Hidri *Language Testing in Asia* (2021) 11:9
https://doi.org/10.1186/s40468-021-00123-8

Language Testing in Asia

RESEARCH Open Access

Linking the International English Language Competency Assessment suite of examinations to the Common European Framework of Reference

Sahbi Hidri

Correspondence: sahbihidri@gmail.com
Department of English, Faculty of Human & Social Sciences of Tunis-University of Tunis, Tunis, Tunisia

Abstract

The study investigated the alignment process of the International English Language Competency Assessment (IELCA) suite examinations' four levels, B1, B2, C1 and C2, onto the Common European Framework of Reference (CEFR) by explaining and discussing the five linking stages (Council of Europe (CoE 2009). Unlike previous studies, this study used the five linking stages altogether to make fair judgements and informed decisions about the practical consequences and validity arguments of this mapping task. Findings indicated that the useful and in-depth discussions of the relevant CEFR descriptors resulted in a deeper awareness of establishing succinct re-familiarisations and re-definitions of the salient features of the different skills and items, thus making them more specific to reflect the CEFR descriptors. The ample alignment activities provided fertile ground for dependable results. For instance, teacher estimates confirmed the cut scores with high agreement percentages, ranging from 74.4 to 99.34. Also, the *FACETS* analyses showed a good global model fit with a high reliability value of the judgement process, only after undergoing rater training sessions. Specifically, the majority of item difficulty estimates were within the typical range, thus indicating that the IELCA examinations were measuring the underlying construct traits; however, the empirical validation called for additional data and further implementation practices regarding other judgements on the levels' boundary for IELCA examinations. Further mapping challenges, implications, and future research were also discussed.

Keywords: Familiarisation, Specification, Standardisation, Standard setting, Validation, Cut-off score, Linking, IELCA examinations, CEFR, CoE, Manual

Introduction

The study investigated the alignment process of the International English Language Competency Assessment (IELCA) suite of examinations to the Common European Framework of Reference (CEFR 2001), using the five linking stages (CoE 2009): (a) *Familiarisation*, (b) *Specification*, (c) *Standardisation Training and Benchmarking*, (d) *Standard Setting* and (e) *Validation*. Linking examinations to international

benchmarks in standard-based educational systems has become prevalent, resulting in a need for more standardised qualifications for people who have different purposes for using the CEFR (CoE, 2009), such as aligning language programmes to benchmarks or recognising examinations for official uses. For many stakeholders, this linkage has remained substantially challenging and contentious since its implementation requires stringent conditions. More interestingly, the benchmark process is a critical component in building evidence to support the interpretation of test scores; however, as pointed out in the literature, building this evidence is not sufficient (Papageorgiou & Tannenbaum (2016) on standard setting), and it is now unwaveringly snowballing into pedagogical actions to be taken by many stakeholders in different geographical areas that have decided to use the CEFR.

In this demanding requirement of a possible paradigm shift in language programmes and examinations, users who plan to map their standard-based educational qualifications onto the CEFR face a plethora of, for instance, relying on accurate judgements to map examinations or curricula. Several arrays of studies (e.g. Fleckenstein, Leucht, & Köller, 2018; Green, 2018; Lim, Geranpayeh, Khalifa, & Buckendahl, 2013) have debated clear evidence on the accurate judgements of the CEFR to qualify for international recognition or 'to inform test scores' users […] about the quality of the evidence provided to them' (Green, 2018, p. 59). The CEFR, meant to benchmark teaching, learning and assessment (North, 2014), has provided detailed methods on how to map qualifications (CoE, 2009) that have been approached from an argument-based validity perspective; however, unlike most of the previous studies that have approached mapping examinations to the CEFR, using one stage only, Standard Setting, this study tackled this mapping differently by including all the five stages. Whatever the mapping nature is, triggering research on addressing decisions about validity arguments, unequivocal interpretations and evidence of linking qualifications to the CEFR is still scarce.

The purpose of this paper was to link the IELCA suite of examinations to the CEFR by describing and discussing the Learning Resource Network[1] approach (LRN, https://www.lrnglobal.org/) to link its four examinations of the English language, including Reading, Listening, Speaking, and Writing with its two routes, Academic and General Training. The Academic Module assesses candidates' ability to access undergraduate or post-graduate programmes in educational institutions, colleges and/or universities of English-speaking countries; however, the General Module assesses a wide range of English language skills appropriate for secondary education, immigration, work experience and other training purposes (IELCA, 2019). A previous study to link IELCA examinations to IELTS was carried out by Hidri (2020), and it showed highly significant correlations between the two tests. This article meant to check whether the linking task led to making informed, valid and rational decisions and accurate judgements about the IELCA examinations by addressing the following questions:

a. How was the mapping task carried out?
b. To what extent did the CEFR linking stages build accurate evidence and judgements to demonstrate validity arguments for IELCA examinations?

