A tale of two centres: A cross-institutional peer review of integrated assessment use in direct entry programs

Cara Dinneen
Macquarie University College

John Gardiner
University of Sydney Centre for English Teaching

Mohammed Sameer
University of Sydney Centre for English Teaching

Jeremy Koay
Macquarie University College

Sharon Cullen
University of Sydney Centre for English Teaching

Tony Hickey
University of Sydney Centre for English Teaching

Alejandra Vazquez
Macquarie University College

Jose Lara
Macquarie University College

Mariela Mazzei
University of Sydney Centre for English Teaching

Direct Entry Programs (DEPs) at Australian universities are designed to enhance students' English language and academic skills, with the primary goals of preparing and assuring students' readiness for university studies. Given that university tasks often require the integration of language skills (e.g., writing assignments using information from lectures, articles, and textbooks), this paper proposes the use of integrated assessment tasks in DEPs to increase the authenticity in their assessment approach by using tasks that replicate actual university learning, teaching, and assessment practices. To encourage the shift from standard practice of separate skills testing to integrated assessment, we present key findings of a literature review in integrated assessment use in language programs and show how integrated assessment is implemented in two institutions in Sydney, Australia. The paper also highlights advantages of cross-institutional peer review and proposes an argument-based validity framework for test design that language centres can use to guide good practice.

Key words: integrated assessment; English language proficiency; direct entry; university preparation; peer review; validity

Introduction

A core function of Australian university centres offering English Language Intensive Courses for Overseas Students (ELICOS) is to prepare international students who have English as an additional language (EAL) for university study. Macquarie University College (MQUC) and The University of Sydney's Centre for English Teaching (CET) are two centres with a long tradition of running direct entry programs (DEPs) for ELICOS students. DEPs have a two-part objective: firstly, to develop students' language skills and literacies in preparation for further studies and secondly, and most importantly, to ensure that on successful completion of the program, students' English proficiency meets the minimum university entry requirements. The Tertiary Education Quality and Standards Agency (TEQSA), the national regulator which oversees ELICOS direct entry programs at university-based centres in Australia, states that institutions must have "formal measures in place to ensure that DEP assessment outcomes are comparable to other criteria used for admission to the tertiary education course of study" (Australian Government, 2018, p.8). Formal measures outlined in the TEQSA Guidance Note can be in the form of benchmarking to proficiency frameworks that have been validated, such as the Common European Framework of Reference (CEFR), and the undertaking of cross-institutional peer review with other English language centres (Australian Government, 2019). This project reports on an example of the latter.

In response to the University English Centres of Australia (UECA) integrated assessment grant initiative in 2022, MQUC and CET undertook a collaborative multi-part project which involved an integrative critical literature review on integrated assessment and contemporary higher education, a cross-institutional study of integrated assessment design and the operationalisation of these assessments. This paper highlights key findings of the literature review through a problem-solution analysis which emphasises the relevance and appropriateness of integrated assessment for students preparing for higher education studies. It goes on to consider integrated assessment design and ways in which the ecological context of the two centres impacts design. The paper then reports on a cross-institutional peer review that was undertaken between the two centres during 2022 and 2023 as a measure of quality assurance, and finally proposes a Framework for Argument-Based Validation for Good Practice in Test Design.

Issues

In Australia, universities determine the English language proficiency level of prospective students from non-English speaking countries applying to degree courses

through a range of mechanisms including international standardised tests of English language proficiency. These include the International English Language Testing System (IELTS), the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT), and, more recently, the Pearson Test of English- Academic (PTE-A) (Gardiner & Howlett, 2016). Within Australia, IELTS has been the dominant university entry test for many years (Read, 2015; O'Loughlin, 2015; Cope, 2011), and ELICOS institutions routinely map their course progression against projected IELTS score equivalencies so that receiving institutions can interpret student results in their EAP programs. Consequently, and perhaps inevitably, over time, the IELTS construct has come to permeate testing constructs found in Australian DEPs. This is evident in the use of separate skills testing, multiple-item listening and reading tests, and the use of criteria from IELTS band descriptors in speaking and writing tasks.

The suitability of separate skills testing for university preparation is questionable, however, because it is not highly reflective of the target language use domain (O'Loughlin, 2015). The importance of constructive alignment of curriculum design, assessment practices, and course delivery methods (Biggs, 1999) that reflect the university context has long been emphasised. Yet, in practice, direct entry programs are often required to use IELTS as a default benchmark for reporting the scores of individual macro skills (i.e., listening, speaking, reading, writing) when conveying DEP test results to university administrators (University English Language Centres, 2020). This administrative requirement has commonly led to assessment design that is incongruent with constructive alignment and the target language use domain. As Read (2022) points out, IELTS emphasises the reliability of single skill scores rather than construct validity for higher education (HE) contexts. More specifically to DEPs, the Australian Universities Quality Agency, presently known as Tertiary Education Quality and Standards Agency (TEQSA), produced *Good Practice Principles for DEPs* in 2009. As Principle 6 describes, "development of English language proficiency is integrated with curriculum design, assessment practices, and course delivery through a variety of methods" (Australian Universities Quality Agency 2009, p.4). Implementing this principle in DEPs presents a dilemma if the multi-model aspects of an integrated skills curriculum are not reflected in assessments and the reporting of integrated skills. DEPs that reflect the integrated nature of university tasks would lead to more integrated assessment tasks and related marking rubrics.

Another issue with separate skill-based assessment is that it does not fully correspond with the language skills required for the target language use domain at tertiary level. University students are expected to produce written pieces or deliver presentations by manipulating and critically evaluating source texts such as lectures, lecture notes, and course reading materials. Drawing on Bachman (2009), effective and useful

language assessments should model, as closely as possible, activities that test takers will need to perform in real-life settings. However, this integration of linguistic and academic skills in assessment in DEPs is still uncommon.

One other issue is that although DEPs across Australia share a common objective, to prepare students for further studies through enhancing students' English language and academic skills, their design and methods of assessment vary (Roche & Booth, 2019). Many university language centres operate closely and prescriptively with their own university boards in the way they design course assessments; however, benchmarking of policies, processes, and assessment or transparency across ELICOS is limited (Roche & Booth, 2019). An *English Australia Best Practice Guide in DEPs* is available to members (Brandon & O'Keefe, 2017), which covers content, methodology, general assessment principles, and alignment to destination programs, but the absence of articulated sector standards for DEPs and shared practices for the validation of programs and assessments means there may be vast differences among DEP standards (Murray & O'Loughlin, 2007). This is problematic for students who are seeking an equitable entry to university but may encounter a diverse range of ways to demonstrate standards through these varying assessment practices in DEPs.

