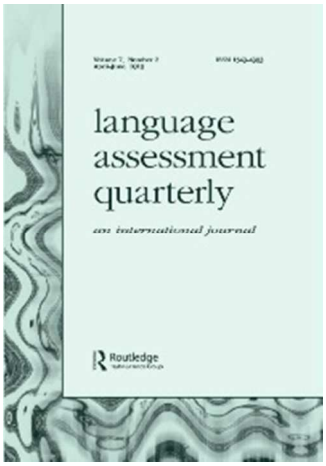

Linking tests of English for Academic Purposes to the CEFR: The score user's perspective

Journal:	*Language Assessment Quarterly: An International Journal*
Manuscript ID	HLAQ-2016-0015.R2
Manuscript Type:	Article
Keywords:	Common European Framework of Reference, English for Academic Purposes, Language policy, Standard setting, University entrance

Linking tests of English for Academic Purposes to the CEFR: The score user's perspective

1. Introduction

The CEFR (Council of Europe – CoE – 2001) has been widely adopted by policy makers in setting language proficiency requirements (see Deygers et al. this volume). As one example of this, to qualify for a Tier 4 (General) student visa to study at degree level, applicants to UK universities from outside the European Economic Area must demonstrate knowledge of English equivalent to at least CEFR level B2 (although universities often choose to set more stringent conditions). In circumstances like these, the stakes for language learners and for assessment are high. Academically qualified students will not be accepted onto degree programmes unless they meet language requirements. However, the destination university determines what constitutes evidence that a student has demonstrated B2 proficiency. This need not be restricted to test scores, or to certificates awarded by a specific accrediting agency. Such uses of the CEFR raise many questions, including how well the framework serves to inform test score users such as university admissions staff, university managers responsible for visa compliance and marketing, officials with responsibility for immigration, test preparation institutes, teachers and students about the quality of the evidence provided to them.

In order to show that their qualifications can be used to establish a learner's CEFR level, assessment agencies have come under pressure to 'link' them to the framework. The CoE has provided extensive guidance on setting CEFR-related standards (CoE 2009). The process is intended to make "the results of the examination in question more transparent to both the users of examination results and to test takers themselves" (CoE 2009, 26). This study explores how four agencies offering tests of English for academic purposes (EAP) have interpreted the CoE guidance

and considers whether relating these tests to the CEFR has provided for users the intended transparency of examination content and level.

2. Literature Review

The objective of standard setting, as Reckase (2009) elegantly expressed it, involves the "translation of policy to numerical cut score" (p.17). The task of a standard setting panel is to work collectively to recommend the score on a test that best reflects policy intentions. A distinction is commonly made between the *content standard*, which sets out what is to be tested, and the *performance standard*, which is expressed as a numerical cut score on the relevant test. In relation to a test of EAP used in admissions, a university's policy might be to admit students who have sufficient language ability to allow them to participate successfully in their courses. The standard setting process involves.

(1) Elaborating the language abilities needed to support academic study: the content standard.

(2) Determining the score (or constellation of scores) on a given test that best corresponds to the level of ability required for success: the performance standard.

Setting performance standards thus depends crucially on the extent to which tests reflect appropriate content standards. This paper considers the role of the CEFR in communicating both content standards (Section 2.1) and performance standards (Section 2.2) to test score users.

2.1. Linking on the horizontal dimension: content comparisons

The approach to linking advocated in the CoE (2009) manual involves three stages. These are *specification*, *standard setting* (which refers in the manual to setting performance standards), and *validation* of the claimed link. Each stage is preceded by familiarisation, which requires participants in the linking procedure to build a detailed knowledge of the CEFR.

The specification stage addresses *what* is tested: content standards. It is concerned with how test content relates to the qualitative descriptive scheme or 'horizontal dimension' of the framework. This includes the question of how test tasks relate to social activities in the target language use domain to which test results are intended to generalise. According to the action-oriented perspective adopted by the CoE, the linguistic resources required to understand textbooks and produce essays in the educational domain are distinct in important ways from those needed to, for example, engage with sacred texts and participate in religious services ('public domain') or to interpret recipes and cook meals ('personal domain'). Results from a test that addresses one context may not readily generalise to another.

Responding to criticism that the CEFR lacks the level of detail required to build test specifications (see for example Weir 2005), North (2014) and the other CEFR authors stressed that the framework is not itself a content standard, but a generative "apparatus to develop a differentiated standard appropriate to the context" (p.62). The CEFR offers guidance for building contextually relevant standards: what [*specified*] learners have to learn to do in order to use a [*specified*] language (or languages) for [*specified forms of*] communication and what knowledge and skills they have to develop so as to be able to act effectively [*in specified contexts*].

Contrary to Kaftandjieva (2007), who believed that limitations in the coverage of the 53 illustrative scales by a test would invalidate any claim of linkage for that test, North (2014) emphasised that tests might draw on the framework selectively. Some of the elements to be specified might be found in the CEFR, but users should add others according to need. In relation to languages for academic purposes, he listed eight CEFR scales as relevant to academic language use, but suggested these could provide no more than a point of departure for specifying the content domain of academic language use. He stressed that because, "two examinations may both be "at

Level B2" and yet differ considerably" (CoE 2009, p.4), the process of linking to the CEFR should result in a profile of a test that captures differences and conveys them to users of the results.

To assist in this profiling effort, categories of communicative language activity and communicative language competence are elaborated in CEFR Chapters 4 and 5 (CoE 2001, pp.43-130) and variations in social context are addressed through the taxonomy of situational features in Table 5 (pp.48-9). The CoE (2009) manual presents these as a series of checklists (CoE 2009, Appendix A Forms A9 to A22). *Content Analysis Grids* are also provided (CoE 2009, Appendix B), supplementing the CEFR with aspects of test task specification found to be lacking in the framework itself (Alderson et al. 2004). Additionally, users are invited to complement the CoE categories with their own.