[1]Recipient of the Queen's Prize for Enterprise: International Trade, 2020

Review of the literature

The CEFR (2001) has impacted the testing industry (Figueras, Kaftandjieva, & Takala, 2013; Fulcher, 2008). It is a framework of reference designed to provide a transparent and thorough foundation for syllabi and curricula (Athanasiou et al., 2016), and a 'reflection and discussion […] to provide a way of describing diversity' (Figueras, North, Takala, Verhelst, & Piet, 2005). For some researchers (e.g. Savski, 2020), the 'ideological contrasts' of the CEFR can reflect the language policy of a country where examination providers and decision-makers in standard-based educational systems use the CEFR for their own agendas. Alternatively, the CEFR has not only suggested a mutual understanding of the crucial characteristics of learner levels, but also provided an array of responses to questions pertaining to educational sectors, national and linguistic boundaries that can enable practitioner dialogue in relation to objectives and language levels (CoE, 2009). The CEFR is divided into three ascending level descriptors, A, B and C with two sub-levels each, and it largely targets different fields such as learning, teaching, assessment, instruction, curriculum design, and textbooks writing (Alderson et al., 2006; North, 2004; North, Figueras, Takala, Van Avermaet, & Verhelst, 2009).

As the manual states (2009, p. 9), a test has to be valid and reliable before carrying out the five stages of the linking process. For instance, in stage one, Familiarisation, a panel of internal and external experts who are supposed to have a priori and in-depth knowledge of the CEFR should be engaged in a series of training activities to become familiar with the validated and calibrated CEFR descriptors. Stage two, Specification, is about the detailed checking of the content and task types of the CEFR examinations (CoE, 2009, p. 10), and this checking task is carried out by raising the panellists' awareness of the arguments of content analysis and familiarity with the CEFR. Panellists should describe the test in terms of internal validity and then provide a broad compendium of information on exam development, grading, data analysis and results; however, providing such information largely depends on the panellists' views of language, language examinations, testing theories and perceptions and practices of CEFR mapping and assessment in general. Stage three, Standardisation Training and Benchmarking, addresses panellists' familiarity with the 'Common Reference Levels' in the spoken and written performances. This stage is tied to stage one, Familiarisation, in that it strengthens the panellists' familiarity with the CEFR descriptors, and it is meant to ensure that the judges' rating is consistent with the different construct traits (McNamara, 1996; McNamara, Knoch, & Fan, 2019); however, in carrying out this grading task, raters find themselves driven into more subsequent moderation rounds to attain consistent rating.

Stage four, Standard Setting, deals with taking the right decision of the cut-off scores or borderline performance to allocate the test-takers' performance to one of the CEFR contextualised descriptors; however, determining these scores rests on fixing and using the appropriate levels efficiently and objectively to match the targeted descriptors. Because of its significant effects (Kane, 2017), Standard Setting should be 'accurate, reliable, valid, useful, and defensible, which is not an easy challenge due to the mix of content expertise, judgement, policy intentions, measurement, and statistical expertise' (Blömeke & Gustafsson, 2017, p. 2). The choice of a specific Standard-Setting method is contingent upon two caveats: (a) format and purpose of examinations and (b) linking scores to the purpose and method of examinations. For instance, the examinee paper

selection method involves the scoring of polytomous items for judges to use exam replies instead of independent ratings to get the average score for each performance and then use this score as the 'minimum passing value' whose total summary is meant to determine each performance standard for the test (Hambleton, Jaeger, Plake, & Mills, 2000, p. 359). The item descriptor matching method (IDM) can reliably reduce subjective evaluations while focusing on the tests' contents (Ferrara, Perie, & Johnson, 2014). The judges need to match test items with the test-takers' ability by highlighting what the test-takers should know and be able to do to reply successfully to an item. The content experts, who are also the panellists, are supposed to check the test-takers' replies and their knowledge and skills of the CEFR and they should also examine the borderline item score and use a scoring rubric to anchor these items at a specific CEFR descriptor by showing a high inter-rater reliability throughout subsequent rounds of consensus agreement. The test items are arranged based on their gradual difficulty level by using an Item Response Theory (IRT) scale. IDM and Basket methods ask for the requirements of an item and its matching with the 'can-do' descriptors (ALTE). However, despite its hypothetical nature, in the Basket method, an applicable method for the receptive skills, listening and reading, as well as grammar and vocabulary, panellists are required to put each item in a basket that matches the targeted CEFR descriptors by considering an abstract examinee's performance who is hypothetically supposed to answer an item correctly at a specific CEFR level. This method can be used in dichotomously binary scored test items such as multiple-choice and constructed items where a range of 0 to 2 partial credit is allowed.

In the last stage, Validation, panellists should check if familiarisation and training reached the objective of the initial planned evaluation by relying on collecting valid evidence from other stakeholders who are involved in the mapping task such as teachers. The manual (2009, p. 90) defines validation as '[…] the body of evidence put forward to convince the test users that the entire process and its outcomes are trustworthy.' Along with procedural and internal validity, panellists should do external validity to confirm the results of the other stages. It is commonly believed that the nature of the Validation outcomes always reflects the panellists' or teachers' views of examinations in general.