Solutions

Reflecting on the current separate skill-based assessment and the lack of consistency among ELICOS institutions, this article proposes two solutions. The first is the use of integrated assessments in DEPs to better meet students' future needs, and the second is a tool that provides language centres with scaffolded support for good practice in integrated assessment design.

Defining integrated assessment

Integrated assessment refers to a form of assessment that requires test takers to use more than one macro skill (e.g., both reading and writing). In the context of DEPs, integrated assessment tasks reflect the types of tasks that university students will engage with in their future academic programs (e.g., Bachelors, Masters, PhDs). This section identifies four features of integrated assessment.

A key feature of integrated assessment tasks is that they replicate target-domain language use. In other words, they "simulate authentic language-use situation" (Plakans, 2020, p.1) with the purpose of assessing students' "capability to use language in a specified space of contexts" (Mislevy & Yin, 2009, as cited in Cumming, 2014, p.3). As such, designing integrated assessment in the DEP context involves identifying tasks that university students are likely to perform. In identifying these tasks, Leki and Carson (1997) found that 40% of writing task types reported by university disciplines encouraged students to write about something they had read.

They also found that at least 52% of the writing tasks EAL students did when studying English was personal in nature and did not truly represent university writing tasks. In the DEP context, this mismatch can be resolved by developing integrated tasks that reflect university tasks.

A second feature of integrated assessment tasks is that they require students to use two or more skills when responding to tasks. An integrated assessment task could, for example, combine reading and writing skills (Delaney, 2008), or writing and listening skills (Plakans, 2012). Combinations involving the integration of more than two skills are also possible. For example, test takers are asked to incorporate information from an audio recorded lecture and an academic reading text in their writing. As Plakans et al. (2019) observe, reading and writing in a university context involve selecting and connecting ideas, and organising texts. Using more than one macro skill to produce a written assignment or oral presentation would reflect the type of assessments that university students are likely to encounter. It is thus essential to design and use tasks in the DEP context that combine two or more skills to simulate university tasks.

A third feature of integrated assessment is the selection of input sources. Texts selected for integrated assessment, according to Knoch and Sitajalabhorn (2013), should be text rich. In other words, input sources should contain paragraphs of text. To reflect an authentic university reading experience, information in the input source may not necessarily have to be completely related to a given task. Such texts would require students to select and include information that is relevant for completing the task. As university students are increasingly expected to engage with multimodal materials, Yang (2012) suggests that visual information (e.g., graphs) may also be included as a form of input source.

The final feature of integrated assessment tasks relates to how the source or input texts are used. While input texts can be used to stimulate thinking in integrated assessment (Plakans, 2020), test takers need to demonstrate comprehension of source materials and an ability to use information from them appropriately in their responses (Plakans, 2015). For Knoch and Sitajalabhorn (2013), meaningful integration of skills in an integrated writing task would involve identifying and selecting relevant ideas from given texts, synthesising or finding connections between ideas from the texts, transforming language from input texts, logically organising ideas to complete assessment tasks, and acknowledging sources. In the DEP context, integrated tasks should require students to use source materials to generate ideas, support their claims, and cite them.

Considering validity

Explicit, evidence-based quality assurance practices must be in place at language centres that offer DEPs (TEQSA, 2019). The validation of test instruments and their uses is an ongoing process of review and leads to the generation of accurate and meaningful test results for stakeholder use (Bachman & Palmer, 1996). Seminal work in validity studies includes the interpretive validity framework (Kane, 1992), and Bachman and Palmer's assessment use argument (2010). The argument-based approach to validation (Kane, 2013) requires the integration of multiple types of evidence to support the interpretation of test results. For example, if a DEP test result of 50% is interpreted as a passing grade for a student who must demonstrate the English proficiency of a Common European Framework of Reference (CEFR) B2 independent user, there must be multiple layers of evidence to support an argument that this interpretation of results is valid.

The Association of Language Testers of Europe (ALTE) (2020) outlines eight types of evidence, called 'aspects' of an argument-based validation of good practice. These are test impact, content validity, construct validity, reliability, criterion-related evidence, fairness, quality of service, and practicality. The scope of this paper does not allow for an in-depth development of a validity argument for the integrated assessment tests presented later in the study. Instead, an overview of six of ALTE's validity aspects for good practice, as they relate to the context of DEPs, is provided and several criticisms of integrated assessment use are highlighted.

Consideration of test impact questions how stakeholders are impacted by using the test. An example of impact is washback — the way that test preparation influences learning and teaching practices in the classroom. Assessment tasks that reflect skills needed for university studies have a positive washback because they lead to the development of highly relevant transferable skills relevant to the DEP students' future studies (O'Loughlin, 2015).

Construct validity focuses on supporting evidence for score interpretation and use of scores (Bachman & Palmer, 2010). In establishing construct validity, developers must demonstrate that the exam format and level of difficulty are appropriate for measuring the components of language competency intended. This is achieved by defining traits of ability that the test measures (ALTE, 2020) and the CEFR performance descriptors are a useful resource for this purpose. One construct validity challenge for integrated assessment tasks is the lack of a fixed genre, which can make it difficult for test designers to develop suitable integrated task rubrics. However, a precedent for developing and validating descriptors in a comparable way to CEFR has been established for a category or genre that currently lacks descriptors (North & Docherty, 2016). This validation process comprises three phases, namely organising descriptors

into categories, such as thematic development and mediation of information, aligning descriptors to CEFR levels, and judging the task performance as described by a descriptor. Despite the absence of a defined genre, the use of academic tasks with reference to source materials is an important part of DEP students' development. As Cumming (2014) points out, if such tasks are not included in DEP assessments, construct under-representation would be the result.

Content-related validation is concerned with the relevance of the test materials and their coverage of appropriate language and competencies (ALTE, 2020). In determining content validity, DEP test developers must argue that their exams are meaningful and the use and interpretation of test scores allow examiners to make good decisions about students' readiness for university study. Although this paper has highlighted the advantages and relevance of integrated assessments in DEP contexts, existing university admission requirements for the reporting of separate skills results can impose rubric and test design constraints that cause substantial variance among university language centres. This is exemplified later in the paper by the distinction in reporting requirements for MQUC and CET which impacts the ways that test scores are derived from similar integrated tasks.

