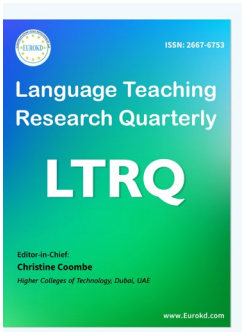

Language Teaching Research Quarterly

2021, Vol. 26, 18–38

A Collaborative Approach to Assuring Standards: Using the CEFR to Benchmark University Pathway Programs' English Language Outcomes

Thomas Roche[1]*, Sara Booth[2]

[1]Southern Cross University, Australia

[2]Peer Review Portal, Australia

Received *20 August 2021* ***Accepted*** *23 November 2021*

Abstract

The past two decades have witnessed growing interest in Higher Education (HE) achievement standards. Globally, the English Language Teaching (ELT) sector provides students with Direct Entry (DE) English language program (ELP) pathways to university study. It has typically relied on commercially available English language tests to ensure achievement standards. Such tests enable providers to demonstrate evidence of DE ELP standards in terms of internationally recognised scores. English language tests, however, represent only one type of external reference point for assuring standards. This paper provides an overview of a sector-led, collaborative approach to an external review of standards of 20 Australian university-operated DE ELPs. Twenty-eight ELP subjects were benchmarked by sixty individual assessors using consensus moderation with the newly expanded Common European Framework of Reference for Languages Companion Volume (CEFR CV). The framework's new subscales, including those on mediation, were employed for assessing student work samples. The case study demonstrates how collaborative benchmarking using the updated CEFR CV can effectively assure DE ELP standards while also identifying areas for program improvement to the benefit of programs, staff, and students.

Keywords: *Achievement Standards, Benchmarking, English Language Teaching (ELT), CEFR, English For Academic Purposes (EAP), Academic Literacies*

thomas.roche@scu.edu.au

doi:10.32038/ltrq.2021.26.02

Introduction

The past twenty years has seen increasing interest in Higher Education (HE) standards of achievement and program learning outcomes comparability. In the United Kingdom (UK) and Australia, for example, the HE sector has historically focused on assuring academic standards through two key assessment approaches: 1) standardised educational testing and 2) external peer review of the assessment. Both approaches are resource and time intensive and are more commonly used within individual institutions rather than in collaborative approaches to assuring standards.

In recent times, both the UK and Australia have moved towards a sector-based approach for assessing student learning outcomes, including the use of external reference points and standards frameworks. This has been motivated in part by a desire to assure learning standards for students and by a desire to achieve cross-institutional comparability (Bloxham & Price, 2015; McCubbin et al., 2021). With academic standards in place, students should be able to achieve comparable learning outcomes across HE institutions. To meet this objective, in the United Kingdom, the Quality Assurance Agency for Higher Education (QAA) has provided a framework for setting and maintaining academic standards through subject benchmark statements; whilst in Australia, legislative changes to the *Higher Education Standards Framework* (HESF) *(Threshold Standards)* of 2015 require institutions to produce evidence of outcomes and standards.

This paper reports on a sector-based approach to ensuring program standards. It outlines a collaborative project involving 20 Australian universities using the recently expanded Common European Framework of Reference for Languages Companion Volume (CEFR CV) (CoE, 2018), a validated language proficiency framework, to benchmark program outcomes and improve the curriculum in their Direct Entry (DE) English language university pathway programs. While this is an Australian study, it has relevance to university DE English Language Programs (ELP) providing pathways to university study globally.

Assuring the standards of DE English language programs

To gain admission to HE programs in Australia, prospective international students need to provide evidence of their English language proficiency. The Australian Department of Home Affairs (DoHA) sets standards of evidence students are required to provide on the application. These are framed in terms of internationally recognised test scores (e.g. IELTS or PTE Academic, DoHA, 2019). Individual universities set their own course-specific English entry requirements (i.e. for Diplomas, Bachelor, Masters) in terms of these same test scores, though in some instances, English requirements may be met on successful completion of a DE ELP or other non-award pathway program (e.g. Foundation or Qualifying Programs). With over 150,000 enrolments in 2018, approximately 30% of commencing international students gained university admission via a DE ELP (DET, 2019).

HE providers in Australia are required by the HESF (TEQSA, 2015) to demonstrate that students' learning outcomes are benchmarked against external standards, reference points describing what students should be able to do and know (Sadler, 2007). External referencing for

the purposes of this paper not only concerns benchmarking program design and methods of assessment, but also student achievement of learning outcomes through student cohort performance analysis and peer review of assessment, including calibration of different markers' grading (see HESF Standard 5.3.4; TEQSA, 2019b). The Australian Business Deans Council (ABDC, 2019) has, for example, already developed its own disciplinary standards against which program outcomes are routinely benchmarked (Watty et al., 2014). Australian universities belonging to two networks, the Innovative Research Universities (IRU) group and the research-intensive Group of Eight (Go8), have undertaken small-scale external referencing exercises (McCubbin et al., 2021). Collaborative external referencing of university DE ELP standards has not been widely reported to date, and, given their important role as pathways to university, they are the focus of this paper.

It is widely accepted that student learning outcomes are best evidenced in assessment performance (Bloxham & Boyd, 2012; Sadler, 2007). In keeping with this, the updated Australian national policy for the English Language Teaching (ELT) sector, *ELICOS Standards 2018* (Australian Government, 2018), includes a general requirement for providers to ensure that their assessment is "valid, reliable, fair and clearly referenced to criteria". The *ELICOS Standard P4.1c* (ii) requires all programs to have formal mechanisms "to ensure that assessment outcomes are comparable to other criteria used for admission to the tertiary education program of study, or for admission to other similar programs of study" (2018). These formal measures include tracer studies, external testing, benchmarking to validated language proficiency frameworks, and external review of the assessment of inputs and outputs (TEQSA, 2019a). There are significant risks to educational quality at HE institutions which do not comply with these standards, including 1) failing to establish the rigour and equivalence of assessment outcomes; 2) failing to address challenges by particular student cohorts; 3) not providing sufficient language support; and 4) not monitoring the attrition rates of DE student cohorts (TEQSA, 2019b).

This paper reports on a case study, a sector-led external benchmarking review of DE ELP standards and student learning outcomes managed through an Australian HE network: University English Centres Australia (UECA), an affiliation of 32 English Centres and Colleges either wholly owned or operated by their parent institution.