Some previous studies have targeted different Standard-Setting methods such as IDM method (Ferrara et al., 2014), the Basket method (Cizek, 1993), examinee paper selection method (Hambleton et al., 2000), Analytic Judgement method (Plake & Hambleton, 2001), Body of Work method (Cizek & Bunch, 2007), Angoff method (Angoff, 1971), Bookmark method (Mitzel, Lewis, Patz, & Green, 2001) and others. Nichols, Twing, Mueller, and O'Malley (2010) addressed human judgement in being subjective, which means that reaching an agreement on a cut score might loom difficult. This result is echoed in the study carried out by Hein and Skaggs (2010), who claimed that panellists faced difficulties in getting the overall cut-off score. Similarly, Papageorgiou (2010) addressed the panellists' discussion to report the cut scores for students and found that selecting an appropriate score is challenging since external factors might vitiate current selections (see the 'Discussion' section on rater (in)consistency). Other studies used teachers' judgements in placing secondary-school participants in levels commensurable with the CEFR (Fleckenstein et al., 2018); and others discussed the mapping of the Dutch state examination onto the CEFR by selecting cut-off scores for students' admission (Bechger, Kuijper, & Maris, 2009). Harsch and Kanistra (2020) used

the IDM method to link a suite of examinations to the CEFR and they found that panellists showed high consistent rating in the independent and integrated tasks. For Kim and Crossley (2020), their model of analysis partly maps with the CEFR macro-functions. In another study on the commensurability of IELCA with IELTS, Hidri (2020, p. 745) aptly used the equipercentile method and reported correlations of .98, .97, .94 and .95 for Listening, Reading, Writing and Speaking, respectively, between the IELCA and IELTS exams. Whatever the nature of Standard Setting might be, it is recommended that the stakeholders involved in the mapping tasks balance some methods to achieve objective and fair results.

Research on the use of stage four of the manual, Standard Setting, to build validity arguments for any type of examinations is abundant (e.g. Athanasiou et al., 2016; Ferrara et al., 2014; Fleckenstein et al., 2018 and many others). Unfortunately, addressing the other four stages, along with this stage, in building these valid arguments has not been given its due momentum despite the fact that the manual (2009) hails all the five stages equally and does not in any way favour any particular stage to carry out the mapping task. What is perhaps particular about the current study is its hypothesis that the use of the manual mapping stages (2009) can ultimately build evidence to demonstrate the validity arguments of the IELCA examinations. This study goes further in trying to show that this evidence of validity arguments remains contingent upon the mandatory inclusion of all the five mapping stages of the manual (2009) as suggested in the following section.

Method

This study aimed to map the IELCA examinations onto the CEFR by following the manual's linking requirements (CoE, 2009). The IELCA entry level starts with entry 3, from B1 to C2. To describe the IELCA examinations, the Reading Module includes three sections of 40 items to be done in 1 h and 20 min. The 30-min Listening contains forty test items with extra 10 min to copy answers on the answer sheet. The listening items are allocated to three sections consisting of up to 8 short listening extracts, all lasting approximately 12 min in total. Candidates listen to the listening input once. The 1-h Writing Module includes two tasks for candidates to write 120–150 words and 180–220 words for tasks one and two respectively. The Speaking Module includes three sections that last 11 min. Table 1 presents an overview of the IELCA examinations[2] and the Standard-Setting methods.

Familiarisation

The *Familiarisation* activities were meant to establish consistency in applying CEFR grading to all components. Recruiting the 14 panellists was carried out according to the manual (CoE, 2009, p. 17–18) and it lasted 6 months. The panel, who was supposed to cover a combination of the 4 component areas, Reading, Listening, Writing and Speaking and for some qualifications, Speaking and Listening combined, were trained so that they could be familiarised with the sections of the CEFR, mainly the ones covering B1 to C2 levels. To demonstrate a representative sample of adequate expertise in language proficiency, the panel included a variety of disciplines ranging from rating, EFL

[2]For a more comprehensive overview of the types of tasks, check the LRN website.