In describing the development of the CoE (2009) manual, Figueras et al. (2005, p.270) recognised that comprehensive profiling on the basis of content would be of value to testing agencies for internal purposes, but might overwhelm score users with information. The CoE (2009) manual therefore includes simplified presentation versions of the Content Analysis Grids to, "provide the basis for good documentation and examination user guides" (CoE 2009, p.30). Another means of reducing the complexity for score users is a summarising *graphical profile* (CoE 2009, Form A23 p.153), which locates the relationship between test and framework on two dimensions. In the example of a completed profile offered by Figueras et al. (2005, p.275), *Sociolinguistic Competence* is tested at A2 while *Reading* is tested at B2.2.

Coste (2007), a CEFR author, pointed to a trade-off between the greater convenience of generic level descriptions (a B2 level learner) and the greater precision, but more limited generalizability of scales focussed on specific activities (a learner who is judged to be B2 in *Goal-oriented co-operation*, but B1 in *Addressing audiences*). He expressed concern that score users may not have much appetite for profiling, or understand the restrictions on the interpretation of the CEFR global

scale. Instead, he suggested, they might prefer to regard it as, "a measuring instrument which can define proficiency levels exactly, calibrating them as precisely as the graduations on a medical thermometer" with, "global, summary labels, signifying that an A2 or B2 learner has attained that level, across the board, in all the skills to which descriptors are attached" (p.39). In other words, unless given clear guidance, users may over-interpret a B2 test score as indicating ability in *all* aspects of the CEFR descriptive scheme rather than just the restricted subset of abilities actually addressed by the test. As Fulcher (2010) observed, this tendency threatens to displace the proper function of test validation. Instead of building a validity argument justifying the use of a test for a specific purpose (such as demonstrating that a potential student will be able to cope with the linguistic demands of university study), there is a risk that a testing agency need only show that a test is linked to the CEFR to persuade users that it is suitable for almost any purpose.

Since its publication, numerous studies have related tests to the CEFR, often making use of the CoE (2009) manual. Specification has proved challenging for these linking studies, including for all of those who reported on this aspect when trialling the draft manual (CoE 2003). Difficulties arose from vagueness and inconsistency of terminology; partial or ambiguous matches between test content and descriptors; and omissions: features covered in a test, but not captured by the CEFR (Khalifa, Salamoura & ffrench 2010, Noijons & Kuiper 2010, Papageorgiou 2010, Wu & Wu 2010). Although several of the CEFR linking reports suggested that specification is a valuable exercise that helps agencies to reflect on the suitability of test content, it appears that the outcomes, even in the simplified form of summaries and graphical profiles, may have limited value for test score users as a means of comparing and differentiating tests. This study considers how the horizontal dimension of the framework has been used by testing agencies working in an EAP context in communicating test content to users.

2.2. Linking on the vertical dimension: level comparisons

Calibration of the illustrative 'Can Do' descriptors (detailed in North 2000), based on teacher judgement and the rating of performance samples, addresses the question of *how well* a learner performs on tested content: the CEFR 'vertical dimension' of proficiency levels ranging from A1 (Breakthrough) to C2 (Mastery). The CoE (2009) manual offers a number of alternative approaches to setting performance standards and attention has centred on the relative benefits of these for determining cut scores.

The fact that different methods can yield very different outcomes is clearly especially problematic in a situation where multiple cut scores (scores representing B1, B2, C1) and multiple tests are involved (Reckase 2009). Adding to the complexity, different organisations are involved in linking to the framework tests of different languages designed for different purposes and populations. To arrive at defensible links, Kaftandjieva (2004) recommended the use of multiple methods, judiciously weighed. Milanovic (2009) proposed that the relationship between test and framework, "should remain tentative and be modified… in light of additional evidence when it becomes available" (p.4). Figueras (2009), concerned that testing agencies have a vested interest in claiming links to the CEFR on questionable evidence, argued that an independent body should adjudicate on the validity of all such claims. The European Association for Language Testing and Assessment (EALTA) (2006) sets out ten requirements for defensible linking including, of particular relevance to this study, "3. Have the test content and the test specifications been analysed in relation to the CEFR descriptors?" and "8. What standard setting procedures have been used to establish cut-off scores for the relevant CEFR level(s)?"

Both critics and defenders of the CEFR have been concerned that score users may interpret results from tests that have been linked to the framework as interchangeable: that, "a score of 'X' on a UK test is equivalent in meaning to a score of "Y" on a US test, and "Z" on an EU test" (Fulcher

2004, p.260). Although such assumptions of equivalence may be convenient and attractive to users (and are implicit in many policy uses of the framework), they are misguided. If scores on two or more tests are to be equated (i.e. if results are to be treated as interchangeable), the tests must be "as similar as possible in content and statistical characteristics" (Kolen & Brennan 2014, p.3). The multidimensional and contested nature of language abilities means that it cannot be assumed that any two tests measure the same construct. Even when developed for similar populations and purposes, tests may involve different content and formats. In relation to their statistical properties, differences in the distribution of scores and in the relationships between components (as between subtests of reading and of listening abilities) also serve to reduce their comparability (Figueras et al. 2005). It must also be considered that tests can differ greatly in technical quality: scores on one may simply be more reliable than scores on another. This study considers whether the CEFR helps score users to compare performance levels across tests.

2.3. Validation of the linking process

The final step in the suggested linking process is empirical validation, which involves corroboration of hypothesised links through evidence from alternative sources, such as teacher ratings or scores on other tests. This study explores the relationships between a number of tests of EAP that have been linked independently to the CEFR.

For such links to be considered valid, they must also be clearly understood by test score users. The CEFR is intended to promote transparency and aid communication. Yet, while there have been many studies relating tests to the framework and reflecting on the process, very little research attention has been given to the presentation of outcomes to test score users. This study investigates how testing agencies working in the relatively self-contained field of English language tests for

access to Higher Education Institutions (HEIs) have chosen to define and communicate the content and level of four widely used international tests of EAP in relation to the CEFR.

Three research questions are addressed:

1. With respect to the horizontal dimension of the CEFR, to what extent do the testing agencies use the CEFR to convey test content to users?
2. In connection with the vertical dimension of the CEFR, what claims do the testing agencies make concerning relationships between test scores and CEFR levels?
3. How far do the independent linking efforts corroborate or validate each other?

Finally, I consider the extent to which the claims made about links between these tests and the CEFR might support users in making judgements about test use for university admissions purposes.