For reliability, developers must evidence claims that test results are precise, stable, consistent, and as free from errors of measurement as possible. DEP practitioners achieve this through rater training, standardisation, moderating results across testing events, and collecting and responding to student and teacher feedback; all of which must be clearly documented in assessment policy and procedures. Additionally, language centres with the requisite expertise and software systematically undertake multi-faceted Rasch analysis to check the reliability of tests, test takers, and rater performance.

Another relevant aspect of validity addressed by this study is criterion-related evidence. DEP assessment developers must evidence ways that test scores correlate with recognised external criterion (Australian Government, 2019), such as the CEFR. While the advent of the extended CEFR Companion Volume's (CEFR-CV) descriptor scales (CoE, 2018) has provided language centres with a valuable referencing resource, criterion validation does remain a significant challenge because the reliability of criteria-based rating rubrics has not received the same amount of attention in integrated assessment studies as other aspects of integrated tasks (Uludag & McDonough, 2022). One of the challenges of developing a rubric for integrated assessment is addressing the accurate integration of source materials and use of language (Plakans & Ohta, 2021). The CEFR-CV descriptors for mediating a text in written and spoken form (CoE, 2018) are a useful resource; however, the authors note the difficulty experienced by markers when rating source integration across proficiency levels.

Most rating criteria for integrated assessment tasks comprise elements of source use, content, idea development, organisation, and language use (Plakans & Ohta, 2021). The use of holistic rubrics that rate the overall task performance in summative tasks has been recommended for practical reasons (Gebril, 2018). Other researchers, however, have found analytic multi-trait scores given by raters to be more consistent and accurate, especially for source usage (Ohta et al., 2018, Shabani & Panahi, 2020). Whether an analytic or holistic approach to rubric design is adopted, the two types of rubrics have merit and limitations, so the choice will depend largely on context and purpose (Shabani & Panahi, 2020).

The practicality of test implementation refers to the use of tests that are appropriately resourced for development and deployment. During the online delivery of DEP assessment in 2020 and 2021 due to the COVID-19 pandemic, for example, both MQUC and CET identified the unsustainability of using separate receptive skills tests containing short answer questions, which had been standard practice. The reduced test security in online delivery was a risk to academic integrity because students could quickly and efficiently share test answers via texting. Additionally, there was little control over test security, which meant that once they had been used, tests needed to be retired for up to two years before being used again. This situation accelerated the need for centres to produce alternative test versions and item banks requiring an unsustainable amount of developer resources. In short, separate skills tests were no longer practicable.

Cross-institutional peer review

In designing the review, MQUC and CET drew on the work of the External Referencing of ELICOS Standards (ERES) Project (Roche & Booth, 2021), which set out to establish cross-institutional comparability of DEP assessment policies, processes, and learning outcomes. The peer review questions below illustrate the simplified framework of four questions that the teams decided on as a means of quality assurance and establishing a validity claim for good practice. Each team acted as an external panel of experts for the other by reviewing test construct, content and criteria, and rating students' performances against the centre's rubrics to provide consensus, disagreement, and feedback.

Peer review questions

For each of the following, please explain your rating. Please list specific suggestions for improvement where appropriate.

1. Does the assessment item map against the stated learning objectives/ outcomes?
2. Is the description of the performance standards (e.g., assessment rubric, annotated work samples) appropriate to the specified learning objectives/ outcomes?
3. Do you agree that the grades awarded reflect the level of student attainment?
4. Are there any other matters you wish to provide feedback on?

The teams provided each other with written and verbal feedback on the questions above.

The design of MQUC and CET integrated assessments

Both MQUC and CET work with their own university boards and administrative directors to ensure the needs of university stakeholders are being met by their programs. These ecological contexts are highly influential in shaping decisions about assessment design. One notable difference between the two centres, for example, is the results reporting requirements. At CET, results are reported according to individual skills of speaking, writing, reading, and listening in line with University of Sydney Admissions requirements, while at MQUC, an overall score is acceptable for the completion of a suite of integrated assessment tasks.

In the design of their integrated assessment tasks, MQUC and CET have met Knoch and Sitajalabhorn's (2013) recommendations; the input texts are language rich, and the test takers are required to synthesise ideas and information from the texts to respond to specific tasks in written or spoken form. Both centres have chosen the CEFR's model of language use to define the linguistic, sociolinguistic, and pragmatic competences to be assessed. In the creation of tasks, learning outcomes, and rubrics, descriptor scales from the expanded CEFR-CV (CoE, 2018) have been adapted to the context of DEPs, which seek to represent language activities from learning and teaching events in higher education.

The development of rating rubrics for the integrated tasks at both centres were part of an iterative process. This process involved adjustments according to teacher and student feedback (O'Grady & Taşkesen, 2022) and features alignment with descriptors from the CEFR-CV's B2 and C1 modes of communication: reception, production, interaction, and mediation. Familiarisation of each rubric scale at the two centres is regularly conducted at the start of the study period through marking group moderation standardisation sessions and teacher marking monitoring (Harsch & Martin, 2012). At standardisation sessions, samples are rated, firstly individually

and then in groups, followed by feedback from the members of the assessment team. This social construct of shared understanding of the rubric and alignment of scores to the accepted standard is likely to result in greater marking reliability (Bloxham, 2009) and reflect the local institutional values (Dimova et al., 2020).

The types of integrated task rubrics used at the two centres have their relative merits. Since the basis for rubric development at both centres was the CEFR- CV (see Rubric Criteria Comparison Appendix A), it is worth noting the tendency of CEFR to use a mix of sentences describing the language user and analytic 'can-do' points. A type of hybrid rubric comprising both holistic and multi-trait analytic criteria for integrated assessment tasks has been validated to score integrated tasks (Yamanishi et al., 2019). At both centres, the integrated task rubrics (see Appendices C, D, and E for rubric excerpts) combine an overarching statement (holistic) with analytic bullet points (analytic). Through mediating these two types of rubric criteria and from the moderation sessions, trained teachers calibrate the score. Although more research is needed in this field (Harsch & Martin, 2013), rater variability may be larger in holistic than multi-trait scores (Ohta et al., 2018). This issue can be ameliorated by extensive rater training with a wide range of exemplars and samples in moderation sessions.