Table 1 IELCA qualifications

Subtest and modules	Level	Number of tasks	Standard-setting method
- General Reading[a] (1 h 20 min)	Entry 3-Level 3 (CEFR B1-C2)	40 items; three sections	Basket and IDM
- Academic Reading (1 h 20 min)	Entry 3-Level 3 (CEFR B1-C2)		Basket and IDM
- Listening (1 subtest of General and Academic) (30 min)	Entry 3-Level 3 (CEFR B1-C2)	40 items; three sections	Basket and IDM
- General Writing[b] (1 h)	Entry 3-Level 3 (CEFR B1-C2)	*Section 1*: (120–150 words)	Examinee paper selection
- Academic Writing (1 h)	Entry 3-Level 3 (CEFR B1-C2)	*Section 2*: (180–220 words) 8 items for each module	Examinee paper selection
- Speaking (1 subtest of General and Academic) (11 min)	Entry 3-Level 3 (CEFR B1-C2)	*Section 1, 2 and 3 8 items for each module*	Examinee paper selection

[a]Reading Academic and General Modules have different text genres
[b]Writing Academic and General Modules have different prompt genres

teaching, standard setting, research and test development to an educational post-graduate level in linguistics and post-doctorate level in assessment and psychometric testing. Table 2 describes the internal and external panellists' scope of experience and education. The external members were examiners, examination/test constructors, or consultants in test development, not currently employed by LRN.

All panellists were required to have at least 5 years of teaching experience. Twelve of the 14 panellists had UK experience, while 9 out of 14 had an EFL abroad experience and then taught in other countries. The panellists were divided into two: Panel 1 covered receptive components (Reading and Listening), and panel 2 covered Speaking, Writing, and Speaking and Listening combined. For reasons of confidentiality and ethicality, the panellists' identities are not disclosed.

Due to the broad range in the location of panellists, the Familiarisation task was carried out via face-to-face and Skype meetings, with more in-depth tasks taking place during meetings. Meetings for both panels were held separately under the guidance of a coordinator expert in language testing and the CEFR. The pre-familiarisation stage, meant to study the relevant sections of the CEFR (B1 to C2), took place 21 days prior to the Familiarisation meeting. All panellists were required to (a) view the CEFR global scale alongside the completion of a matching exercise so that they could be (re)familiarised with language and keywords pertaining to each level, (b) orientate to the 'can-do' descriptors within the Self-Assessment Grid through the completion of gap-filling activities, (c) be familiarised with the qualitative aspects of spoken language use

Table 2 Panellists' profile (n = 14)

Experience	Education
Teaching experience, including experience in CEFR levels, from Elementary (A1) to Proficiency (C2) (100%)	CELTA
Examining/rating experience (54%)	DELTA
Standard-setting experience (29%)	BA
Item writing/test development (31%)	MA
Research (38%)	PhD

through the completion of titling and gap-filling activities, (d) match descriptors' levels from the illustrative scales in the CEFR and (e) complete activities taken from the website http://www.helsinki.fi/project/ceftrain/index.php.66.html, thus demonstrating knowledge of the salient characteristics of each level. As C2 tasks were limited, past papers released to the public domain from other examining boards were used to benchmark proficiency items. Panel 1 worked on Reading and Listening. For Listening, alternative proficiency sample response (C1) focused on task 1 (C1), task 2 (B1), task 6 (B1) and task 4 (B2), while Reading focused on tasks 1 to 9 also with alternative proficiency sampled response C2. For panel 2, they addressed Speaking and Writing, and then Speaking and Listening combined. Speaking focused on task 1 (C1), task 4 (B2), task 8 (B1) and task 9 (B1). Listening had task 2 (B1). As for Writing, it included tasks 1 to 6 and 8 to 11 with alternative proficiency sampled responses C2. All tasks covered all targeted levels of the CEFR.

Two meetings, which were part of the Familiarisation process, were held at the LRN head office in London on 10th and 11th April 2014, and they started with reviewing the main issues identified from the remote Familiarisation activities. During the meeting, the groups were paired off to recap the salient features of the CEFR per level while noting the relevance of each feature to the LRN examinations and relevant subtests for the linking process. A matching exercise followed this task to reinforce the conclusions rounded off by the pair work and group discussion activities. The rating instruments used commenced with a global overview of the candidates' levels before tapping into the level more deeply through using Table C2 from the CEFR to give a more comprehensive profile on candidates' performance. The plus levels were used too, where a candidate's response was considered stronger than B1 but not quite B2. The process of reaching a consensus view also required referring to transcripts of spoken production or key areas of the written script where panel members had used specific areas of response to support their judgement. The discussion points were easier to refer to and scrutinise in cases of disparity since all individual and pair activities had been recorded.

The training sessions took place from 10th to 15th April 2014, followed by the Familiarisation meetings on 8th and 9th May 2014. Each day was devoted to a single component spanning from B1 to C2 of the CEFR. In overviewing training on Reading and Listening, re-familiarisation was conducted for each area by highlighting the Familiarisation results. This task led to the training session of A1-C2, where relevant, on the CEFR by covering the salient characteristics for reception, particularly B1-C2. The panel members were required to (a) complete the table with the correct headings transferred from the table (CoE, 2009, p. 124), (b) complete matching activities that connected levels with descriptors and (c) complete gap-filling activities in relation to the 'can-do' descriptors. The main features observed were then recapped, highlighting the differences between B1 and B2, and then B2 and C1. The differences between C1 and C2 were evident among panel members.