3. Methods

3.1. Selection of tests

To evaluate the CEFR as a means for expressing standards in a common metalanguage understood across diverse settings, this study investigated how testing agencies working in the field of EAP used the framework when representing their tests to score users. The tests to be included in the study were selected on the basis that they were:

1. Widely used by HEIs in destination countries as evidence for the English language proficiency of international students entering degree programmes.
2. Accessible to a global test taking population.

Twenty-four institutions were identified for inclusion in the study. Figures for 2012 from the Organisation for Economic Co-operation and Development (OECD) (2014) show four English speaking countries – the USA, the UK, Australia and Canada – accounting for approximately 40% of global international student enrolments. Five HEIs featuring in the QS World University

Rankings 2015-2016 were selected from each of these four countries. One HEI was chosen from each of four countries identified by the OECD where English is not the national language, but where English medium courses are widely available for international students: Denmark, Finland, Sweden and the Netherlands. The websites of the selected HEIs were consulted to establish which tests are accepted as evidence that students meet English language requirements for direct entry to academic degree programmes.

Two tests were recognised by all 24 HEIs: the International English Language Testing System: Academic (IELTS), provided by the IELTS partnership (the British Council, IDP: IELTS Australia and Cambridge English Language Assessment: CELA) and the TOEFL iBT test, provided by the Educational Testing Service (ETS). The Pearson Test of English, Academic (PTE-A), provided by Pearson Language Assessment, was accepted by 16. Cambridge English: Advanced (CAE), provided by CELA, was accepted by 17 (none in the USA and three in Canada). Other tests were recognised by institutions in one or two of the countries surveyed. For example, the Canadian Academic English Language Assessment (CAEL) was generally accepted in Canada, but not elsewhere. Other evidence for an acceptable level of English language ability recognized by some HEls included results from national qualifications such as the German *Abitur* or French *Baccalaureat*, successful completion of an extended period of English medium education or local preparatory courses such as the Kings College, London Foundation programme or Murdoch Institute of Technology EAP course. However, only four tests: IELTS, TOEFL iBT, PTE-A and CAE appeared to satisfy the criterion for widespread recognition.

All four tests are accessible to test takers worldwide (Table 1). TOEFL iBT offers 50 test dates per year and IELTS 48. The PTE-A website reports that the test was administered on 363 days in 2013 while CELA lists 38 dates in 2016 on which CAE will be administered in either a computer-based or paper-based format.

Table 1. Testing agencies' claims concerning the availability and recognition of four tests of EAP (October 2015)

Test	*Availability*		*Recognition*	
	Countries	Centres	Countries	Institutions
TOEFL iBT: www.ets.org/toefl	165	Over 4,500	Over 130	Over 9,000
IELTS: www.ielts.org	140	1,000	Over 145	Over 9,000
PTE-A: pearsonpte.com	Over 50	200	56	Over 1,700
CAE: www.cambridgeenglish.org	130	2,800	"Worldwide"	"Thousands"

Because of the high stakes associated with the results, it is essential that tests used in admissions should meet recognised standards for security and quality of measurement. The CoE (2009) manual makes the point that "a claim [of linking to the CEFR] makes little sense unless it is accompanied by evidence of good practice, internal validity and adequate quality procedures for all the steps of the test development and administration cycle" (p.38). With this imperative in mind, the following criteria were used in determining whether tests should be included. Publicly available evidence was sought for:

1. provision of multiple parallel forms of the test
2. security of test production, administration and results
3. reliability of test scores

All four agencies publish extensive information on the qualities of their testing systems, explaining for users how they meet the criteria for security and reliability. These include such features as field testing of test material, protocols for the production and treatment of test materials, fairness reviews, biometric ID checks, inspection and monitoring regimes, online results verification services, systems for detecting misconduct and support for validation studies. Indices published by

the testing agencies (Table 2) demonstrate that all four tests also achieve acceptable levels of reliability, satisfying criterion 3.

Table 2. Reliability estimates reported by four testing agencies

Test	Scale	α	SEM
TOEFL iBT	0-120	.94	5.64
IELTS	0 – 9	.96	0.23
PTE-A	53-79*	.97	2.32
CAE	0-100	.93	2.89

* Reliability estimates for PTE-A are reported only for scores in the range 53–79

3.2. Analysis of documents presenting claims of CEFR linking

The study involved the selection and analysis of documents from testing agencies that presented their claims regarding CEFR alignment. Areas of specific interest for the study included information for score users concerning:

- The relationship between test content and the CEFR descriptive scheme.
- Scores claimed to correspond to the B2 level of the CEFR.
- Recommendations regarding appropriate score thresholds for acceptance onto university programmes.
- Claims of bilateral links made between scores on the tests concerned.
- Evidence cited in support of CEFR linking claims.

An initial web search with the terms "[*test name*] CEFR score," revealed that all four agencies publish either a web page or linked document, or both, claiming a relationship between test scores and the CEFR (see Appendix A). The CoE (2009) manual makes it clear that such claims should not

be taken on trust, but should be supported by "both theoretical and empirical evidence" (CoE 2009 p.7). Supporting evidence in the form of linked or referenced research papers providing more technical backing for the claimed relationships was therefore also included within the scope of the study.

4. Results and Discussion

4.1. Linking on the horizontal dimension: content comparisons

Concerning content coverage, analysis of user-oriented documents revealed a clear division between the two older tests of EAP – TOEFL, launched (in its original paper-based version) in 1964 and IELTS launched in 1989 – and their two more recent competitors – CAE, launched in 1991 and PTE-A, launched in 2009. Relatively little attention is given to connections between the CEFR and the content or design of either IELTS or TOEFL iBT. The links made are largely limited to vertical score: level correspondences. The developers of PTE-A and CAE, in contrast, emphasise the integral part played by the framework in test development and operational test production systems.