ELP External Reference Points

Internationally, several established external reference points are used for benchmarking ELP learning outcomes, including the Cambridge English Scale, Pearson's Global Scale of English, and the CEFR. The CEFR was originally developed as a reflective tool for describing threshold standards in language learning (North, 2014), with the primary goal of achieving score comparability across tests and programs across Europe. Since then, it has been used to facilitate score transparency between a wide range of university admissions tests across Europe (Deygers et al., 2018c). The CEFR is a user-oriented proficiency scale, which describes language use through three basic categories with six levels: *Basic user* (A1 & A2), *Independent user* (B1 & B2) and *Proficient user* (C1 & C2). It contains generalised descriptions of what learners are

likely to be able to do at any given proficiency level in the form of positive can-do statements. As such, it is well suited to be used as an external reference point for assessing DE ELP standards.

The CEFR is not without critics. Researchers have noted the challenge of using the CEFR as an external reference point due to the descriptional inadequacy of the wording of the scales (Fulcher et al., 2011, p. 8). For example, descriptors have not been produced for every level of the scale with gaps in levels for some language features (Alderson et al., 2006); the terminology used has been criticised as impressionistic (Alderson, 2007; Fulcher, 2004), vague and incoherent (Harsch & Rupp, 2011). It can be difficult to apply the scales in a range of contexts. For instance, it is widely acknowledged that until 2017, versions of the CEFR underrepresented the complexity of academic writing (McNamara et al., 2018). These inadequacies can lead to various interpretations of language used against the levels by assessors. Beyond concerns with CEFR descriptors themselves, in order to use the CEFR as an external reference tool, users need to identify the CEFR subscales which match the construct of English they are using in their program design and assessment (Harsch & Martin, 2012; Harsch & Rupp, 2011). In order to improve assessment against the CEFR, assessment designers should be involved in developing subscales together, first consider the purpose of the scales (Knoch et al., 2021), identify the relevant elements of those subscales and then be trained together to align judgements (Deygers et al., 2018b).

In 2018 the CEFR CV (CoE, 2018) was released to clarify descriptive elements of the CEFR with the inclusion of new scales not in the original CV. In doing so, it addressed some, but not all, of the concerns in the literature (Deygers, 2021). There are, however, some grounds for using the new CEFR CV for benchmarking DE English language programs' standards. Firstly, the CEFR CV remains user-oriented. The framework's descriptors are readily comprehensible to both educators and language learners because they describe the real-world language abilities, which are the aim of most language learners (North, 2014). Secondly, the CEFR CV provides improved graduation of standards through a more elaborate description of the reference levels, including 'plus' levels for B1+ and B2+ and more descriptors for the 'C' levels (Goodier, 2018) which are particularly suited to DE pathways into linguistically demanding university courses. A third reason for using the CEFR is that it is widely used internationally. It has been increasingly used beyond Europe (e.g. in the Americas: e.g. Canada and Colombia see Normand-Marconnet & Bianco, 2015; and in Cuban higher education, see Harsch et al., 2020). It has been adopted in policy in Thailand as a frame of reference for assessing university graduates' English language proficiency (Wudthayagorn, 2021); used in Vietnam to develop the national Vietnamese Standardised Test of English (VSTEP) (Nguyen & Hamid, 2021); adapted in Japan to formulate the CEFR-J; and, used to inform the development of the Common Chinese Framework of Reference for Languages (CCFR) (Read, 2019). Therefore, if these standards are adopted in a benchmarking project, it would make the programs' DE English achievement levels readily interpretable globally. Finally, the CEFR CV includes new scales for mediation (Deygers et al.,

2018a; Deygers, 2021) describing standards for relaying information and others' ideas, a skill critical for university students (e.g. paraphrase and referencing) (Ahmed & Roche, 2021)[1].

Following a literature review and UECA member survey (Roche & Booth, 2019), UECA decided to adopt the CEFR CV as their external frame of reference for assuring DE English program standards. This paper reports on a collaborative project employing the updated CEFR CV as an external reference point for benchmarking university DE ELPs and findings from that project.

Materials and Methods

Participants

Twenty Australian UECA institutions participated in the national benchmarking review with a focus on written assessment standards (See Appendix A). Typically, three participants from each institution joined as assessors (60 in total). The assessors all held a recognised undergraduate degree, as well as a Teaching English to Speakers of Other Languages (TESOL) qualification, or an undergraduate degree in education with a TESOL method focus as a minimum educational qualification.

Project aims

The project aimed to establish cross-institutional comparability of learning outcomes (Bloxham & Price, 2015; McCubbin et al., 2021) by using subscales of the CEFR CV to review assessment items and student performance on those items (Bloxham & Boyd, 2012; Sadler, 2007). This is a methodology widely used in higher education benchmarking (Sankey & Padró, 2016; Syme et al., 2021). The benchmarking involved 20 participating Centres and was established as UECA's *External Referencing of the ELICOS Standards* (ERES) project. The project's key aims were to:

1. Benchmark assessment policies and processes across UECA member Australian universities
2. Externally peer review assessment and student work samples in English Language programs to compare achievement standards
3. Build capacity for Australian English Language Centres to participate in external referencing and exchange activity to improve their own educational performance
4. Develop institutional and national actions and share good practice with other institutions

This paper reports on the project: describing how the programs were benchmarked, the resulting findings, including recommendations for program improvement identified; and, the benefits and challenges of using the CEFR for this purpose. It also then discusses implications for future benchmarking of DE ELPs.

Methodology

At a macro-level, the project employed a mixed methodological benchmarking approach based on the Australasian Council on Open, Distance and e-learning's (ACODE) benchmarking

[1]The 19 new mediation descriptor scales refer to involves the (re)processing of an existing text, and accounts for language used to relay information and or synthesise a text (CoE, 2018, *Section 2.1.3*, p.14).

approach (Sankey & Padró, 2016). Similar approaches have also been used for benchmarking university pathway programs for domestic students in Australia (e.g. for enabling or bridging programs see Syme et al., 2021). The key features of ACODE's benchmarking methodology, originally used to develop standards for technology-enhanced learning, includes the development of key performance indicators (KPIs) and key performance measures (KPMs) and peer reviewer questions. These are outlined below.

A steering committee (the UECA Committee) agreed to a set of aims (see 2.2), procedures (see 2.3.1) and outputs as well as underlying principles and a timeline (see Table 1) as is typically done in benchmarking projects (Booth & Coolbear, 2015; Morgan & Taylor, 2012). The benchmarking project team developed the underlying principles which participating centres agreed to: mutual respect, a willingness to share and learn from participants; and, a shared commitment to quality improvement and enhancement. Assessment tasks and work samples submitted by participants were treated as confidential documents. Table 1 outlines the five key project phases for the national benchmarking project (May 2018-Aug 2019).