MQUC uses five integrated tasks in its suite of direct entry assessments, two of which were used in the peer review.

i) *The integrated reading and listening to writing* task is a timed examination with two input sources: a 450-word academic article for a general audience at a C1 level of difficulty, and 700-word audio text on the same topic at a B2+ level of difficulty. Both are discursive texts on a controversial topic with opinions, explanations, justifications, and exemplifications included. Students demonstrate their comprehension of the texts by drawing on relevant information to produce a 450-to-500-word discussion essay in response to an essay prompt.

ii) *The integrated listening to speaking* task input material is an 800-word dialogue between two people about problems and solutions on a particular aspect of university study. Each dialogue contains a main problem, plus details about two specific instances of the problem. A solution is provided to address each of the two specific instances and final advice is given to solve the overall problem. Students listen to the dialogue and make notes about key ideas and supporting details. This is followed by an interactive 8 to 10-minute pair discussion about the problem, instances of the problem, personal responses or experiences with such a problem, and suggested solutions, with individual follow-up questions by the examiner at the end.

CET uses two integrated assessment tasks in its DEP course. Both were used in the comparative study.

i) *The critical response task* is an integrated listening and reading to writing task with four input sources: one recorded seven-minute talk at the B2 level and three C2 reading texts around 2,000 words in total. The topics of the texts should be related, even tangentially, to some aspect of the themes covered in Weeks 6-8 of the course. Text One is usually an important and longer text that identifies key ideas and provides definitions of any recurring key words. Texts Two and Three should further explain the key ideas but also have a different or overlapping perspective. A simpler or less academic text from a reputable source may be useful for converting into a spoken style transcript of around 550-650 words for the listening input. All input sources are given to students two days before the assessment. In the critical response task, students draw on relevant information from the input sources to write a 500-700-word response. Students are required to synthesise information from different texts based on a question. The question should ask students to evaluate ideas and express a position. This question helps identify the filter that readers use to extract relevant ideas from the texts. These ideas can then be used as evidence to support their argument in response to the question. Students receive a writing score only.

ii) *The interactive speaking task* is an integrated reading to speaking task with three C2 input reading texts provided to students two days in advance. These texts, from reputable sources and with a combined total of 3,000 words, must be related to the broad topics covered in Weeks 6-9 of the DEP. In the assessment, students are shown a question and have 10 minutes preparation time before a 15-minute group discussion (3-4 students). Following the discussion, students are given five minutes to study a small section of one of the given reading texts (around 200 words). Each student is then asked three follow-up questions based on this section and receives a speaking score awarded by two teachers.

Peer Review Outcomes

The MQUC and CET panel of experts in this project comprised five experienced EAL teachers in each team. Preliminary project planning was conducted via Zoom meetings, followed by in-person meetings (see agreed parameters for the study in Appendix B). An initial face-to-face consensus moderation meeting took place at MQUC campus and due to the richness of discussion, a second meeting was held at CET two weeks later. These moderation meetings ensured a shared understanding of the rubric and marking interpretation between the project group members that were familiar with the test materials and those from the other institution (Bloxham,

2009). This moderation through discussion and clarification thus enhanced rater confidence and reliability (O'Connell et al., 2016; Bloxham, 2009). Calibrating rater judgements through moderation has also been implemented with CEFR external benchmarking of writing in various educational contexts (Harsch & Martin, 2012; Deygers & Van Gorp, 2015), and such moderation proved crucial in using the rubric adequately in our project.

In response to the first of the peer review questions, it was agreed that for both centres the assessment tasks and rubrics had been successfully mapped against the stated learning objectives/outcomes and importantly, that the appropriate assessment standards as referenced against the CEFR, were evident. The peer review confirmed that broadly the assessments in place were a valid measure of students' proficiency for admission to the tertiary education course of study, as required by the regulator (Australian Universities Quality Agency, 2009). These findings can also be presented to the relevant university governance boards as a formal measure ensuring that the centre's DEP assessment outcomes are comparable to relevant criteria (here the CEFR). Further detail, including suggestions for improvement to individual tasks, are discussed in the following sections.

MQUC Outcomes

- *Integrated reading, listening, and writing test*

CET raters questioned MQUC's use of one reading text and one listening text for this task and suggested that the use of two reading texts of different genres would provide a better assessment of students' reading capabilities. In response, MQUC have amended the test specifications so that the reading input consists of two shorter texts with different genres and register. They will trial it with an upcoming cohort. The students currently have 90 minutes' writing time for the synthesis of two texts, and the MQUC team may need to consider a longer writing period according to task complexity.

CET raters also noted ambiguity in the rating criterion for task response, which caused difficulty in making an assessment. As a result, MQC have amended the rubric to indicate more clearly the number of ideas that could be addressed. Additionally, CET recommended modifications to the mediation of language criterion regarding paraphrasing techniques and language accuracy which have now been implemented.

- *Integrated listening to speaking test*

CET raters commented that the input dialogue was very structured (see problem-solution text attributes described above) and may be predictable to students. The MQUC team did not feel this to be an issue since all genres have predictable patterns

and the focus of the task is on students relaying details of the dialogue, which could not be predicted, in a follow-on discussion. CET also noted that follow-up questions of variable levels of difficulty needed to be better distributed among the candidates, and this has now been implemented. There were also minor modifications made to the language and fluency components of the rating rubric to make vocabulary, grammar, and fluency more explicit.

Final commentary by the CET team was about the lower levels of complexity, or difficulty of the input texts in the MQUC tests compared to the CET tests. The expert panel questioned the authenticity of the tasks if the texts were not representative of the level of complexity students are likely to encounter in undergraduate and postgraduate studies, and asked whether there was an assumption that the skills demonstrated in this task were transferrable to other academic tasks. MQUC's response to this is that yes, the skills, knowledge, and abilities demonstrated by students when undertaking these integrated tasks can be taken and applied to learning and teaching situations at university. These skills and competencies were drawn from the CEFR's model of language use and mapped to its reference levels and corresponding language user proficiency profiles (CoE, 2018). An important distinction about the complexity of the texts is that time is a key constraint with the test construct. Unlike university study and the CET integrated tasks, where students have time on their own to read or translate texts as required, these integrated skills tests are timed, invigilated examinations and students have no access to digital tools to support comprehension nor assist them in summarising and paraphrasing information ahead of time.

CET Outcomes

- Critical response task

MQUC raters raised questions about the specifications and task. They highlighted a mismatch in the number of learning outcomes in the task sheet and the specifications. In response, CET agreed to align the task sheet to the specifications. MQUC raters mentioned that more personalisation such as "The critical response essay tasks requires you to …" could improve the task instructions for students. CET agreed with this suggestion. A difference in approach to essay structures between the two centres became apparent when MQUC questioned the varied essay structures present in the CET student samples. While the MQUC DEP develops a specific discursive essay structure, CET uses numerous samples instead of exemplars to demonstrate the variations in how students can write introductions and conclusions. However, one clear requirement in the introduction was the inclusion of a clear position.