User-oriented brochures that present TOEFL and IELTS to HEIs (ETS 2011; IELTS Partners 2013) make no mention of the CEFR. Neither organisation includes references to the CEFR on score reports, although both do provide information online about the relationship between their tests and the framework (IELTS Partners 2015, ETS 2015). Tannenbaum and Wylie (2008) and Papageorgiou et al. (2015) are cited on the ETS website as the sources for the TOEFL iBT: CEFR comparisons. Tannenbaum and Wylie (2008) expressed concern about potential discrepancies between the definition of proficiency in the framework and the content of the test, but omitted a specification phase, arguing that TOEFL iBT coverage of reading, writing, listening and speaking made it sufficiently congruent with the CEFR to justify standard setting. Papageorgiou et al. (2015) added to the specification aspect, noting that, "without satisfactory content alignment… there is

little justification for conducting a standard-setting study" (p.9). The authors identified similarities between the TOEFL iBT Speaking descriptors at level 3 of the scoring rubric and B2 descriptors taken from CEFR scales for *Grammatical Accuracy, Vocabulary Control, Vocabulary Range* and *Spoken Fluency*. Lim et al. (2013) reported a similar standard setting study for IELTS. Like the ETS researchers, they found it difficult to reconcile the "multiplicity of contexts… languages, and… uses" (p.35) of the CEFR with standard setting approaches developed for application to defined contexts for test use.

It is notable that neither the TOEFL iBT nor the IELTS study employed the tools provided by the CoE (2009) to profile test content. Moreover, the IELTS website cautions against using CEFR levels as the basis for determining entry requirements. Instead, it advises receiving HEIs to, "look at the IELTS bandscore descriptors and use the IELTS Scores Guide DVD to ascertain the appropriate level of language ability required for their institution or course" (IELTS Partners 2015, online). The IELTS Partners (2013) suggest that a score of 6.5 is "probably acceptable" for "study on "linguistically less demanding academic courses" (i.e. Agriculture or Technology, but not Law or Medicine). ETS (2011) similarly refers users to a CD providing sample TOEFL iBT test taker responses at various levels together with advice on setting locally appropriate cut scores.

The user-oriented brochures for the other two tests – Pearson (2012) and CELA (2015)– both refer to the CEFR as a standard. Pearson (2012) describes it as "a standard across Europe and in many countries outside of Europe" (p.38) and CELA (2015) calls it "the international standard for describing language ability" (p.3). In a supporting document (Pearson 2010), it is stated that PTE-A is, "designed to measure language competence according to the principles of the CEF[R] and to address specifically language competencies in the range from upper B1 to lower C2" (p.3). CELA (2015) opens with the statement that CELA exams are aligned with the CEFR and that CAE is "targeted at Level C1," which "gives students the English skills needed for academic success" (p.3)

as well as for the workplace and international travel, reflecting a more general construct than for the other three tests.

Pearson (2010) explained how test content was based on the CEFR. The 144 "considerations" found in the CEFR (reflection boxes prefaced "users of the Framework may wish to consider and where appropriate state…") were used as a checklist to guide test design. Zheng and de Jong (2011) presented extracts demonstrating how the CEFR was embedded in the test specifications. Features of the guidelines for item writers such as "Content domain," "Discourse type" and "Topic" reflect categories found in the CoE (2009) specification forms. Training for item writers, item reviewers and human raters followed the stages of familiarisation, specification and standardisation that were emerging in the draft CoE (2003) manual. Item writers were asked to identify a CEFR level for each item submitted and their predictions were tested against empirical data from field-testing. Nonetheless, in the presentation of test content for users that emerged from this process, there is minimal reference to the framework.

Pearson (2012, pp.40-41) pairs Can Do statements from the CEFR global scale A1 to C1 (CoE 2001, p.33) with the PTE-A scale. The document outlines how the CEFR B2 statements relate to academic study: B2, it states, "was designed as the level required to participate independently in higher level language interaction" (Pearson 2012, p.40). However, PTE-A score reports do not refer to the CEFR levels, only reporting section and overall scores on the Pearson Global Scale of English (GSE). Pearson (2012) advises that receiving HEIs should determine entry requirements and reports that a score on the GSE in the range 51-61 is typically required for undergraduate studies, 57-67 for postgraduate studies and 59-69 for MBA studies.

No standard setting studies have been reported for CAE, but according to the CELA website (UCLES 2015a), "there is growing evidence to support the view that the Cambridge English exams embody or reflect the CEFR in a variety of ways" (www.cambridgeenglish.org/research-and-

validation/fitness-for-purpose/). This claim is based on three strands of evidence: historical, conceptual and empirical. The historical perspective notes CELA involvement in the development of the CEFR and the parallel ALTE Can Do project described in Appendix D of the CEFR (p.253ff.). The CEFR (CoE 2001) draws a connection between the CAE assessment criteria and its own illustrative scales (p.194). Material from CELA tests is offered by the CoE (2011) to exemplify CEFR levels for standard setting purposes with CAE exemplifying C1. This implies that item banking methods used to maintain consistent standards across test forms also embed the CoE levels. Similarly to PTE-A, manual alignment procedures are said to be incorporated into routine test operations (Khalifa & ffrench 2009).

CAE *Statements of Results* report a CEFR level together with scores on the Cambridge English Scale and a grade. CELA also includes CEFR levels in its recommendations on entry requirements, suggesting that C1 (identified with a passing grade on CAE) is suitable for direct undergraduate and postgraduate entry to English medium HEIs, with B2 (represented by Cambridge English: First) being suitable for access to studies below degree level (pre-sessional/ foundation intensive English courses). CELA (2015) provides Can Do descriptors, drawn from the CEFR-related ALTE Can Do project, to aid interpretation. Although some CAE tasks have been profiled against CEFR categories to illustrate the levels of the framework (CoE 2005), CELA (2015) does not present them in this way. The focus of each test part is explained – for example, in Part 2 of the Listening test, test takers must demonstrate that they have "understood detail and can identify specific information within the recording" – but it is not made explicit for the user how each test section or task relates to CEFR categories.

Although both PTE-A and CAE claim a more intimate relationship with the framework than IELTS or TOEFL iBT, the test providers make little use of the CoE (2009) profiling tools or metalanguage to communicate their content to users. The links to the CEFR are presented in global

terms, not in the form of a differentiated profile. In this context at least, the CEFR is not used as the foundation for direct, transparent comparisons between the content standards informing different tests. Users wishing to make such comparisons would need to look elsewhere, or to recover information presented in different terms by the four agencies and make the comparisons for themselves.