Table 1
Six Project Phases (May 2018-Dec 2019)

Phase 1: Introduction to Project	**Phase 2: Project Management**	**Phase 3: Self-Review**	**Phase 4: Review and Calibration**	**Phase 5: Final Report**	**Phase 6: Review and implement recommendations**
May-July 2018	Aug-Dec 2018	Jan 2019-Apr 2019	May-Jun 2019	Jun-Aug 2019	Sep-Dec 2019
Literature Review. Agreement on aims, principles, outputs, timeline, KPIs and KPMs for benchmarking template (including introductory workshop on 13th July 2018)	Sign collaboration agreements Identify institutional coordinators Provide support documentation on CEFR	Self-review of assessment policies and processes Self -review Report Peer Review Schedule Update support documentation on CEFR	Peer review workshop Compare assessment policies and processes Calibrate results Identify good practice, improvement, and enhancement	Final Report National and Institutional actions UECA Committee endorsement Project Evaluation	Centres review individual reports, consider suggested areas for improvement and implement recommendations.

The next step was to develop the scale for the benchmarking of program outcomes and student samples. The project leads adopted an iterative approach (Harsch & Martin, 2012; Deygers et al., 2018b) to develop a CEFR sub-scale for the UECA benchmarking project. Assessors developed the subscales together (Deygers et al., 2018b). First, they identified a sub-

set of CEFR CV (CoE, 2018) scales that aligned the construct of English that the participating Centres were using in their program design and assessment (Harsch & Seyferth, 2020; Knoch et al., 2021). The subscales used were taken from the CEFR CV and included the written assessment grid (2018, 173-235) and written reports and essays grids (2018, 77).

Collaborative peer review questions

In Phase 1 of the project, a meeting with centre managers was arranged to discuss what was to be benchmarked (KPIs) and how (KPMs). The following peer review questions were developed as measures of the KPMs:

KPI#1: English Standards across DE English pathway programs

- *KPM1.1: What internal processes and policies are in place for moderating assessment in DE Programs? Are these effective?*
- *KPM1.2: What external reference points are used to validate assessment in DE Programs? Are these effective?*
- *KPM1.3: What formative and summative assessment tasks are used in DE Programs, and how do these assessments map against the stated learning outcomes? Are these effective?*

KPI#2: Monitoring and Tracking for Continual Improvement of DE English pathway programs

- *KPM2.1: What processes are in place to monitor student progress and assist students at risk in DE Programs? Are these effective?*
- *KPM2.2: What data is collected and analysed from students and stakeholders to ensure continual improvement in DE Programs? How effective is this process?*
- *KPM2.3: What strategies are in place to track student success after completing a DE Program? Are these effective?*

KPI#3: Calibration of Assessment and Student Work Samples across DE English pathway programs

- *KPM3.1 Are the Unit Learning Outcomes (ULOs) for Program Level Outcome (PLOs) clearly specified and appropriate?*
- *KPM3.2 Are the Unit Learning Outcomes appropriate at the Grade/Exit levels (as benchmarked against the appropriate CEFR levels)?*
- *KPM3.3. Does the assessment task/s design enable students to demonstrate attainment of the relevant ULO's and relevant PLO's?*

These were developed so that peer reviewers could provide responses in the form of 'yes'; 'yes, but'; 'no, but'; and 'no' ratings, which were then quantified as part of the external review process. Reviewers were also able to provide an additional free-text response to elaborate on these ratings, identify areas of good practice or suggest areas for improvement.

Collaborative peer review process

There are a number of peer review processes that HE providers can use to demonstrate that their program is fit for purpose and that students meet threshold standards (Bloxham & Price, 2015; Sefcik et al., 2018). Peer-review used by the Quality Verification System (QVS) developed by the Australian Group of 8 Universities (Go8, 2013) and the Academic Calibration Project run by

the Australian Research-Intensive Universities in 2014, for example, use a single, randomly assigned but non-blind external reviewer. However, studies of single-peer reviewers assessing HE standards in the UK (Bloxham et al., 2015) and using the CEFR (Deygers & Van Gorp, 2015; Harsch & Hartig, 2015) have found that individuals can vary in their interpretation of those scales. All judgements, such as assessments of standards, are characterised by uncertainty (Kahneman et al., 1982), and if they are not calibrated, they can exhibit unacceptable levels of inter-rater reliability between assessors (Lumley, 2002). In light of this, pairs of reviewers rather than single assessors were used in the *ERES* project.

Establishing standards for language programs is a complex process (Harsch & Kanistra, 2020). The importance of training assessors together to align judgements has also been noted in the ELT literature (Deygers & Van Gorp, 2015; Deygers et al., 2018b). To reduce uncertainty and calibrate judgements (Figueras et al., 2005) in the *ERES* project, all external reviewers were sent the new *CEFR CV* and a series of three *UECA Benchmarking Project Guidance Documents* with direction on interpreting the CEFR standards. The documents were developed by qualified and experienced English for Academic Purposes educators and assessment specialists to guide interpretation of the CEFR and the rating of student work samples. To further support that judgments made by assessors were comparable in terms of external standards, the UECA national benchmarking project also adopted consensus moderation practices: "a process for assuring that an assessment outcome is valid, fair and reliable and that marking criteria have been applied consistently" (Bloxham, 2009, p.4). Studies have found consensus moderation to increase both inter-rater reliability and assessor confidence (O'Connell et al., 2016). Due to this, it has become an established practice in HE external reviews of assessment to support assessors make judgments that are comparable in terms of criteria and standards (Bloxham et al., 2015; Booth, 2017). Calibrating assessor-judgements through such a process has been used with CEFR external benchmarking of writing in other educational contexts focusing on high school student writing (Harsch & Martin, 2012) and with university staff (Deygers & Van Gorp, 2015).