In terms of the input material, MQUC raters observed that the listening input material

was too challenging in comparison to the reading texts. CET agreed to adopt their suggestion of adding more signposts to the scripts to make the recordings more natural.

- Interactive speaking task

MQUC raters suggested that aspects of the task be clarified, indicating that specifications may need further elaboration. They expressed concern in relation to the complexity of the task and the workload practicality for the assessors. Although the CET team accepts that it was a challenging task, the synthesis of input sources that closely simulates the "authentic language-use situation" (Plakans, 2020, p. 1) in the university target-domain was highly valued, and both teachers and students have access to the input texts more than 48 hours before the assessment. In addition, they noted that using two teacher assessors reduces the load on evaluators. Even though the MQUC raters strongly believe that this task should contribute to a reading score, adopting this suggestion would require a change in the reporting structure to the University of Sydney. As in the critical response task, CET agreed to the personalisation of the instructions for students. MQUC raters questioned whether the complexity of the task lends itself to students reading rather than producing more natural speech. CET agreed to simplify the task and the follow-up questions to improve natural speech. MQUC also questioned how to evaluate students who lift sections from input materials and prolonged reading from notes. This can be problematic at times, but chunks that are obviously read from texts are not rewarded, and follow-up questions are used as a backup if the content is minimal in the discussion.

Considerations for other centres

This project has highlighted a range of considerations in integrated skills testing and in conducting cross-institutional peer reviews. The first consideration is that there are two different but reasonable approaches to integrated assessments which other centres could consider in terms of score reporting for university admissions and providing student access to input for tests (Shabani & Panahi, 2020). University admissions reporting requirements can comprise: (1) integrated tasks which provide a single integrated score or (2) integrated tasks which provide a score for either writing or speaking. In either case, it is important that DEP providers ensure university admissions staff are informed about appropriate test score interpretations. Among admissions staff there may be an expectation for language skills to be reported separately in the way that international proficiency providers, such as IELTS, report scores. Such legacy administrative practices, however, should not drive pedagogical decisions in contemporary DEP design. To retain currency and relevance in a fast-changing world, assessment design needs to be closely linked to language use in the target language use domain (Bachman & Palmer, 1996), which in the case of DEPs

is higher education learning and teaching environments. Additionally, as Ajjawi et al. (2023) affirm, assessment design should hold psychological authenticity for the learner, providing opportunity for them to judge the relevance and value of a task within the broader social realm of their future place of work or study.

Another consideration is related to input texts (Plakans, 2015, 2020; Knoch & Sitajalabhorn, 2013; Yang, 2012). Student access to input readings and audio texts can be given during or prior to the assessment, considering such factors as the length and complexity of the texts, security, authenticity, reduction of cognitive load, and student fatigue due to time constraints, as well as practical aspects of invigilation and resource management.

A third consideration is regarding quality assurance and good practice in test design (ALTE, 2020; Bachman & Palmer, 2010; Kane, 2013). The Argument-Based Validation for Good Practice Framework (Appendix F) is drawn from ALTE's Principles of Good Practice (2020) and defines important aspects of validity, the questions, and the evidence that language centres can use to validate good practice in test design.

The project led to three recommendations for centres undertaking a similar benchmarking study:

i) expand the range of questions in the four peer review questions noted above to cover validity arguments pertaining to fairness, practicality, and impact of test design (see Appendix F);

ii) conduct cross-institutional standardisation sessions prior to beginning the marking so that teams are familiar with the workings of rubrics; and

iii) clarify feedback processes to consider whether the centre will provide one feedback report, or multiple reports from each team member.

Finally, the project resulted in three recommendations for designing integrated tasks:

i) align specifications and instructions for students and teachers because it will provide a clear task objective and appropriate washback;

ii) design the rubric and the criteria carefully to encapsulate the task learning objectives. Centres should also consider the relative merits of multi-trait analytic rubrics, holistic rubrics, or hybrid style rubrics depending on the purpose of the task; and

iii) consider the academic task authenticity against the task difficulty level that is manageable for their student cohort.

Benefits to centres

Both centres have found the overarching integrated assessment project to be enormously beneficial for the assurance of quality in assessment design, implementation, and validation. Rather than protecting assessment innovations as business-sensitive artefacts, and perpetuating ELICOS centre-silos, the teams found that the sharing and peer review of assessment artefacts, practices, and processes has fuelled innovation and driven quality improvement, while ensuring appropriate standards are in place in integrated assessment design and delivery.

REFERENCES

Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2023). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education*, 1-12. https://doi.org/10.1080/02602938.2023.2271193

Association of Language Testers in Europe. (2020). *Principles of good practice.* https://pt.alte.org/resources/Documents/ALTE%20Principles%20of%20Good%20Practice%20Online%20version%20Proof%204.pdf.

Australian Government. (2018). *ELICOS Standards 2018.* Retrieved from https://www.legislation.gov.au/F2017L01349/asmade/text

Australian Government. (2019). *Guidance note: ELICOS direct entry.* Retrieved from https://www.teqsa.gov.au/guides-resources/resources/guidance-notes/guidance-note-elicos-direct-entry

Australian Universities Quality Agency. (2009). *Good practice principles for English language proficiency for international students in Australian universities*. TD/TNC 108.846. Retrieved from https://www.voced.edu.au/content/ngv:51168#

Bachman, L. F. (2009). Generalizability and research use arguments. *In Generalizing from Educational Research* (pp. 137-158). Routledge.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in the real world: Developing language tests and justifying their use.* Oxford University Press.

Bachman, L., & Adrian, P. (2022). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Biggs, J. (1999). *Teaching for quality learning at university.* Open University Press.

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education, 34*(2), 209-220. https://doi.org/10.1080/02602930801955978

Cope, N. (2011). Evaluating locally-developed language testing: A predictive study of 'direct entry' language programs at an Australian university. *Australian Review of Applied Linguistics, 34*(1), 40-59. https://doi.org/10.1075/aral.34.1.03cop

Council of Europe (CoE) (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment companion volume with new descriptors.* Council of Europe. Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Cumming, A. (2014). Assessing Integrated Skills. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (Vol. I). John Wiley & Sons. https://doi.org/DOI:10.1002/9781118411360.wbcla131

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5-43. https://doi.org/doi.org/10.1016/j.asw.2005.02.001

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7,* 140-150. https://doi.org/doi:10.1016/j.jeap.2008.04.001

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing, 32*(4), 521-541. https://doi.org/10.1177/0265532215575626

Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, implementation, and development.* Routledge.