4.2. Linking on the vertical dimension: level comparisons

Each of the testing agencies has chosen to present the vertical alignment with the CEFR in a rather different way (Table 3). These differences could be attributed to differences between the agencies in their approaches to measurement.

Table 3. Vertical links claimed between tests of EAP and the CEFR

CEFR level	TOEFL iBT (0–120) *	PTE-A †	CAE Certificate	Cambridge English Scale ‡
C2	-	80 – **85** – 95	Grade A	200–210
C1	24 – 22 – 25 – 24 – **95**	67 – **76** – 84	Grade B/C	180–199
B2	18 – 17 – 20 – 17 – **72**	51 – **59** – 75	**Level B2**	160–179

* Figures represent scores for *Reading* (0–30)-*Listening* (0–30)-*Speaking* (0–30)-*Writing* (0–30)-*Total* (0–120) (bold)

† Figures represent score levels indicating the ability to perform the *Easiest*-*Average* (bold)-*Most Difficult* tasks at CEFR B2 level

‡ The Cambridge English Scale is a range of scores used to report results on CELA exams targeting

different levels of the CEFR. Cambridge English Scale scores for CAE range from 142 to 210

Sources: UCLES (2015b), ETS (2015), Pearson Education (2012)

Separate alignments are suggested for TOEFL iBT scores for each of the four tested skills in addition to the total score. Adjustments have been made to the score comparisons presented. Until 2014, the suggested cut scores for iBT TOEFL, based on Tannenbaum and Wylie (2008), were 87 for B2 and 110 for C1. The current cut scores of 72 and 95 thus represent a shift of 15 points downwards (twice the standard error of measurement) on the TOEFL scale. In discussing this change, Papageorgiou et al. (2015) noted that B2 is sometimes set (as by UK Visas and Immigration - UKVI) as the minimum level required for entry to university. However, locally defined iBT TOEFL score entry requirements in the USA, Canada, UK and Australia are generally lower than the 87 points recommended in Tannenbaum and Wylie (2008) as the B2 cut score. Thus students might have satisfied a university's requirements for entry, but be denied a visa because their TOEFL iBT score located them below the B2 threshold. The TOEFL iBT: CEFR cut score recommendations were changed to "better capture the current practice in university entry requirements among the surveyed institutions" (Papageorgiou et al. 2015, p.12).

The IELTS website (IELTS Partners 2015) explains that scores are reported on a nine-band scale (with half-band increments), rather than in the form of scaled scores as is the case for the other three tests. The boundaries between bands do not correspond directly to CEFR levels (as embodied by CAE and other CELA tests). Rather than associating IELTS scores with a CEFR level, the relationship is represented graphically, showing IELTS bands and CEFR levels as overlapping ranges on a visual scale. Current information from the IELTS partners suggests that B2 (160 to 179 on the Cambridge scale) ranges from the upper portion of IELTS band 5.0 (which is identified as 154 to 162 on the Cambridge scale) to the lower portion of the 6.5 range (176 to 185). As with TOEFL iBT, the IELTS website refers to a developing understanding of the IELTS-CEFR

relationship informed by research. In this case, the shift was upwards: the threshold between B2 and C1 moved above IELTS 6.5.

PTE-A is the only one of the four that defines for users what it means to be 'at' a level in terms of a probability of success in relation to Can Do descriptors. Pearson (2012) suggests scores that correspond to the ability to carry out relatively *Easy*, *Average* and *Difficult* tasks at each CEFR level. This style of reporting relates to the calibration of descriptors in the development of the CEFR illustrative scales. The B2 level includes descriptors that were judged to be relatively easy (e.g. *Can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options*) and others that were judged to be difficult, close to the C1 boundary (e.g. *Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech*).

Unlike the flexibility of IELTS and TOEFL iBT, the fixed relationship between CAE grade C and C1, would seem to leave little scope for modifying that relationship in the light of experience in the way suggested by Milanovic (2009). CELA (2015) explained that (unlike the other three tests under consideration) CAE forms part of a linked 'suite' of examinations, each targeting a different CEFR level. Results are reported both as a grade and, from 2015, as a score on the Cambridge English Scale, which connects CAE to the other examinations in the suite. CAE is targeted at C1 so that a Pass or 'C' grade on CAE (a score in the range 180 to 192 on the Cambridge English Scale) is also taken to represent a C1 level of ability. Test takers who score just below C on CAE (160 to 179) receive a certificate stating they demonstrated ability at Level B2 (the level targeted by Cambridge English: First - FCE).

The differences in the presentation of CEFR links shown in Table 3 reflect the variation in the approaches to linking adopted by the four agencies. Such variation reflects the non-directive, "open, dynamic and non-dogmatic" (CoE 2001, p.18) nature of the CEFR. However, when taken together

with the recent changes suggested in score: CEFR correspondences, it may result in a somewhat confusing picture for test score users.

4.3. Aligning scores between tests: Empirical validation?

In addition to relating test scores to the CEFR, the four testing agencies also offer evidence of how scores on their tests align with those on one or more of their competitors' tests. Following the stages of the CoE (2009) manual, this could serve as external validity evidence to corroborate the claims of linking to the framework.

The evidence presented by ETS as backing for the claimed relationship between TOEFL and IELTS comes from a study (ETS 2011), which employed an equipercentile linking method. This involved identifying the IELTS and TOEFL iBT scores that would screen out a similar percentage of the test takers on both measures. 1,153 test takers participated and were asked to report their scores on IELTS or TOEFL iBT. A moderately strong correlation of 0.73 was found between the scores on the two tests (see Table 4).

Table 4. TOEFL iBT test takers reporting scores on both TOEFL iBT and IELTS (n=1,153)

	IELTS mean	TOEFL iBT mean	Correlation
Overall	6.6	83.6	0.73
Listening	6.8	20.9	0.63
Speaking	6.3	20.0	0.57
Reading	6.8	21.2	0.68
Writing	6.1	21.6	0.44

Source: Linking TOEFL iBT ™ Scores to IELTS® Scores – A Research Report. (Educational Testing Service 2011).