Panel 2 training for Speaking, Speaking and Listening, and Writing was again conducted for each area while fleshing out the familiarisation meeting results in April and recapping on the observations found. Panel 2 training session covered the salient characteristics, mainly focusing on B1-B2 for Listening and C1-C2 for Speaking. Panel members were required to (a) complete the bold points of the table with the correct points transferred from Table C1, (CoE, 2009, p. 184); (b) complete matching activities

that connected the level with the descriptor; (c) complete gap fill activities as to the 'can-do' descriptors for Speaking, Listening and Writing; (d) match the CEFR descriptors and marks allocation from the mark schemes for Speaking and Writing components of IELCA; and (e) carry out two rounds of CEFR level and mark scheme allocation as in Table 3.

The data collected during the Familiarisation stage was analysed using Cronbach alpha and intra-class correlation coefficient with the CEFR levels being assigned by the panellists and converted into numbers. A two-facets Rasch analysis was performed using the FACETS 3.71.2 (Linacre, 2013a), and it implemented the ratings that raters awarded to examinees to estimate individual rater severity and task difficulty.

Specification

The *Specification* phase, meant to form a claim on linking each subtest's content in relation to the CEFR, involved the reinforcement of knowledge and content analysis built in the Familiarisation stage. The LRN's inhouse test development team passed to the panel for agreement and further completion. After agreeing on the relevant forms, the group was given sufficient time to work in groups and highlight key answers to be inserted into the forms. The panel was then divided into two groups, with each group being assigned one person to feedback into the session where observations were highlighted for discussion. Table 4 outlines the forms used to build the content analysis evaluation (CoE, 2009, p. 126–147).

Through completing the Specification stage, the construct of each subtest sparked reflection of candidate competency through the embedment of the list below into items, rubrics and mark schemes:

- Genre and background of each text/extract setting
- Communication tasks
- Text and extract length
- Difficulty scale of text and extract
- Rubric for lexical and grammatical range and accuracy
- Difficulty scale in relation to the CEFR scales
- Socio-linguistic (e.g. linguistic markers, register, adequacy and dialect), strategic and pragmatic competencies

Standardisation, training and benchmarking

Standardisation took the form of benchmarking judgements through additional samples at all points on the scale, including the plus levels, which reinforced the

Table 3 CEF level and mark scheme round allocation

Speaking	Speaking and Listening	Writing	Apply level to CEFR	Apply grade to LRN mark scheme
B1 × 2	B1 × 2	B1 × 2		
B2 × 2		B1 × 2		
C1 × 2		B2 × 2		
C2 × 2		C1 × 2		
		C2 × 2		

Table 4 IELCA content analysis evaluation

IELCA content analysis evaluation	Level claimed	Qualifications forms
Reading comprehension	B1 to C2	A1-General Examination Description
Listening comprehension	B1 to C2	A2-Test Development
Spoken interaction	B1 to C2	A3-Marking
Spoken production	B1 to C2	A4-Grading
Written interaction	B1 to C2	A5-Reporting Results
Written production	B1 to C2	A6-Data Analysis

familiarisation and training process. It was observed with some panel members that the application of the CEFR was second nature and that the metalanguage used throughout the manual was of common usage, whereas other panel members, despite their familiarity with the CEFR levels, used a different type of language. Most importantly, the terminology that panel members used unequivocally echoed their knowledge of the CEFR principles.

Standard setting

In the *Standard-Setting* stage, the panel, who were also the examiners, used the Basket method and IDM method in the Reading and Listening sections. This selection was made individually, with each member who was given sufficient time to cast judgement with the level and rationale behind the level assigned. The method used for the productive skill was the examinee paper selection. Samples chosen for the Speaking and Writing sections are born from live test sittings and reflected substantially different bands in the mark scheme corresponding to each section, except for the Speaking section, where the boundaries between the different CEFR levels were not clear.

Validation

For empirical *Validation*, the manual (2009) suggests some ways for undertaking the linking process. For external validity evidence, the panellists collected teacher judgements as the external criterion to be implemented towards the examinations' internal validity. Table 5 details the LRN competency profiles of the candidates as part of the recruitment process.

It was essential that examiner reliability fall in line with LRN quality assurance standards. All examiners had to attend examiner induction training and complete post-training, whereby their rating performance was analysed statistically through the many-facet Rasch model. On completion of this task, they were then assigned as examiners/raters, which helped identify lenient, severe and inconsistent examiners/raters. All examiners/raters were given a chance to improve as examiners/raters through regular monitoring and feedback in addition to annual standardisation followed by

Table 5 Candidates Profile in the External Validation Stage

	Total no.	Age group	Gender		Origin			
			Female	Male	Indian	Malaysian	Nigerian	Pakistani
General	304	16-51	88	216	114	43	83	64
Academic	293	15-51	93	200	83	40	120	50

performance appraisal. However, where inconsistencies were identified, examiners/raters were not entered into the pool for that allocation of sessions until their rating performance was considered consistent. For external validation, the results of teacher estimates were chosen as external criteria. All teachers were asked to provide judgements regarding students' language proficiency levels, which were compared with the students' actual test scores.