English Australia (2017). *Guide to best practice in 'direct entry' programs*. https://www.englishaustralia.com.au/documents/item/197#

Gardiner, J., & Howlett, S. (2016). Student perceptions of four university gateway tests. *University of Sydney Papers in TESOL, 11,* 67-96.

Gebril, A. (2018). Intergrated-Skills assessment. *The TESOL Encyclopedia of English Language Teaching.* https://doi.org/DOI:10.1002/9781118784235.eelt0544

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*(4), 228-250. https://doi.org/10.1016/j.asw.2012.06.003

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281-307.

Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes. *Assessing Writing, 18,* 300-308.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527.

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

Leki, I., & Carson, J. (1997). "Completely different worlds": EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly, 31,* 39-69. https://doi.org/doi.org/10.2307/3587974

Levinson, S. C. (1995). Three levels of meaning. In *Grammar and meaning: Essays in honour of Sir John Lyons* (pp. 90-115). Cambridge University Press.

Murray, D., & O'Loughlin, K. (2007). *Pathways - Preparation and Selection.* Paper presented at the National Symposium: English Language Competence of International Students, Sydney. https://internationaleducation.gov.au/research/Publications/Documents/NS_PathwaysPreperationSelection.pdf

North, B., & Docherty, C. (2016). Validating a set of CEFR illustrative descriptors for mediation. *Cambridge English: Research notes,* (63), 24-33.

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., & Watty, K. (2016). Does calibration reduce variability in the assessment of accounting learning outcomes? *Assessment & Evaluation in Higher Education, 41*(3), 331-349. https://doi.org/10.1080/02602938.2015.1008398

O'Grady, S., & Taşkesen, Ö. (2022). Developing a rating scale for integrated assessment of reading-into-writing skills. *Language Learning in Higher Education, 12*(1), 159-183. https://doi.org/10.1016/j.asw.2022.100609

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing, 38,* 21-36.

O'Loughlin, K. (2015). 'But isn't IELTS the most trustworthy?': English language assessment for entry into higher education. In *International education and cultural-linguistic experiences of international students in Australia* (pp. 181-194). Australian Academic Press.

Plakans, L. (2012). Writing integrated assessment. In G. Fulcher & F. Davidson (Eds.), *Handbook of Language Testing.* Routledge.

Plakans, L. (2015). Integrated second language writing assessment: Why? What? How? *Language and Linguistics Compass, 9*(4), 159-167. https://doi.org/10.1111/lnc3.12124

Plakans, L. (2020). Assessment of integrated skills. *The Encyclopedia of Applied Linguistics.* https://doi.org/10.1002/9781405198431.wbeal0046.pub2

Plakans, L., Liao, J., & Wang, F. (2019). "I should summarize this whole paragraph": Shared processes of reading and writing in iterative integrated assessment tasks. *Assessing Writing, 40,* 14-26.

Plakans, L., & Ohta, R. (2021). Source-based argumentative writing assessment. In A. Hirvela, & D. Belcher (Eds.), *Argumentative writing in a second language: Perspectives on Research and Pedagogy,* 64-81. Ann Arbor: University of Michigan Press.

Read, J. (2015). *Assessing English Proficiency for University Study*. Palgrave Macmillan UK.

Read, J. (2022). Test Review: The International English Language Testing System (IELTS). *Language Testing, 39*(4), 679-694.

Roche, T., & Booth, S. (2019). *External Referencing of ELICOS Direct Entry Program Standards: UECA National Report 2019.* Retrieved from https://ueca.edu.au/initiatives/

Roche, T., & Booth, S. (2021). A Collaborative Approach to Assuring Standards: Using the CEFR to Benchmark University Pathway Programs' English Language Outcomes. *Language Teaching Research Quarterly, 26,* 18-38.

Shabani, E. A., & Panahi, J. (2020). Examining consistency among different rubrics for assessing writing. *Language Testing in Asia, 10*(1), 1-25.

Tertiary Education Quality and Standards Agency. (2019, 5 June). *Guidance note: ELICOS direct entry.* https://www.teqsa.gov.au/guides-resources/resources/guidance-notes/guidance-note-elicos-direct-entry

Uludag, P., & McDonough, K. (2022). Validating a rubric for assessing integrated writing in an EAP context. *Assessing Writing, 52,* 100609.

University English Language Centres (2020). *UECA Benchmarking 2020: Phase 1 Direct Entry ELICOS Assessment Practices Summary Report*. Retrieved from https://ueca.edu.au/benchmarking-2020-phase-1/

Yamanishi, H., Ono, M., & Hijikata, Y. (2019). Developing a scoring rubric for L2 summary writing: a hybrid approach combining analytic and holistic assessment. *Language Testing in Asia, 9*(13), 1-22. https://doi.org/10.1186/s40468-019-0087-6

Yang, H. (2012). Modeling the relationships between test-taking strategies and test performance on a graph-writing task: Implications for EAP. *English for Specific Purposes, 31,* 174-187. https://doi.org/10.1016/j.esp.2011.12.004

DOI: https://doi.org/10.61504/OKTA3655

Cara Dinneen is the Education Manager for English Language Programs at Macquarie University College and Convenor for the English Australia Assessment SIG. She holds a Master of TESOL, and Graduate Certificates in Educational Research and Business Educational Leadership, and has 22 years' experience in teaching and leadership in Australia, Oman, and Spain.

cara.dinneen@mq.edu.au

Orchid ID. 0000-0002-2497-565X

John Gardiner is a teacher on direct entry programs at the University of Sydney Centre for English Teaching (CET). He holds a Dip. Teach, Grad. Cert. TESOL, and Master of Arts (TESOL) and is a member of the Assessment Quality Team at CET. His interests include language testing, action research, and course development.

john.gardiner@sydney.edu.au

Mohammed Sameer writes assessment tasks, runs standardisation sessions for teachers, and teaches various courses at University of Sydney's Centre for English Teaching + Learning Hub. He holds a PhD in Education and an MA in Linguistics. His interests include language testing, curriculum development, and sociolinguistics.

mohammed.sameer@sydney.edu.au

Jeremy Koay is a pracademic who teaches at the Australian Catholic University, Hanoi University, Macquarie University College, and the University of Canberra. He has taught at tertiary level and in professional contexts in Australia, Laos, Malaysia, Myanmar, New Zealand, and Vietnam. His research interests include Discourse Analysis and TESOL.