While field-testing PTE-A, Pearson (2012) asked test takers to report their scores on IELTS or TOEFL iBT. Of the 10,000 test takers involved, over 2,400 had taken IELTS and 140 had taken TOEFL iBT. A moderately strong correlation was found between PTE-A and both IELTS (r=.76) and TOEFL iBT (r=.75) (see Table 5).

Table 5. PTE-Academic field test participants reporting scores on alternative tests

	Test takers	**Mean score**	**Correlation with PTE-A***
TOEFL iBT	140	92.9	0.75
IELTS	2432	6.49	0.76

* based on self-reported scores

Source: Pearson (2012 p.44)

Although cautioning that predicting scores on one test based on scores on another is problematic, Pearson (2012) used the results to draw up concordance tables relating PTE-A to IELTS and TOEFL iBT (Table 6). These tables were reported to take account of the "less than optimal effort of test takers during field testing where test results have no direct relevance to the test takers" (Pearson 2012, p.50) (a recognised threat to the validity of such exercises). However, no details were given of how these adjustments were made.

Table 6. Pearson (2012) Score concordance table for PTE Academic, TOEFL iBT and IELTS

PTE-A	IELTS	TOEFL	PTE-A: CEFR
42	5.5	54	
50	6.0	74	
51		76	Easy B2 tasks
58	6.5	85	

59		87	
65	7.0	95	Average B2 tasks
67		98	Easy C1 tasks
73	7.5	106	
75		109	Difficult B2 tasks
76		110	Average C1 tasks
79	8.0	114	
80		115	Easy C2 tasks
83	8.5	119	
84		120	Difficult C1 tasks
85			Average C2 tasks
86	9.0		
95			Difficult C2 tasks

Source: Pearson (2012, p.49)

As a member of the IELTS partnership, CELA has access to data for both IELTS and CAE. They presented the case for the relationship between the two tests in two documents available through the CELA website: University of Cambridge Local Examinations Syndicate (UCLES) (2011) and UCLES (2013). In the latter it is claimed that the link between IELTS and CAE is based on three sources: item banking for Reading and Listening papers; Common Scales for Writing and Speaking papers; and score comparisons.

Item banking refers to the practice of trialling test material to establish its measurement characteristics before including it in operational tests. UCLES (2013) reported that Reading and Listening items for Cambridge exams (including IELTS, CAE and CPE) are all located on a common measurement scale before being assigned to a suitable test. The Reading and Listening

papers of CAE and CPE are thus said to be linked to IELTS through the process of test construction. The second source, the Common Scales for Writing and Speaking, refers to the comparability of mark schemes employed in the different tests. Rating scales employed in both IELTS and CAE draw on the same qualitative analyses of the features of test taker performance at different levels. Unfortunately, CELA (UCLES 2013) does not provide evidence showing how this common provenance affects the comparability of standards. Reflecting the empirical evidential source, Lim et al. (2013), invited registered IELTS test takers to take CAE and vice versa. In all, 186 test takers from 24 countries participated. A correlation of r=.87 was reported between scores on IELTS and CAE. Table 7 presents the concordance between CAE and IELTS.

Table 7. Concordance between CAE scores and grades and overall IELTS scores.

IELTS Score	CAE Score	CAE Grade
8.0	80	A
7.5	74	B
7.0	67	C
6.5	58	
6.0	52	

Source: Comparing scores on Cambridge English: Advanced (CAE) and IELTS (UCLES 2013)

It is unclear from the CoE (2009) manual how closely scores on different tests would need to correlate and how nearly CEFR-linked cut scores would need to align to support a claim of mutual validation. Nonetheless, the evidence from this study indicates that scores from the different tests are far from interchangeable and that the testing agencies interpret the CEFR levels in markedly different ways. There is a high risk that a learner shown just to meet the requirement for CEFR B2 on one test would receive a different CEFR level on another test.

5. Conclusions

In the context of EAP testing for admissions, the CEFR seems to have provided a useful tool for test development and reflection within testing agencies. It also provides test score users with a broad, if approximate, sense of the relative level of test scores and of entry requirements. However, test score users should also be aware that tests may be located at the same CEFR level, yet assess different abilities; one agency's B2 may overlap with another's A2; and all links are provisional and subject to change. Test score users should be clearly warned not to rely on CEFR level correspondences as a basis for high stakes decision making.

A principled approach to setting entry requirements involves weighing the attributes of the applicants, the demands of the course and the support available for students. Following established standard setting practice (APA, AERA & NCME 2014), those responsible for HEI admissions should judge the value of tests as a sources of evidence to inform their decisions regardless of any claimed association with the CEFR.

In relation to the horizontal, content dimension, all four tests are offered for use in determining an adequate level of English for study purposes and all address four communicative skills. All four offer test score users comprehensive information, including descriptions of the test tasks, sample materials and scoring criteria which make it clear that the tests are structured differently, target a different range of levels of proficiency, include different task types, employ different approaches to scoring and have different measurement characteristics. The strength of the correlations reported between them does suggest that scores on one can be broadly predictive of scores on one or more of the others (at least within a certain range), but the correlations also suggest there are substantive differences in what is measured by each: 'B2' implies something rather different in each case.

These differences between the tests offer a proving ground for the CEFR as a profiling device. In a competitive marketplace, each agency strives to differentiate its test from those of its competitors

(see for example ETS 2009). The CEFR descriptive scheme and associated tools such as the Content Analysis Grids (CoE 2009, Appendix B) might seem an obvious means of demonstrating how tests take account of relevant contextual aspects of language use, facilitating comparisons between them. However, if the specification phase of linking does offer the opportunity to elaborate the theoretical rationale, aims and content coverage of an examination in a form that supports ready comparisons with competitors (CoE, 2009), it is an opportunity that the agencies spurn. Even for the two tests avowedly based on the CEFR, little use is made of its terms or reporting forms to explain test characteristics. Perhaps this reflects the distance between the context free generalisations of the framework and the contextually embedded nature of test tasks; perhaps the content analysis grids are simply too detailed for communicating to users, or perhaps each agency, reflecting a distinctive understanding of the construct, prefers to package information in its own way. In this context, the CEFR does not provide the kind of common, structured representation that would help test score users to judge the suitability of tests for admissions purposes and to make principled comparisons between them.