Results

The study tried to link the IELCA examinations to the CEFR by corroborating the relevant evidence for this mapping task. The different linking phases led to valid and reliable results. In the Familiarisation phase, panel members needed to orientate more to specific areas related to Reading, Listening and Speaking. In the Reading component in Task 2 (B1), the Familiarisation meeting outcomes led to re-familiarising the salient features of 'identifying cues and inferring' and 'reading for information and argument' from B1 to B2+. Some panel members pitched this exercise at B1 for 'reading for information and argument' due to the items requiring a higher competency level than the text itself. An in-depth discussion involving the relevant descriptors highlighted the need for panellists to re-evaluate and define 'identifying cues and inferring', 'reading for information and argument' and the redefinition of the requirements of the items in the task. In the Listening component area in Task 2 (B1), the familiarisation meeting outcomes led to re-familiarising the salient features of 'identifying cues and inferring'.

Table 6 presents the analyses indicating the judges' scores in understanding the CEFR scales.

Some panel members thought the exercise was more difficult than the target level due to one of the speakers' accent and some of the unknown vocabulary and phrasal verbs. However, B1 level requires listeners to infer the meaning of unknown words through the context with the pace of delivery of this particular task beyond the

Table 6 Judges' scores in the Familiarisation tasks (n = 14)

(B1–C2 CEFR)												
L1	**L2**	**L4**	**R1**	**R2**	**R3**	**S1**	**S4**	**S5**	**S6**	**W2**	**W4**	**W6**
1	3	6	1	9	3	1	5	3	1	5	1	3
1	3	1	1	3	3	2	5	3	1	4	1	3
1	3	5	1	3	4	1	5	3	1	5	6	3
						1	5	5	3	5	6	9
1	4	5	6	9	4							
						1	3	5	1	6	1	3
						1	3	5	5	5	1	3
1	3	5	1	3	3	1	5	3	1	5	1	3
1	3	5	1	3	3	1	4	5	5	5	1	9
2	3	5	1	3	3	1	5	3	1	5	1	3
						1	5	3	1	5	1	3
1	3	5	2	8	9							
1	3	5	1	3	3							
1	3	5	1	9	3							

1, raters from R1 to R14 = Ma, Ol, Em, Ra, Sa, Ja, Ro, Ni, So, Tr, Ek, Ca, Ma and Jo; 2, ratings were coded by the following scheme: c2 = 1, c1 = 2, b2 = 3, $b1^{+}$ = 4, b2 = 5 and $b2^{+}$ = 6

capability of B1, with no background noise or interference. In the Speaking component area in Task 8 (Liam), the familiarisation meeting outcomes led to re-familiarising the salient features of 'fluency' and 'coherence'. Some judges thought Liam's level was slightly higher than B1. Also, the comparison to Liam's partner bore some confusion in pitching Liam's level independently from Larry. Liam's confidence, pronunciation and general application of language served as a mask for some panel members in enabling themselves to focus on the underlying features of his level.

To explore the source of rater effect in familiarised CEFR levels from B1 to C2, the two-facet Rasch model measurement was used with the global fit results displaying the dataset that sufficiently fit this model well (i.e. rater and task). Rater severity and task difficulty were equally calibrated on the same interval, i.e. logit scale, which allowed to interpret the results of unifying the reference scale. The data-model fit was evaluated by examining unexpected responses (see Table 7 for the *FACETS* analysis) given the model's assumptions. A palatable model can be attained when about 5% or less of (absolute) standardised residuals are ≥ 2, and about 1% or less of (absolute) standardised residuals are ≥ 3 (Linacre, 2013b). According to Table 7, five unexpected responses had standardised residuals larger than 3.0, taking 1.9% of the ensemble of responses (260) and indicating a good global model fit.

Table 8 presents rater code, rater severity, error and infit and outfit mean-square values and other group statistics (mean, standard deviation of the mean, separation index fixed chi-square with degree of freedom and significance level) accompanying each map. The results of the two-facet calibrations for averaged ratings on the B1-C2 tasks are mapped out in Table 8 where columns 1–5 display the logit scale, estimates of intra-rater severity with raters at the top being more severe than those at the bottom, task difficulty (tasks located higher were more difficult to receive high ratings than tasks located lower in the column) and the nine-point rating scale transformed from the raw rating scale from C2 to B1, respectively. Rater effect can be evaluated through the mean of measure (logit), standardised error of mean, chi-square (with degree of freedom and significance level), separation index and separation reliability.