jeremy.koay@mq.edu.au

Sharon Cullen is an English language teacher with many years' experience in direct entry programs and assessment development as a member of the Assessment Quality Team at the Centre for English Teaching + Learning Hub at the University of Sydney. She is particularly interested in test evaluation and validity.

sharon.cullen@sydney.edu.au

Tony Hickey is an Education Manager of University Pathways at the University of Sydney Centre for English Teaching and coordinates the Centre's Assessment Quality Team. He has 35 years of experience in various education roles and holds a BA, Dip Ed, CELTA, DELTA, and IDLTM.

tony.hickey@sydney.edu.au

Alejandra Vazquez is a Senior Teacher at Macquarie University College, English Language Programs. She has 30 years' experience as an ELT teacher both in Australia and Argentina. She holds a Master of Applied Linguistics (TESOL). She is passionate about language testing, curriculum development, and leading teaching teams.

alejandra.vazquez@mq.edu.au

Jose Lara is a Senior Teacher for English Language Programs at Macquarie University College. He holds a Master of Applied Linguistics (TESOL) and has 25 years of experience in Venezuela, Saudi Arabia, Bangladesh, and Australia. His interests include educational technologies, learning designing, active reading, and student engagement.

jose.lara@mq.edu.au

Mariela Mazzei manages the Direct Entry Course at CET ensuring that students successfully transition to life at university. She is responsible for supporting teachers through professional development activities and maximising teacher and student engagement in the program. She manages curriculum and assessment projects, monitors student welfare and academic progress, and oversees intervention programs at the centre.

mariela.mazzei@sydney.edu.au

Appendix A

Rubric Criteria Comparison

The following table maps the two language providers' (MQUC and CET) integrated assessment task marking criteria to the CEFR's illustrative descriptor scales.

Centre	Integrated Test Type	CEFR-CV	Marking Criteria
MQUC	Integrated Listening and Speaking Test	Oral comprehension	Communicative effectiveness
		Oral production	Mediation language
		Interaction	Delivery (pronunciation)
	Integrated Reading, Listening, and Writing test	Oral comprehension	Task response (comprehension and use of source material)
		Reading comprehension	Structure
		Written production (reports and essays)	Academic style
		Mediating a text	Mediation language (synthesis)
			Citation and attribution
CET	Interactive Speaking Task (reading input/ speaking output)	Reading comprehension	Content/relevance (including critical thinking)
		Overall oral interaction	Discourse management (interaction, fluency & coherence)
		Sustained monologue: putting a case	Pronunciation
		Mediating a text: spoken production	Vocabulary & grammar (choice, range & accuracy)
		Mediating a text: spoken interaction	
	Critical Response Task (reading & listening input, written output)	Reports and essays	Content/relevance
		Mediating a text	Use of sources
		Coherence and cohesion	Connection of ideas
		Vocabulary control/ range	Academic vocabulary
		Grammatical accuracy	Grammar

Appendix B

Agreed parameters for the peer review.

Program	Samples selected from students who completed a Direct Entry program with the requirements of CEFR B2+ or IELTS 6.5 (no sub-score below 6.0) or higher.
Integrated Skills Tests	Reading and listening into writing (MQUC-ELP Integrated Skills Test, CET Critical Response) Reading or listening into speaking (MQUC-ELP Integrated Listening Speaking Test, CET Interactive Speaking Task)
Assessment inputs	1. Course outline (how the task fits in the program) 2. Test specifications 3. Assessment task instructions 4. Rating rubrics
Assessment outputs	Samples of deidentified student work at different levels of performance (without scores displayed)
Raters	MQUC-ELP: Alejandra Vazquez, Jose Lara, Cara Dinneen, Jeremy Koay CET: John Gardiner, Mohammed Sameer, Sharon Cullen, Tony Hickey
Communication channel	Documents shared via Teams.
Timeline	Process
By Friday 24 February	Teams select and prepare assessment inputs and outputs for the study and make available on the shared Teams site.
By Friday April 28	Teams review assessment inputs and complete the Peer Feedback Questionnaire (Appendix A)
Early May 2023	A consensus moderation meeting is held to review feedback questionnaires, discuss allocated marks, and reach agreement on overall academic achievement standards.

Appendix C

Excerpts from CET's Marking Rubrics

Centre for English Teaching Integrated Speaking Task (reading input, speaking output)			
Criteria	**Score Band**	**Overarching Statement**	**Specific Descriptors**
Content & Relevance	60 - 64	Ideas are generally relevant, with some understanding and development of ideas based on the text.	Generally relevant information from the text is chosen in relation to the question/ discussion. Some lapses in accuracy and clarity of summarised /paraphrased information from text, may include instances of direct reading/lifting from text. Some development of ideas in texts through explanation, elaboration, and linking to other sources, personal knowledge, and experience.
	65 - 69	Ideas are relevant, with only occasional lapses in accuracy and clarity of paraphrases and development of ideas based on the text.	Relevant evidence is chosen from text in relation to question/discussion Occasional lapses in clarity of summarised/ paraphrased information from text (but accurately reported). Some development of ideas in texts through explanation, elaboration, and linking to other sources, personal knowledge, and experience.

Centre for English Teaching Critical Response Task (reading & listening input, written output)			
Criteria	**Score Band**	**Overarching Statement**	**Specific Descriptors**
Content & Relevance	60 - 64	Ideas are mostly relevant, but may be unclear, and demonstrate an adequate critical response to the question.	A position is expressed in response to the question, but it requires more clarity and/or strength. Ideas are sometimes relevant to the question, but key ideas may be omitted or unclear. Evidence from the input sources is sometimes clear but not always selected and synthesised sufficiently to support ideas.
	65 - 69	Ideas are relevant and demonstrate an adequate critical response to the question.	A position in response to the question is mostly expressed clearly. Ideas are often relevant to the question, but some could be more developed. Evidence from the input sources is often clear and selected and synthesised adequately to support ideas.