When it comes to the vertical dimension of proficiency levels, in spite of the adjustments made by the agencies, discrepancies between their recommendations on CEFR score correspondences persist. It is particularly interesting to observe how the UK immigration policy may have affected adjustments in the agencies' interpretation of the B2 level as a relatively liberal benchmark for university entrance. Current advice to HEIs from Pearson (2015) suggests that a score of 51 should be accepted as meeting the UKVI B2 requirement, although according to Pearson (2012), this score would predict failure on most B2 level tasks. The Pearson (2012) concordance table suggests that PTE-A 51 (easy B2 tasks) corresponds to 76 on TOEFL iBT. On the other hand, ETS currently identifies 72 on TOEFL iBT as the lower bound of B2. UKVI accept 5.5 on IELTS as evidence for a B2 level (Home Office 2015). This is identified by CELA (2015) as 162 on the Cambridge English

scale or two points above a B2 passing score on FCE. In the ETS score comparison table, IELTS 5.5 corresponds to just 46 on TOEFL iBT. On the Pearson concordance table it corresponds to 54 on TOEFL iBT and 42 on PTE-A, which Pearson (2012) suggests is below the threshold for B1. In short, one agency's B2 may be another's A2/B1: the outcomes of the different linking approaches do not support each other closely and do not provide convincing mutual validation.

One cause of the discrepancies is probably "the Achilles' heel of standard setting" (Kaftandjieva 2004, p.4): its arbitrariness and the tendency for different methods to bring different results. Another may be the under-specification of the CEFR levels with each agency arriving at its own interpretation. There have been shortcomings in the implementation of linking methodologies and the agencies have certainly not been slow to criticise the work of the others. De Jong (2009), for example, in presenting the Pearson research, questioned the basis for the IELTS: CEFR correspondences, while Lim et al. (2013) were critical of both Pearson and ETS. Perhaps, as suggested by North (2014), the methods advocated in the CoE manual are simply unsuitable for the purpose. On the other hand, neither Pearson nor CELA employed the standard setting methods criticised by North, but claim to embed a rich understanding of the CEFR in their testing operations. Nonetheless, Pearson and CELA reached mutually contradictory conclusions.

Findings suggest that more work is needed to help users to understand the implications and limitations of the CEFR as a tool for interpreting score outcomes. Priority should be given to the content and quality of assessment procedures rather than to the CEFR levels. The value of the framework for test developers, supported by the CoE (2009) manual, is not yet matched by its value as a tool for test score users.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, D.C.: American Educational Research Association.

Cambridge English Language Assessment (2015). Exams for Higher Education: Selecting students for foundation, undergraduate and postgraduate courses. Cambridge, UK: UCLES

Coste, D (2007). Contextualising uses of the Common European Framework of Reference for Languages. Paper presented at Council of Europe Policy Forum on use of the CEFR, Strasbourg, 8 February 2007.

Council of Europe (2003). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assesment (CEF). Manual:Preliminary Pilot Version*. DGIV/EDU/LANG 2003, 5. Strasbourg: Language Policy Division

Council of Europe (2005). *CD-ROM Relating Language Examinations to the CEFR: Reading and Listening Items and Tasks Pilot Samples.* Strasbourg: Language Policy Division

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual.* Strasbourg: Language Policy Division.

de Jong J.H.A.L. (2009). Unwarranted claims about CEF alignment of some international English language tests. Paper presented at the Sixth Annual Conference of EALTA, Turku, Finland 5th June, 2009

EALTA (2006). EALTA guidelines for good practice in language testing and assessment. Online document: Retrieved from www.ealta.eu.org/guidelines.htm, 6 September 2015.

Educational Testing Service (2011). *A Guide to the TOEFL® Test for Institutions.* Retrieved from http://www.ets.org/s/toefl/pdf/guide_toefl_test_institutions.pdf, 6 September 2015.

Educational Testing Service. (2009). *How to select the best academic English-language assessment.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2010). *Linking TOEFL iBT scores to IELTS scores—A research report*. Princeton, NJ: Educational Testing Service.

Educational Testing Service (2015). *Compare TOEFL® Scores.* Retrieved from www.ets.org/toefl/institutions/scores/compare 6 September 2015.

Figueras, N. (2009). Language Educational Policies Within a European Framework. In Alderson, J.C. (Ed.), *The Politics of Language Education: Individuals and institutions.* (pp. 203-221). Bristol: Multilingual Matters.

Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem: Cito/EALTA.

Figueras, N., North, B., Takala, S., Verhelst, N., and Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing*, *22*(3), 261-279.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, *1*(4), 253–266.

Fulcher, G. (2010). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. In Psyaltou-Joycey, A. & Matthaioudakis, M. (Eds.) *Advances in Research on Language Acquisition and Teaching* (pp. 15-26). Thessaloniki, Greece: GALA,

Galaczi, E D, ffrench, A, Hubbard, C, and Green, A.B. (2011). Developing assessment scales for large-scale speaking tests: a multiple method approach, *Assessment in Education: Principles, Policy and Practice*, *18*(3), 217-237.

Geranpayeh, A., & Taylor, L. (Eds) (2013). Examining listening: Research and practice in assessing second language listening, Cambridge, UK: Cambridge ESOL/Cambridge University Press.

Home Office (2015). Tier 4 of the Points Based System: Policy Guidance. Version 11/2015. London: HMSO.

IELTS Partners (2013). *IELTS Guide for educational institutions, governments, professional bodies and commercial organisations*. Retrieved 6 September 2015 from www.ielts.org/PDF/Guide_Edu- Inst_Gov_2013.pdf.

IELTS Partners (2015). *Common European Framework.* Retrieved 6 September 2015 from www.ielts.org/researchers/common_european_framework.aspx.

Kaftandjieva, F. (2004). Reference supplement to the preliminary pilot version of the manual for relating language examinations to the to the Common European Framework of Reference for Languages: learning, teaching, assessment. Section B: Standard Setting. Strasbourg: Language Policy Division.