In the *ERES* project, student work samples submitted for benchmarking were first assessed independently by two judges at each institution. These were then referenced against the descriptors of the CEFR scales so that assessors could judge the work as being either B1+, B2, B2+, or C1 and took notes providing supporting evidence of these levels in the text. Then two assessors met to compare their provisionally allocated marks. Following this, assessors engaged in a discussion about how marks should be allocated to justify their rating before reaching an agreement on the assessment outcome for each piece of submitted work. This allowed for a consensus CEFR level to emerge for each submitted item, based on the evidence present in the submitted work (Harsch & Hartig, 2015). A national calibration workshop was held where over 40 participants from 20 participating institutions shared experiences of rating and assessed de-identified samples of student work. Assessors worked in pairs and then in groups to see if they could achieve consensus in applying threshold learning standards to the students' work. Percentage Agreement (PA) is considered a useful measure of agreement when external criteria are being employed in assessment (McHugh, 2012). Assessor agreement on student work

samples presented at the calibration workshop was above 80%, which indicated strong agreement in their assessment. Each participating Centre was allocated between 2-3 partner institutions to review, with each review undertaken by a minimum of two trained assessors at partnering institutions.

Benchmarking material

Each participating English Language Centre submitted the following evidence for each program review:

- A unit/program outline
- A context statement outlining Unit or Program Learning Outcomes (PLOs), including a mapping of relevant written assessment tasks to these outcomes
- Relevant written assessment task sheets (for students)
- Relevant written assessment rubrics and marking guide (for teachers)
- De-identified student-written assessment examples:
- For each written in-program formative assessment item worth 20% or more, three samples each of a Pass and a (just) Fail with the accompanying marking rubric; and/or
- For each written Exit or Capstone/summative assessment, three samples each of up to three grades (Exit levels) with marking rubrics; and
- A table detailing numbered samples with awarded grades and nominal CEFR level.

Appendix A provides detail on the types of assessment items submitted by the centres. Each participating English Language Centre submitted evidence for their review using the Peer Review Portal (2019), which is a cloud-based review management system. The Peer Review Portal has been referred to by TEQSA as an optional online support mechanism for external referencing and peer review (See TEQSA, 2019b).

Results

The collaborative peer review resulted in detailed findings of individual centre's policies, processes and assessment standards, as well as recommendations for program improvement, which were shared with participating centres in institutional reports via the Peer Review Portal. The aggregated and de-identified results, which provides an understanding of current practice across the group, is presented here.

KPI#1English standards

Broadly, across all participating DE ELPs (n=20, 100%), effective policies and practices were noted as in place for moderating assessment marking and ensuring consistency of grading. Key areas of good policy and process practice included widespread use of assessment rubrics, online marking/scoring systems, which further helped standardise approaches to assessment across centre locations, routine formal assessment induction processes for new assessors/teachers and assessment validation approaches, including validation checklists. This finding is perhaps in some ways unsurprising, given that centres are often quality assured by an industry peak body, such as NEAS, which provides quality reviews programs with a view to them upholding standards, supports centres in demonstrating quality in their programs and services. Four centres

(20%) though received recommendations for improvement, such as the need to review marking rubrics and make them clearer for both assessors and students. External reviewers found the assessment rubrics in these centres to be, in some instances, overly complex, specifying numbers of errors rather than performance standards, and in some cases, the wording of the rubrics was unclear to the peer reviewers.

The collaborative peer review process identified that all twenty centres used external reference points to provide transparency on students' assessment achievement. The majority of the group relied on a range of external reference points. (e.g. Pearson's Global Scale of English, Cambridge English Scale Score, as well as public domain IELTS rubrics). This suggests that many of the centres were using test rubrics to set proficiency levels for their programs. Submitted policy and procedure documents indicated program reviews were widely used across all centres (n=20, 100%) to good effect for ensuring the quality of program design and assessment practices, as were advisory committees in a smaller number of centres. Fifteen centres (75%) were given suggestions for improvement, including recommendations for some centres to map their program learning outcomes more clearly to an external frame of reference for students in terms of "can do" statements rather than to a single test score or proficiency band such as an IELTS score.

The benchmarking identified a wide variety of formative and summative assessment tasks being used across the group (e.g. essays, reports, and critical reviews) with a range of 5 to 20 assessment tasks (1-3 of which were typically writing tasks) per centre for each DE pathway program. Despite this range, the program learning outcomes showed a great deal of similarity across the participating centres, with common academic language practices developed and assessed in all programs (n=20, 100%), including using academic language appropriately (e.g. vocabulary, collocations, syntax) and building an argument using scholarly sources. These practices were assessed in tasks through students' use of paraphrasing, summarising, synthesising, citation and reference. In one centre, digital literacy was explicitly stated as a learning outcome (e.g. Use a range of digital literacy tools and practices to appropriately access, assess and disseminate online information.). Some areas of improvement for individual centres were also identified, including ensuring individual assessment task sheets were clearly mapped against learning outcomes. A quarter of participating centres (25%) were recommended to review their volume of assessment and consider if they were over assessing their students.

Monitoring and tracking for continual improvement

All of the 20 participating DE pathway programs had effective processes for monitoring student progress and clear and actionable policy and procedure on supporting at-risk students. In some instances, peer reviewers identified inconsistencies in the use of at-risk processes across classes and locations. Many centres, it was noted, could do more to leverage technology to provide a more consistent and better-documented form of support for at-risk students.

The majority of the university English language centres collected and analysed feedback from students and stakeholders to ensure continual improvement in their DE programs. Good practice examples included learning management systems that were used to build student profiles. There

was evidence of robust statistical analysis of student performance data in a number of centres (e.g. Rasch analysis of student performance across discrete item tests was used in 2 centres (10%)). The majority of centres (n=16, 80%) used external quality reviewers, such as National English Language Teaching Accreditation Scheme (NEAS). Some areas for improvement within individual centres included a need for better sharing of student performance data across campus locations, and some centres were encouraged to set up more formal and routine reporting on program retention and attrition trends.

Though the majority of participating centres (n=13; 65%) tracked student success in their subsequent university studies after completing a DE pathway program, there was variation in how this was done across the group. In some centres, this happened routinely with reports shared at program review meetings or academic boards, whereas in other instances, it was done through informal feedback from program coordinators. Six participating centres (30%) did not track DE student performance after completing the program. The benchmarking reviews recommended the implementation of automated reports and processes for ongoing tracking of DE program students' success rates, retention rates and academic achievement, thereby providing a better understanding of student and DE English pathway program performance.

KPI#3: calibration of assessment and student work samples

Phase 4 of the project involved collaborative peer review of assessment tasks, samples of student work against the CEFR as an external frame of reference. For all participating centres (n=20, 100%), the ULOs were assessed as being clearly specified and appropriate for the PLOs. Peer assessors also agreed that the assessment tasks enabled students to demonstrate attainment of the relevant ULOs and relevant PLOs for all of the centres (n=20, 100%), but this included a number of "Yes, but" ratings (n=8, 40%). Feedback from external assessors in these instances highlighted that the identified ULOs lacked specificity for language features, such as grammar and vocabulary use, that they would need to make a confident judgement on standards of student work. However, the majority of feedback indicated the tasks were well-designed to ensure students were able to demonstrate their English language proficiency at the appropriate levels.