The severity span between the most lenient rater Sa (logit = .94) and the most severe rater Ol (logit = − .67) was 1.61 logits which was confirmed by the fixed chi-square value (with degree of freedom) of 58.6 (13), significant at p = .000 level. However, the relatively small value of separation (r = 2.06) indicated that the extent of rater severity variation was not large. Examining the variable map, Fig. 1, revealed that rater severity measures (column 1) were roughly located between two horizontal levels of − 1 to + 1 logit values, thereby indicating a certain degree of variation in rater severity levels among the 14 raters. Columns 2 and 3 included the tasks measured and the scale to map these performances. The individual fit of rater performance was evaluated

Table 7 Unexpected responses (7 residuals) (CEFR Level B1-C2)

Category	Score	Expected	Residual	Standardised residual	Rater	Task
2	2	1.0	1.0	4.4	10 Tri	14 LT1
6	6	2.6	3.4	3.4	5 Sa	17 RT1
2	2	1.1	.9	3.2	2 Ol	20 ST1
6	6	2.0	4.0	3.9	3 Em	25 WT4
6	6	2.8	3.2	3.2	4 Ra	25 WT4

Table 8 Raters measurement report in the Familiarisation activity (n = 14)

Rater code	Raters	Rater severity measure (in logits)	Error	Fit	
				Infit MnSq	Outfit MnSq
R2	Ol	− .67	.49	.79	.87
R8	Ni	− .58	.30	.23	.25
R13	Ma	− .56	.47	.33	.26
R10	Tr	− .55	.32	.98	.92
R6	Ja	− .54	.41	.31	.38
R11	Ek	− .54	.25	.23	.35
R3	Em	− .40	.82	.82	.87
R7	Ro	− .40	.75	.74	1.03
R9	So	− .25	1.24	.86	.85
R1	Ma	− .03	1.42	.83	1.68
R4	Ra	.35	1.69	1.42	.83
R14	Jo	.35	1.23	.97	2.03
R12	Ca	.76	1.62	1.08	1.31
R5	Sa	.94	2.59	2.82	2.96

Group statistics: M = − .15, SD = .52, Separation = 2.06, Reliability =.81, fixed Chi-Square (df) = 58.6(13), p = .00

referring to rater fit indices (column 5, Table 8), presenting two mean-square statistics showing data-model fit for each rater: Infit and outfit with the former being sensitive to an accumulation of unexpected ratings and the latter to individual unexpected ratings. Both can value from 0 to infinite, but with an expected value of 1 (Linacre, 2002; Myford & Wolfe, 2003).

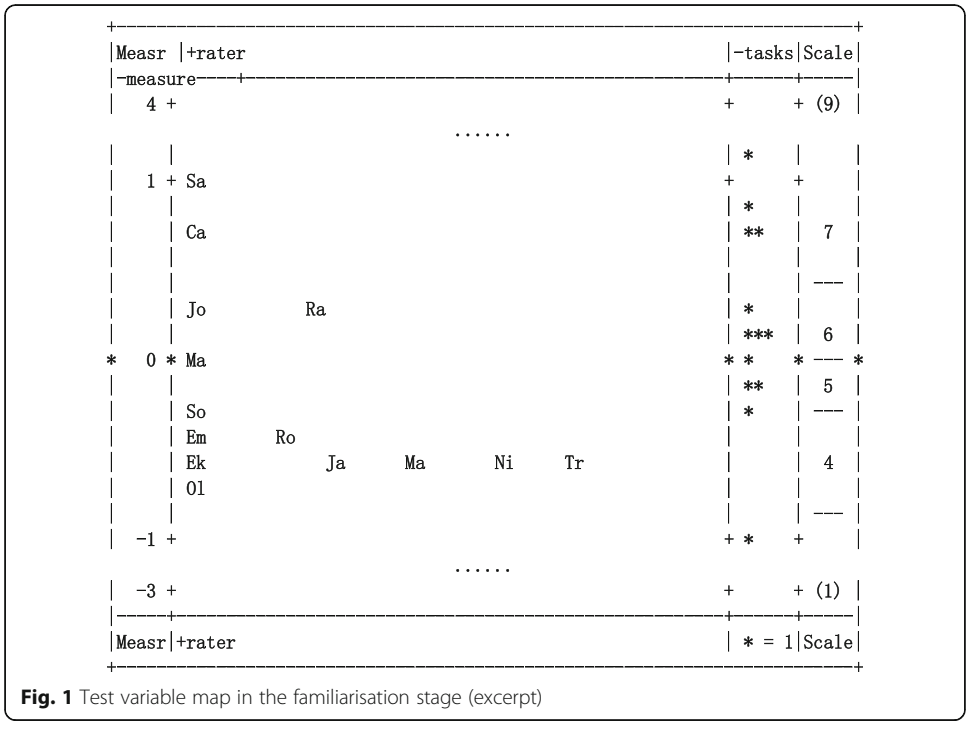

Fig. 1 Test variable map in the familiarisation stage (excerpt)

As can be seen in column 5, individual rater infit values ranged from .23 (Ni) to 2.82 (Sa) and the outfit values ranged from .25 (Ni) to 2.96 (Sa). If applying the wide range of lower and upper limits control, four raters were sceptical of insufficient fit (i.e. Ni, Ma, Ek and Ja) and at least one rater (i.e. Sa) was suspected of central tendency or halo effect for overfitting (see Engelhard, 2009; Myford & Wolfe, 2003). Table 9 provides a further summary of rater fit description.