Appendix D

MQUC Integrated Skills Test Learning Outcomes and Marking Rubric Excerpt

Reading Comprehension Learning Outcomes	Listening Comprehension Learning Outcomes	Writing Learning Outcomes
R1: Can understand articles and reports with contemporary problems in which writers adopt stances or viewpoints	L1: Can understand a recorded, structured mini-lecture on an academic topic for a general audience	W1: Can accurately synthesise information and arguments from several sources
R2: Can distinguish key ideas, opinions, and supporting examples	L2: Can distinguish key ideas, opinions, and supporting examples	W2: Can write a discussion essay expressing multiple points of view
	L3: Can understand the speaker's point of view	W3: Can use a range of language features with precision
		W4: Can cite and attribute sources appropriately

Sample criterion from the rating rubric

MQUC Integrated reading, listening to writing test	Fail	Pass	Distinction	High Distinction
Task response (35%) Demonstrates good understanding of the lecture and reading passage by using key ideas, examples, and explanations from the lecture and reading passage to discuss benefits and limitations.	Includes insufficient key ideas from the lecture and/or reading passage.	Addresses all parts of the question including at least ONE key idea from the lecture AND reading passage.	Addresses all parts of the question including at least TWO key ideas from the lecture AND reading passage.	Fully addresses all parts of the question skilfully using the key ideas from the lecture and reading passage.
	Misrepresents key ideas from the lecture and/ or reading passage OR many of these ideas are unclear and/or irrelevant (i.e., not from the lecture or reading passage).	Explains key ideas mostly clearly despite some inaccuracies; includes some examples and explanations from input texts but does not sufficiently develop the key ideas	Explains key ideas clearly and accurately; adequately develops ideas with examples and explanations from the lecture and reading passage.	Explains key ideas clearly and accurately; fully develops ideas with examples and explanations from the lecture and reading passage.

Appendix E

MQUC Listening to Speaking Test Learning Outcomes and Marking Rubric Excerpt

Learning Outcomes to be tested

Listening Comprehension	Speaking	Interaction
L1: Can follow an informal discussion between two speakers on topics normally encountered in academic life	S1: Can convey meaning through intelligible pronunciation and intonation	IN1: Can synthesise and relay information from a spoken text without changing meaning
L2: Can distinguish key ideas, opinions, and supporting examples	S2: Can manipulate language structures in sustained interactions	IN2: Can initiate, maintain, and end discourse with effective turn taking
L3: Can understand the speaker's point of view	S3: Can express own ideas and opinions clearly	IN3: Can help the discussion along, confirm comprehension, and ask for clarification

Sample of a criterion from the rating rubric

MQUC Listening to Speaking Test		Fail	Pass	Distinction	High Distinction
Communicative effectiveness (45%) Engages in extended discussion by relaying, explaining, and opining upon the relevance of detailed information from a recorded dialogue.	**Listening comprehension (20%)** d interactions by third parties on familiar and unfamiliar topics encountered in social and academic life.	Demonstrates limited understanding of the key ideas from the dialogue.	Demonstrate a satisfactory understanding of the key ideas from the dialogue.	Demonstrates good understanding of the key ideas from the dialogue.	Demonstrates clear and precise understanding of the key ideas from the dialogue.
	Conversational interaction (25%) Establishes conversational relationship with speaking partner through sympathetic questioning, expressions of agreement, and indications of reservations or disagreement.	Uses a limited range of conversational techniques to manage interaction; may rely mostly on interlocutor.	Uses conversational techniques to manage interaction, with varying degrees of success.	Uses a range of conversational techniques to manage sustained interaction.	Uses a broad range of conversational techniques to manage sustained interaction.

Appendix F

Framework for Argument Based Validation for Good Practice in Test Design

Validity Aspect	Questions to ask	Considerations	Examples of evidence for ELICOS Centres
Construct Validity	To what extent do the test results conform to the model of communicative language ability that underlies the test?	Map learning outcomes to be tested to CEFR model of language use.	• Test specifications with clear statement of learning outcomes. • Rating rubrics demonstrating a clear means of measuring achievement of learning outcomes. • Student samples demonstrating a range of levels of achievement.
Content Validity	To what extent does the test cover the full range of knowledge and skills relevant to real-world situations and authentic language use?	Understand the TLU domain at the English Centre & the university. Select authentic texts.	• Review literature reviews, policy and practices in HE. • Undertake focus group studies with HE stakeholders.
Reliability	To what extent are test results and feedback precise, stable, consistent, and free from errors of measurement?	Test specifications, rater training, moderation, rater analysis, item response analysis.	• Manual of procedure for rater training, standardisation, and moderation of test results. • Log of activities, outcomes, problems, and solutions for each delivery. • Rasch analysis of item response and rater performance
Criterion-related evidence	To what extent do test scores correlate with a recognised external criterion which measures the same area of knowledge or ability?	Map to CEFR levels of communicative language competencies.	• Benchmarking studies i. Mapping of outcomes to external, validated outcomes ii. Comparative study of test-taker performance on other validated tests • Cross-institutional peer review • Tracer studies with university • Focus group study with students and academics
Concurrent validity	To what extent does our test correlate to established, valid tests which are measuring the same skill/s?	Have test takers with a recent TOEFL/IELTS/ PTE test score take the placement test & correlate results.	
Predictive validity	How accurately does the test predict test takers' future performance?	Monitor effectiveness of placement decisions based on placement test results.	

Validity Aspect	Questions to ask	Considerations	Examples of evidence for ELICOS Centres
Fairness	To what extent do test design, validation, development, administration and scoring procedures minimise construct-irrelevant variance?	The consistent use of clear test specifications, examiner and test-taker instructions, procedures for marker training and moderation and regular test analysis and review.	Testing manual for design, implementation, moderation, and management of test results.
Quality of service	To what extent can we assure the provision of secure examination materials, the confidentiality of examination data and results, and procedures to hand enquiries about results and appeals?	Administrator's handbook covering all parts of testing (see ALTE good practice guide).	Administrative manual for Education Managers, Senior Teachers, Coordinators, and teachers detailing processes and procedures for test development, test implementation, test marking, and the management and communication of test results.
Practicality	Are the resources in place sufficient to meet the needs of quality control for the exam in its intended design?	Assess staffing and/ or funding resources and if resources are exceeded consider modifying the test design or securing an increased allocation of resources.	Documented quality assurance processes and outcomes for each delivery, monitoring success rates, highlighting issues, and proposing action plans.
Impact	What are the positive impacts of the test at the macro level (general educational processes) and micro level (individual stakeholders)?	Macro: positive washback, constructive alignment, professional support for assessment and teaching teams who use the test. Face validity among students, teachers, administration, and parents.	Constructively aligned syllabus. Student surveys that include rating and commentary on the perceived relevance, clarity, and student preparedness for DEP assessment tasks. Teacher feedback on aspects of learning and teaching, and appropriateness of the assessments.