Peer reviewers then assessed the samples of student work against the CEFR CV subscales prior to seeing the centre's awarded marks. The collated feedback reports showed that of the 20 participating centres, 16 (80%) were found by peer reviewers to be "yes" assessing to standard with a further four assessed as "yes, but" (20%) or as assessing broadly to standard with a small number of student work samples identified as being either borderline or at another proficiency level. Feedback from reviewers across the centres demonstrated consensus that high, medium and lower samples fit the overall awarded proficiency ratings. In many instances, assessors indicated they would have appreciated a key or instructions on using the centres' rubric for checking their marks against the rubric. This feedback indicates that there was a high degree of similarity in DE program assessment standards across the participating centres with a small number of exceptions.

A survey of peer reviewers and centre directors from participating centres was undertaken in Phase 5 to evaluate the project. In response to the survey, over 90% of respondents agreed that

participation in the benchmarking project had enabled their centre to validate its policies, processes and assessment standards. One respondent noted the greatest value in the project was in receiving "*Confirmation from partner institutions that our assessments were well designed and appropriate."* While other respondents indicated the chief benefit of the project was in the professional development experienced by participating staff as "*it led to a consolidation of knowledge across Centre teams, and there is a stronger sense of professionalism emerging in participating staff"*. Participating institutions reported other benefits from the project, including receiving feedback on areas for improvement "*such as making parts of rubrics clearer and providing training packages for teaching staff so as to understand more about marking*." When asked if centres would be interested in undertaking further external referencing the majority of respondents were affirmative, though there was a sense more was to be done on improving methods for assessing writing before focusing on benchmarking other skills (e.g. speaking). Feedback indicated that assessor training could be improved by including annotated student work samples to help standardise CEFR level interpretation.

Discussion

While collaborative approaches have been used for benchmarking other university DE ELP outcomes (e.g. for domestic enabling or bridging programs see Syme et al., 2021), this paper presents an example of a collaborative peer review of university DE ELPs and establishes cross-institutional comparability of those programs' learning outcomes in Australia. It is novel in using the expanded CEFR CV for this purpose in that context. The project found that effective processes and policies were in place for moderating assessment as well as monitoring student progress across all participating centres' DE ELPs.

DE ELP learning outcomes

Across the group, there was almost universal agreement about the learning outcomes of the DE ELPs, with very similar wording used to describe English language proficiency and academic skills outcomes across individual centres, though there were differing degrees to which academic integrity and digital literacy were stressed as part of these. Importantly, the peer review confirmed that the majority of centres' ULO's were clearly aligned with assessment tasks (Biggs, 2014), at the stated proficiency levels. This case study has also shown how the described collaborative process can identify where programs were not designed or operating to standard. To improve program quality, recommendations to individual centres were made by expert reviewers. Following the benchmarking, individual centres were then able to consider and respond to the specific feedback provided on their program. A number of these key findings for participating centres are outlined below.

The benchmarking identified that participating centres currently rely on a range of external reference points (e.g. Pearson's Global Scale of English, as well as public domain IELTS rubrics) for setting their PLOs (and in some instances ULOs). In a limited number of cases, expert reviewers identified areas where centres' PLOs were not clearly aligned with assessment tasks. Centres were then recommended to review their assessment tasks and, where appropriate,

reword task sheets and rubrics. Some of the above external reference points used by centres are connected to high stakes English proficiency tests. This practice, no doubt to some extent, reflects the broader national context whereby the Australian visa granting authority DoHA sets English language standards for student visa application purposes in terms of test scores (e.g. IELTS). Individual universities frame than their own program English entry requirements; these are typically formulated in terms of those same DOHA recognised test scores. While this provides a sector-wide and indeed internationally shared understanding of PLOs in terms of test proficiency levels, it also results in programs frequently being driven, in terms of learning outcomes, assessment practices and teaching content, by high-stakes test constructs rather than by the academic English skills students need to learn in order to achieve in university studies. For example, some centres assessment rubrics were described in language, which can be understood as coterminous with coherence, lexical range/accuracy, and grammatical range/accuracy. While these are undeniably fundamental structural elements common to a number of academic English written texts (e.g. reports and essays), and there is evidence that assessment tasks that measure students' English proficiency against these features are good predictors of performance in students' first year of study (Humphreys et al., 2021); it was notable that other important elements of academic English texts were often absent from rubrics across the group (e.g. paraphrasing, referencing). This would enable staff teaching into DE programs to further develop international students' understandings of academic integrity practices (i.e. text authorship and ownership practices) used in Australian universities, which they often struggle to use appropriately (Flowerdew & Li, 2007). As a result of their absence in PLOs and ULOs, it became apparent that some centres did not focus in their program learning outcomes on those higher-level English language skills which are essential to university study (Ahmed & Roche, 2021). In addition to the above noted textual features which were missing from the ULOs and PLOs, it was suggested by some reviewers that given the increasingly digitised higher education sector into which DE pathway students matriculate; centres should consider more explicitly addressing digital literacies in their ULOs and PLOs. A number of centres could then operationally improve the relevance of their programs to university study by migrating more of the delivery of their assessment tasks to the online Learning Management Systems (LMS) used at their parent institution and explicitly embedding additional learning outcomes relating to academic literacies (including digital literacy) in their program design and assessment practices.

It is important to note that some centres felt that aligning their PLOs, assessment practices and teaching material with high-stakes test descriptors was appropriate, given these were the measures of proficiency their parent institutions saw as the relevant standards for admissions purposes. The review of PLOs and ULOs in the benchmarking project suggests that rather than having their English construct determined by default by test scales, participating centres should revisit the construct they are using to underpin their course design and consider if that construct, as expressed in their PLOs and ULOs, is relevant in terms of the English skills their students need in their future studies (Knoch et al., 2021),

In terms of assessment practices, the collaborative benchmarking identified great variability in the volume of assessment used across the centres. In the review, some centres were seen to use over 18 discrete assessment tasks over the course of a 10-week program; others used only 5 with nested sub-tasks. This finding raised the question about the appropriate volume of assessment of a 10-week DE ELP; while no standards were set in this project about appropriate assessment volume, the issue was identified for further exploration.