Kaftandjieva, F. (2007). Quantifying the quality of linkage between language examinations and the CEF. In C. Carlsen & E. Moe (Eds.) *A Human Touch to Language Testing* Oslo: Novus Press, 33–43.

Khalifa, H, & ffrench, A. (2009). Aligning Cambridge ESOL examinations to the CEFR: issues and practice, *Research Notes 37*, 10–14.

Khalifa, H, Salamoura, A and ffrench, A (2010). Maintaining alignment to the CEFR: FCE case study. In Martyniuk, W. (ed.). *Aligning Tests with the CEFR. Reflections on using the Council of Europe's draft Manual.* Cambridge, UK: Cambridge University Press.

Kolen, M.J., and Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Lim, G.S. (2012) Developing and validating a mark scheme for writing, *Research Note*s 49, 6-10.

Lim, G.S., Geranpayeh, A., Khalifa, H. and Buckendahl, C.W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing 13*(1), 32-49.

Martyniuk, W (ed.) (2010). *Aligning Tests with the CEFR. Reflections on using the Council of Europe's draft Manual.* Cambridge, UK: Cambridge University Press.

McNamara, T., and Roever, C. (2006). Language testing: The social dimension. Oxford: Blackwell.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes*, 37, 2-5.

North, B. (2000). *The Development of a common framework scale of language proficiency.* New York, Peter Lang.

North, B. (2014). *The CEFR in Practice*. Cambridge, UK: Cambridge University Press.

OECD (2014) *Education at a Glance 2014: OECD indicators.* OECD Publishing. http://dx.doi.org/10.1787/eag-2014-en

Papageorgiou, S. (2010). Linking international examinations to the CEFR: The Trinity College London experience. In Martyniuk, W. (ed.). *Aligning Tests with the CEFR. Reflections on using the Council of Europe's draft Manual.* Cambridge, UK: Cambridge University Press.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., and Cho, Y. (2015). *The Association Between TOEFL iBT® Test Scores and the Common European Framework of Reference (CEFR) Levels* (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.

Pearson (2010). *Aligning PTE Academic test scores to the Common European Framework of Reference for languages. Research Note*. Retrieved 2 October 2015 from http://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf

Pearson (2012). *PTE Academic: Score Guide*. Retrieved from http://pearsonpte.com/wp-content/uploads/2014/07/PTEA_Score_Guide.pdf.

Pearson (2015). *UK Higher Education Institutions Information Pack Prepared June 2015*. Retrieved from pearsonpte.com/wp-content/uploads/2015/07/PTE-Academic-HEI-Brochure-June-2015.pdf

Reckase, M. D. (2009). Standard setting theory and practice: Issues and difficulties. In N. Figueras & J. Noijons (Eds.). *Linking to the CEFR levels: Research perspectives* (pp. 13–21). Arnhem: EALTA.

Saville, N (2005). An interview with John Trim at 80. *Language Assessment Quarterly 2*(4), 263–288.

Shaw, SD and Weir, CJ (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge, UK: Cambridge ESOL/Cambridge University Press.

Tannenbaum, R.J. and Wylie, E.C. (2008). *Linking English Language Test Scores onto the Common European Framework of Reference: An Application of Standard Setting Methodology.* ETS Research Report 08-34. Princeton, NJ: Educational Testing Service.

Taylor, L (2004a). Issues of test comparability, *Research Notes*, *15*, 2-5.

Taylor, L (2004b). IELTS, Cambridge ESOL examinations and the Common European Framework, *Research Notes*, *18*, 2-3.

Taylor, L (Ed) (2011). Examining speaking: Research and practice in assessing second language speaking, Cambridge, UK: Cambridge ESOL/Cambridge University

Taylor, L and Jones, N (2006). Cambridge ESOL exams and the Common European Framework of Reference for Languages (CEFR), *Research Notes, 24,* 2-5.

UCLES (2011). Methodology used for benchmarking Cambridge English: Advanced to IELTS. Retrieved from www.cambridgeenglish.org/images/28892-cae-comparison-methodology.pdf 6 September 2015

UCLES (2015a). International language standards explained. Retrieved from www.cambridgeenglish.org/cefr/ 6 September 2015

UCLES (2015b). Fitness for purpose. Retrieved from www.cambridgeenglish.org/cefr/ 6 September 2015.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281-300.

Wu, J., & Wu, R. (2010). Relating the GEPT reading comprehension tests to the CEFR. . In Martyniuk, W. (ed.). *Aligning Tests with the CEFR. Reflections on using the Council of Europe's draft Manual.* Cambridge, UK: Cambridge University Press.

Zheng, Y., & De Jong, J.H.A.L. (2011). Research Note: Establishing construct and concurrent validity of Pearson Test of English Academic. London: Pearson.

Appendix A

Documentary sources

Test	Web pages	Key score user documents	Research referenced on website
CAE	www.cambridgeenglish.org/cefr www.cambridgeenglish.org/research-and-validation/fitness-for-purpose	Cambridge English Exams for Higher Education (16 pages)	Taylor and Jones (2006) Saville (2005) Weir and Milanovic (Eds) (2003) Galaczi, ffrench, Hubbard and Green (2011) Hawkey, and Barker (2004) Jones (2000) Jones (2001) Jones(2002) Khalifa, ffrench and Salamoura (2010) Geranpayeh and Taylor (2013) Khalifa and Weir (2009) Lim (2012) Shaw and Weir (2007) Taylor (2011)

IELTS	.	IELTS Guide for Educational Institutions, Governments, Professional Bodies and Commercial Organisations (16 pages)	Hawkey, and Barker (2004) Lim et al. (2013) Milanovic (2009) Taylor (2004a) Taylor (2004b)
PTE-A	pearsonpte.com/institutions/scores	PTE Academic Score Guide (72 pages) UK Higher Education Institutions Information Pack (21 pages)	Pearson (2010) Pearson (2015)
TOEFL	www.ets.org/toefl/institutions/scores	A Guide to the TOEFL Test for Institutions (9 pages)	Tannenbaum and Wylie (2008) Papageourgiou, S., Tannenbaum, R., Bridgeman, B., Cho, Y. (2015)