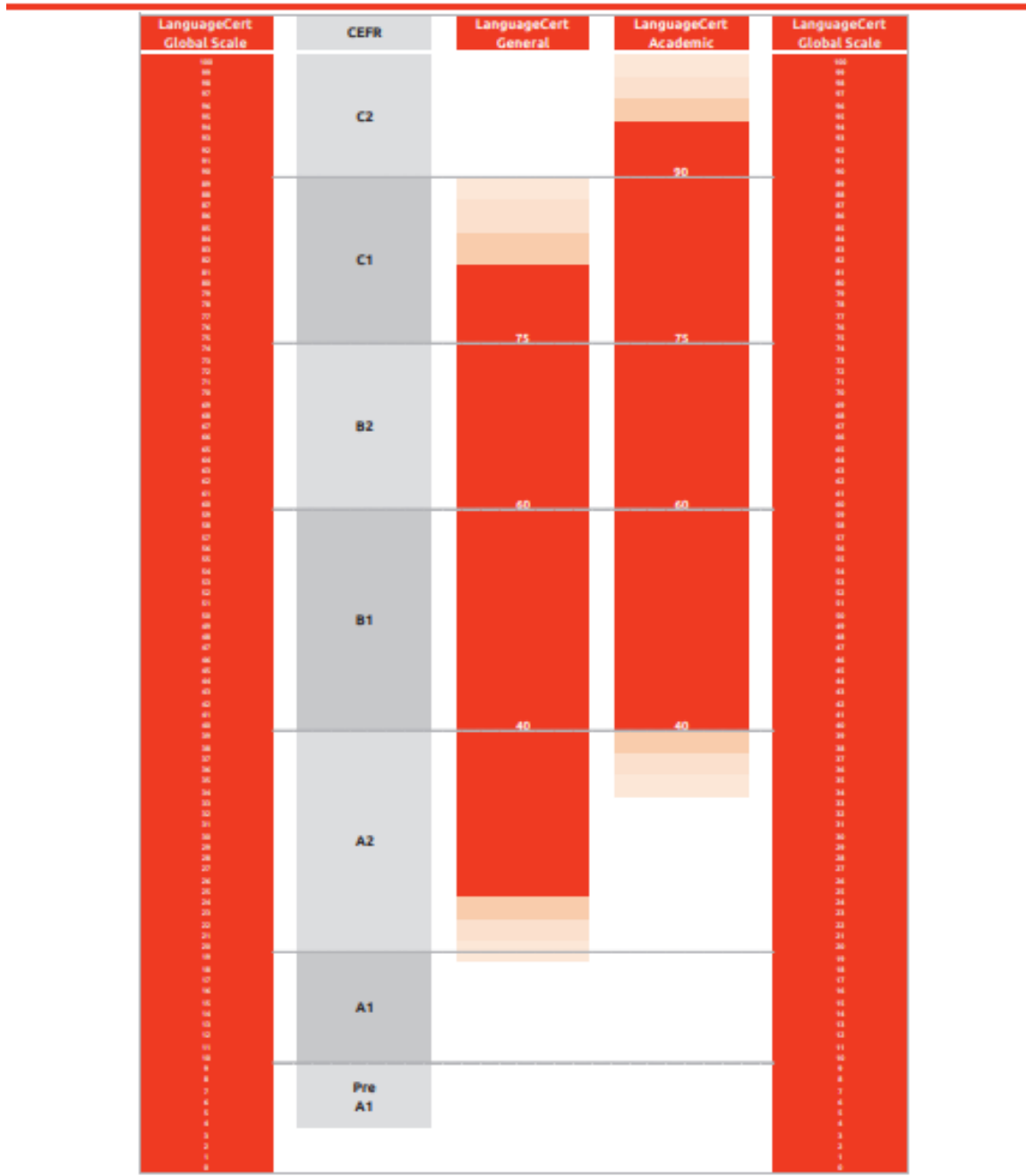
Candidate Performance Feedback (Writing)

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

Candidate Performance Feedback (Speaking)

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

CEFR Level	Scaled Score	Performance Descriptors (Listening, Reading, Speaking, Writing)
C2	90 - 100	<ul style="list-style-type: none"> Can understand with ease any kind of spoken language, provided there is familiarity with the accent. Can read with ease virtually all forms of the written language, including abstract or linguistically complex texts. Can produce clear, smoothly flowing well-structured speech with an effective logical structure which helps the recipient to notice significant points. Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.
C1	75 - 89	<ul style="list-style-type: none"> Can understand an extended speech even when it is not clearly structured and when relationships are only implied. Can read and understand long and complex texts, appreciating distinctions of style. Can give clear, detailed presentations on complex subjects, integrating sub themes, developing points and rounding off with an appropriate conclusion. Can write clear, well-structured texts on complex subjects, underlining relevant issues, expanding and supporting points of view with subsidiary points, reasons and examples, and rounding off with an appropriate conclusion.
B2	60 - 74	<ul style="list-style-type: none"> Can understand extended speech and lectures and follow complex lines of argument provided the topic is reasonably familiar. Can read and understand articles and reports in which the writers adopt particular attitudes or viewpoints. Can give clear, detailed presentations on a range of subjects related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples. Can write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options.
B1	40 - 59	<ul style="list-style-type: none"> Can understand the main points of clear standard speech on familiar matters. Can read and understand texts that mainly consist of high frequency everyday language. Can reasonably fluently give a straightforward description of subjects within his/her field of interest, presenting it as a linear sequence of points. Can write a text on a subject of personal interest, using simple language to list advantages and disadvantages and give his/her opinion.
A2	20 - 39	<ul style="list-style-type: none"> Can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance Can read and understand very short, simple texts such as personal letters Can give a simple description of people, daily routines, likes/dislikes etc. as a short series of simple phrases and sentences linked into a list. Can write a series of simple phrases and sentences linked with simple connectors like 'and,' 'but' and 'because'.
A1	10 - 19	<ul style="list-style-type: none"> Can recognise very familiar words and phrases when people speak slowly. Can read and understand very simple sentences on familiar topics. Can produce simple mainly isolated phrases about people and places. Can write simple isolated phrases and sentences.
<p>The above descriptors are adapted from the Common European Framework of Reference for Languages (2018). Text from these is reproduced by kind permission of the Council of Europe.</p>		



THIS IS NOT A CERTIFICATE

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

LanguageCert is a business name of PeopleCert Qualifications Ltd, UK company number 09620926.

Copyright © 2023 LanguageCert

All rights reserved. No part of this publication may be reproduced or transmitted in any form and by any means (electronic, photocopying, recording or otherwise) except as permitted in writing by LanguageCert. Enquiries for permission to reproduce, transmit or use for any purpose this material should be directed to LanguageCert.

DISCLAIMER

This publication is designed to provide helpful information to the reader. Although care has been taken by LanguageCert in the preparation of this publication, no representation or warranty (express or implied) is given by LanguageCert with respect as to the completeness, accuracy, reliability, suitability or availability of the information contained within it and neither shall LanguageCert be responsible or liable for any loss or damage whatsoever (including but not limited to, special, indirect, consequential) arising or resulting from information, instructions or advice contained within this publication.



Language
Cert

[languagecert.org](https://www.languagecert.org)