Another issue that arose across the group from the benchmarking was the general absence of routine tracking mechanisms for monitoring student achievement after the successful completion of their DE ELP. Few of the centres were able to report on how their students performed after completing their courses in comparison with other international students entering on the basis of recognised proficiency test scores. Centres should investigate methods for tracking their students in their first year of post-DE study. These data will provide further validation of their PLOs and assessment standards. A number of participants reported that obtaining tracking data is not always straightforward as it requires centres to work across a range of operational units in their parent institution and comes with a resource implication.

The CEFR CV sub-scales as an external frame of reference

Through the national benchmarking project, UECA member centres developed a shared understanding and interpretation of English language learning standards through the use of an agreed-to external frame of reference: a set of sub-scales from the CEFR CV. One of the benefits for participants from this project was the shared focus on the CEFR CV's new sub-scales capturing the academic literacy practices often taught in university DE English programs. For example, academic writing frequently requires combining and synthesising different source texts (Cho & Bridgeman, 2012; Hyland, 2006) and in recognition of which, many DE ELPs focus on developing students' ability to summarise and paraphrase informative texts and use other authors' ideas in their own academic writing (Roche, 2017). As noted in the section on DE ELP learning outcomes, these are also fundamental academic English language practices international students often struggle to develop without explicit educative interventions (Ahmed & Roche, 2021; Flowerdew & Li, 2007); and, that commercial high-stakes English tests such as IELTS and TOEFL, do not assess. As such, many in the group found that CEFR CV descriptors provided a broader view of what language is to be taught and assessed, moving discussions of standards from a test-driven focus on sub-skills (speaking, reading, writing and listening) to higher-level language skills necessary for academic study; thereby, encouraging educators to consider these as core to DE ELP learning, assessment and outcomes (O'Sullivan in Plenter-Vowles, 2018).

While the CEFRV CV proved to be a useful frame of reference for setting written assessment standards, some issues with the framework also arose during the project. Despite being updated, and now better capturing the features of academic writing, as seen in reports and essays, problems remain with the scales (DeGeygers, 2021). For example, there are still gaps in the scale; some features lack descriptors at certain levels (e.g. B2+ Grammar and Vocab range). There are occasionally ill-matched descriptors for academic language (e.g. C1+ appropriateness

"including emotional, allusive and joking usage") which remain distracting, or at worst, confusing for assessors. The greatest issue for reviewers, though, proved to be the descriptors of the B2 band, which are broad and account for a breadth of proficiency that other scales break down further (e.g. in terms of a test score IELTS 5.5-6.0). Many centres were assessing programs that had learning outcomes lower than a B2+, and for those which did, this broad B2 band did not enable them to adequately distinguish between lower-level learners' proficiency levels. In other words, descriptional inadequacy (Fulcher et al., 2011) still characterises elements of the CEFR CV (DeGeyers, 2021). To remediate these deficiencies, further work could be done developing a UECA CEFR CV-based rating sub-scale which would form the basis of a more reliable rating (Lumley, 2002). Such scales would need to more accurately describe the relevant features of common assessment tasks (e.g. an essay, an annotated bibliography, a report). Best practice indicates that assessors should be involved in the further development of CEFR CV sub-scale wording (Harsch & Kanistra, 2020; Harsch & Martin, 2012). Sector validated examples, with annotations of how these meet the scale, could then be used across centre members to further standardise assessments of student work.

Future collaborative benchmarking of DE ELP programs

As reported above, nearly all of the surveyed expert peer reviewers agreed that taking part in the collaborative benchmarking project was of value. The project achieved its overarching aim to establish cross-institutional comparability of learning outcomes, and in doing so, validated centre policies, processes, and assessment standards. The three layers of participants as used here greatly facilitated this achievement: a steering committee which set the project aims then reviewed and approved subsequent tools associated with the project (e.g. the procedures, principles and CEFR CV subscales employed to assess standards); a small project leadership team for operationally driving the project (developing procedures, principles, a timeline and collating findings); and finally, expert peer reviewers for compiling material for review and assessing student work samples. The online Peer Review Portal greatly facilitated the efficient, paperless exchange of documents as well as the production of reports for individual centres and the steering committee. During the project, however, it became clear that expert peer reviewers benefited from more regular drop-in sessions than were initially planned for participants. The literature on benchmarking identifies the importance of assisting individual examiners to assess standards consistently (Elder et al., 2005; Sadler 2007) through encouraging the use of assessor training to lessen unintended variance between assessors (e.g., Lumley, 2002). While the training employed in this project was well-received, participants indicated in feedback that they would have liked more opportunities to calibrate their judgements. In order to maintain good levels of consistency in judgments on standards in student work samples, moderation sessions should be used prior to the assessment of student work, and in addition, a sample of work can be used in a moderation session after a round of review to check rater judgements (Bloxham & Price, 2015; Sadler, 2013). Forming more regular scheduled rater training sessions could also help develop a community of practice amongst participants for informal sharing of best practices. Whether UECA chooses to continue using the CEFR CV subscales identified here or further adapt those,

assessors will need to continue rater training in order to minimise rater variability. Research published elsewhere (Deygers & Van Gorp, 2015; Harsch & Kanistra, 2020; Harsch & Martin, 2012) show how a CEFR-based scale can be co-constructed by novice raters as well as how rater judgments can be collected and reported on in an effort to improve rater judgments and validate the scale.

Conclusions

This paper demonstrates how a collaborative approach can be employed to assure university DE programs' English language standards. It has outlined a relevant, transparent and systematic process for benchmarking DE ELP performance measures for comparing learning outcomes and student achievement standards. For the group, this collaborative project determined that each of the participating programs shared core equivalences in terms of learning outcomes and English standards. Using the CEFR CV as an external point of reference enabled participating centres to confirm they were producing graduates with requisite levels of English language proficiency and that graduates of their programs were able to produce extended written academic texts, such as essays and reports, showing they can clearly distinguish their own ideas from those in source material through paraphrasing and referencing in their writing. It is of note that these fundamental academic language skills (e.g. mediation - paraphrase and synthesis) are currently beyond the measure of many internationally recognised high-stakes discrete-skills tests. Despite these affordances, the current CEFR CV scales are not without shortcomings. Some of the descriptors remain vague or are missing detail at particular levels. The breadth of the CEFR B2 level, for example, is wholly inappropriate for benchmarking programs that require a proficiency distinction within that broad band. Benchmarking institutions and groups need to carefully weigh the advantages of the scales identified here with the challenges they present. The project as described here also establishes the first iteration of relevant CEFR CV sub-scales for the UECA group, which could be further developed for further cross-institutional benchmarking of university DE ELPs. Such a tool could provide much needed independent expert assurance of quality learning outcomes across the sector. The project also provided all centres with detailed, expert feedback, with areas for improvement identified for each centre. These changes can be considered, implemented, and their impact monitored in future peer reviews or through the centres' own internal quality processes. Finally, this collaborative peer review has consolidated the work of individual ELP academics and centres across a shared professional organisation, clarifying the value proposition of UECA's programs to parent institutions and government agencies. The project has shown how best practices can be shared across a professional network through collaborative peer review to the benefit of institutions' programs, their staff and ultimately their students, whether that be in Australia or further abroad.

References

Ahmed, S. T., & Roche, T. (2021). Making the connection: Examining the relationship between undergraduate students' digital literacy and academic success in an English medium instruction (EMI) university. *Education and Information Technologies*, 1-20. https://doi.org/10.1007/s10639-021-10443-0

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659-663. https://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.2007.00627_4.x

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly: An International Journal, 3*(1), 3-30. https://doi.org/10.1207/s15434311laq0301_2

Australian Business Deans Council (ABDC). (2019, September 19). *Learning Standards*. https://abdc.edu.au/teaching-learning/learning-standards/

Australian Government (2018, September 13). *ELICOS Standards 2018*. https://www.legislation.gov.au/Details/F2017L01349

Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of higher education*, *1*(5), 5-22. https://search.informit.org/doi/pdf/10.3316/informit.150744867894569

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

Bloxham, S., & Boyd, P. (2012). Accountability in grading student work: Securing academic standards in a twenty-first century quality assurance context. *British Educational Research Journal*, *38*(4), 615-634. https://doi.org/10.1080/01411926.2011.569007

Bloxham, S., Hudson, J., den Outer, B., & Price, M. (2015). External peer review of assessment: an effective approach to verifying standards? *Higher Education Research and Development, 34*(6), 1069-1082. https://doi.org/10.1080/07294360.2015.1024629

Bloxham, S., & Price, M. (2015). External examining: fit for purpose? *Studies in Higher Education*, *40*(2), 195-211. https://doi.org/10.1080/03075079.2013.823931

Booth, S. (2017). A cost-effective solution for external referencing of accredited courses of study. *TEQSA 2017 Conference Proceedings*, Melbourne: pp.126-145. https://www.teqsa.gov.au/teqsa-conference-2017

Booth, S., & Coolbear, P. (2015, March 9-11). *Enhancing our understanding of the potential of international peer review benchmarking for quality improvement*. In *Proceedings of 2015 AAIR Annual Forum*. University of Tasmania, Australia.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421-442. https://doi.org/10.1177/0265532211430368

Council of Europe (CoE) (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment Companion Volume with New Descriptors*. Strasbourg: CoE. Retrieved 29.08.2018, from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Department of Education, Skills and Training (DET). (2019, June 17). *End of Year Summary of International Student Data 2018*. https://internationaleducation.gov.au/research/international-student-data/pages/default.aspx

Department of Home Affairs (DoHA). (2019, June 17). *Subclass 500: Student Visa. Visit Australia to participate in a course of study*. https://immi.homeaffairs.gov.au/visas/getting-a-visa/visa-listing/student-500

Deygers, B. (2021). The CEFR Companion Volume: Between Research-Based Policy and Policy-Based Research. *Applied Linguistics*, *42*(1), 186-191. https://doi.org/10.1093/applin/amz024

Deygers, B., Van den Branden, K., & Van Gorp, K. (2018a). University entrance language tests: A matter of justice. *Language Testing*, *35*(4), 449-476. https://doi.org/10.1177/0265532217706196

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

Deygers, B., Van Gorp, K., & Demeester, T. (2018b). The B2 level and the dream of a common standard. *Language Assessment Quarterly*, *15*(1), 44-58. https://doi.org/10.1080/15434303.2017.1421955

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018c). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly*, *15*(1), 3-15. https://doi.org/10.1080/15434303.2016.1261350

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly: An International Journal*, *2*(3), 175-196. https://doi.org/10.1207/s15434311laq0203_1

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22*(3), 261-279. https://doi.org/10.1191/0265532205lt308oa

Flowerdew, J., & Li, Y. (2007). Language re-use among Chinese apprentice scientists writing for publication. *Applied linguistics*, 28(3), 440-465. https://doi.org/10.1093/applin/amm031

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonisation. *Language Assessment Quarterly*, *1*(4), 253-266. https://doi.org/10.1207/s15434311laq0104_4

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29. https://doi.org/10.1177/0265532209359514

Goodier, T. (2018, June). The CEFR Companion Volume launch conference, May 2018 – Tim Goodier. Eaquals. https://www.eaquals.org/eaquals-blog/the-cefr-companion-volume-launch-conference-may-2018-tim-goodier/

Group of Eight (Go8) (2018, Sep 8). *Go8 Quality Verification System: Assessment Review Guidelines*. The Group of Eight, Canberra.

Harsch, C., Collada Peña, I. D. L. C., Gutiérrez Baffil, T., Castro Álvarez, P. & García Fernández, I., (2020). Interpretation of the CEFR Companion Volume for developing rating scales in Cuban higher education. *CEFR Journal - Research and Practice*, *3*, 86-98. https://cefrjapan.net/publications/journal

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly, 12*(4), 333-362. https://doi.org/10.1080/15434303.2015.1092545

Harsch, C., & Kanistra, V. P. (2020). Using an innovative standard-setting approach to align integrated and independent writing tasks to the CEFR. *Language Assessment Quarterly*, *17*(3), 262-281. https://doi.org/10.1080/15434303.2020.1754828

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., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test centered approach. *Language Assessment Quarterly*, *8*(1), 1-33. https://doi.org/10.1080/15434303.2010.535575

Harsch, C., & Seyferth, S. (2020). Marrying achievement with proficiency–Developing and validating a local CEFR-based writing checklist. *Assessing Writing*, *43*, 100433. https://doi.org/10.1016/j.asw.2019.100433

Humphreys, P., Haugh, M., Fenton-Smith, B., Lobo, A., Michael, R., & Walkinshaw, I. (2012). Tracking international students' English proficiency over the first semester of undergraduate study. *IELTS research reports online series*, 41. https://www.ielts.org/for-researchers/research-reports/online-series-2012-1

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing, 38*(4), 602-626. https://doi.org/10.1177/0265532221994052

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, *19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

McCubbin, A., Hammer, S., & Ayriss, P. (2021). Learning and teaching benchmarking in Australian universities: the current state of play. *Journal of Higher Education Policy and Management*, 1-18. https://doi.org/10.1080/1360080X.2021.1934244

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica: Biochemia Medica*, *22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

McNamara, T., Morton, J., Storch, N., & Thompson, C. (2018). Students' Accounts of Their First-Year Undergraduate Academic Writing Experience: Implications for the Use of the CEFR. *Language Assessment Quarterly, 15*(1), 16-28. https://doi.org/10.1080/15434303.2017.1405420

Morgan, C., & Taylor, J. A. (2012). Benchmarking as a catalyst for institutional change in student assessment. In K. Coleman & A. Flood (Eds.), *Marking time: leading and managing the development of assessment in higher education* (pp.25-39). Common Ground Publishing.

Normand-Marconnet, N., & Bianco, J. L. (2015). The Common European Framework of Reference down under: A survey of its use and non-use in Australian universities. *Language Learning in Higher Education, 5(*2), 281-307. https://doi.org/10.1515/cercles-2015-0014

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching*, *47*(2), 228-242. https://doi.org/10.1017/S0261444811000206

Nguyen, V. H., & Hamid, M. O. (2021). The CEFR as a national language policy in Vietnam: Insights from a sociogenetic analysis. *Journal of Multilingual and Multicultural Development*, *42*(7), 650-662. https://doi.org/10.1080/01434632.2020.1715416

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

Peer Review Portal (2019, Jul 12). https://www.peerreviewportal.com

Plenter-Vowles, K. (2018, Jan 27). The CEFR Companion Volume with New Descriptors: Uses and Implications for Language Testing and Assessment. VIth EALTA CEFR SIG, Dublin, Ireland. https://www.researchgate.net/publication/330741857_The_CEFR_Companion_Volume_with_New_Descriptors_Uses_and_Implications_for_Language_Testing_and_Assessment_VIth_EALTA_CEFR_SIG

Read, J. (2019). The influence of the Common European Framework of Reference (CEFR) in the Asia-Pacific region. *LEARN Journal: Language Education and Acquisition Research Network*, *12*(1), 12-18. https://files.eric.ed.gov/fulltext/EJ1225686.pdf

Roche, T. B. (2017). Assessing the role of digital literacy in English for Academic Purposes university pathway programs. *Journal of Academic Language and Learning, 11*(1), A71-A87. https://journal.aall.org.au/index.php/jall/article/view/439

Roche, T. & Booth, S (2019). *External referencing of ELICOS direct entry program standards: UECA National Report 2019*, Southern Cross University. https://ueca.edu.au/initiatives/

Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education: Principles, Policy & Practice*, *14*(3), 387-392. https://doi.org/10.1080/09695940701592097

Sadler, D. R. (2013). Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principles, Policy & Practice, 20*(1), 5-19. https://doi.org/10.1080/0969594X.2012.714742

Sankey, M., & Padró, F. F. (2016). ACODE Benchmarks for technology enhanced learning (TEL): Findings from a 24 university benchmarking exercise regarding the benchmarks' fitness for purpose. *International journal of quality and service sciences*. https://www.acode.edu.au/pluginfile.php/550/mod_resource/content/8/TEL_Benchmarks.pdf

Sefcik, L., Bedford, S., Czech, P., Smith, J., & Yorke, J. (2018). Embedding external referencing of standards into higher education: collaborative relationships are the key. *Assessment & Evaluation in Higher Education, 43*(1), 45-57. https://doi.org/10.1080/02602938.2017.1278584

Syme, S., Davis, C., & Cook, C. (2021). Benchmarking Australian enabling programmes: assuring quality, comparability and transparency. *Assessment & Evaluation in Higher Education*, *46*(4) 572-585. https://doi.org/10.1080/02602938.2020.1804825

Tertiary Education Quality and Standards Agency (TEQSA). (2015, June 17). *Higher Education Standards Framework (Threshold standards) 2021: Contextual overview of the HES Framework 2021*. https://www.teqsa.gov.au/contextual-overview-hes-framework

Tertiary Education Quality and Standards Agency (TEQSA). (2019a, June 5) Guidance Note: ELICOS Direct Entry Version 2.0. https://www.teqsa.gov.au/latest-news/publications/guidance-note-elicos-direct-entry

Tertiary Education Quality and Standards Agency (TEQSA). (2019b, June 25) Guidance Note: External Referencing (including Benchmarking). https://www.teqsa.gov.au/sites/default/files/guidance-note-external-referencing-v2-5-web.pdf?v=1581308237

Watty, K., Freeman, M., Howieson, B., Hancock, P., O'Connell, B., De Lange, P., & Abraham, A. (2014). Social moderation, assessment and assuring standards for accounting graduates. *Assessment & Evaluation in Higher Education, 39*(4), 461-478. https://doi.org/10.1080/02602938.2013.848336

Wudthayagorn, J. (2021). An Exploration of the English Exit Examination Policy in Thai Public Universities, *Language Assessment Quarterly*. https://doi.org/10.1080/15434303.2021.1937174

Appendix A.

Twenty participating University English Centres Australia (UECA) Institutions

Institution	Writing Assessment Tasks Reviewed
Australian Catholic University	Research Essay
Australian National University	Body Paragraph (incl. Essay Plan) and Final Essay
Central Queensland University	Research Essay and Research Report
Curtin University	Critical Response and Research Report
Flinders University	Critical Review
The University of Melbourne	Critical Response Written Exam, Research Essay, Critical Response Written Exam
Monash University	Summary and Report Writing tasks
Queensland University of Technology	Mid-program Timed Essay, Final Timed Essay, Research Assignment: Report or Essay
Royal Melbourne Institute of Technology	Referenced Essay and Critical Response
Southern Cross University	Exit Writing Exam and Final Report
Swinburne University of Technology	Writing assessment (under exam conditions)
The University of Adelaide	Final Integrated Reading and Writing Task (under exam conditions); Final Exam Essay
University of Wollongong	Final Essay, Exam
University of New England	Research Project and Essay
University of Newcastle	Writing Assessment
University of Sydney	Comparative Summary and Critical Response Essay
University of Tasmania	Final Short Essay Writing Exam
University of Western Australia	Final writing exam
Victoria University	Mid-Program and Writing Exam
Western Sydney University	Research Report and Research Essay

Acknowledgements

Not applicable.

Funding

This work was supported by University English Centres Australia's (UECA) External Referencing of the ELICOS Standards [ERES] Project funding.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access