Retraining at this stage came in the form of more sampling taking place for B1-C2 levels, followed by discussion against the relevant descriptors of the CEFR (CoE, 2009). Following the CoE (2009, p. 36–54), to be assured of confidence in the panel members for the Standard-Setting phase, it was essential to follow all the linking stages to complete the training of the panel members. Given the discrepancies highlighted, the following steps were followed with each panel member equipped with:

- The Salient Characteristics for reception (CoE, 2009: Table A2, p. 124 and 125)
- The Global Oral Assessment Scale (CoE, 2009: Table C1, p. 184) of the CEFR to maintain their global impression of candidate level for spoken performance
- The Oral Assessment Criteria Grid (CoE, 2009: Table C2, p. 185) for spoken performance
- The Supplementary Criteria Grid–plus levels (CoE, 2009: Table C3, p. 186) for spoken performance
- The Written Assessment Criteria Grid (CoE, 2009: Table C4, p. 187) for written performance
- All relative descriptor levels of the CEFR
- The Association of Language Testers in Europe (ALTE) 'Can Do' Statements
- A selection of subtest item responses-Writing and Speaking (not used in standard setting)

This stage was then followed by a second round of data collection using Cronbach alpha and intra-class correlation coefficient. Results in Table 10 indicated that the judges were highly reliable in using the CEFR scales at the end of the final training session.

In the Specification phase, changes to related subtests and supporting specification documents are detailed in italics on the relevant mark schemes. First, the Specification process enabled the team to objectively recognise gaps in the mark scheme against the CEFR descriptors. Second, the C1 Task Achievement descriptor in IELCA speaking, relating to Qualitative Factors for Reception-Table A3, CEFR, was lengthened to-'able to understand prompt and respond appropriate to the prompt *through identifying speaker's attitude and opinion through tone of voice, intonation, and stress*.' Third, the B2 Task Achievement descriptor in IELCA Speaking, relating to Qualitative Factors for Reception-Table A3, CEFR, was lengthened to 'where clarification is needed can use

Table 9 Summary of the rater fit description

Logits	Infit	Outfit	Notes
Less than .50	Ni Ma Ek Ja	Ni Ma Ek Ja	Insufficient variation
Between .50 and 1.50	Ol Tr Em Ro So Mar Ra Jo Ca	Ol Tr Em Ro So Ra Jo Ca	appropriate variation
Larger than 1.50	Sa	Sa Jo Ma	Central tendency or halo effect

Table 10 Judges' rating reliability

	Listening	Reading	Speaking	Writing
ALPHA	.99	.99	.97	.99
ICC	.95	.94	.78	.95

strategies in order to understand.' Finally, C1 Language and Appropriacy descriptor for IELCA Speaking, relating to the Factors for Grammatical Accuracy scale was amended from 'Language is generally accurate...' to '*Language is consistently accurate with rare grammatical errors, although there may be minor slips*'.

In the Standardisation stage, the methodology advised in the manual (2009) was followed spanning a 6-month period. Due to a lack of a sufficient number of C2 items in IELCA subtests of Reading and Listening, the cut score for C2 was higher, which was also confirmed by teacher judgements' cut scores. This threshold was addressed again in line with LRN's qualification review procedure, with further Standard-Setting rounds to support the substantial claims made. The accuracy in these claims on the CEFR levels and grading boundaries was high with the accurate predictions given on grade boundaries for IELCA C2. For example, the prediction for a C2 grade in IELCA Writing fell between 75 and 83.

In the Standard-Setting phase, the cut scores for all components of IELCA are presented in Tables 11, 12, 13, 14 and 15. While the cut scores for B1 and B2 levels seemed reasonable, considering the number of questions in each section and the spread of levels across the different CEFR levels that the test targets, an insufficient number of C2 level items in the tests posed concerns over any possible unsubstantial claim made for this level. Table 11 shows very high cut scores for Reading and Listening C2 level. Test providers were well aware that the suggested cut scores were very high and were cautious in their claims at the C2 level. However, they were unwilling to make any suggestions at this level until further data was collected to support the C2 level cut scores established at the end of the Standardisation stage. All Standard-Setting forms contained space allocation for CEFR level and mark scheme descriptors, where required.

Table 12 gives the cut scores suggested for the writing and speaking sections of the qualification. Based on the samples of the Standardisation stage, cut scores were suggested for the General and Academic Writing sections. The cut score between B1 and B2 for Speaking indicated in Table 12 was a result of external validation.

Tables 13 and 14 present the Writing and Speaking subtests data where the CEFR levels were converted into numbers to allow for statistical analyses of the data. The mean ratings on the numeric scale with their corresponding CEFR level and standard deviation are provided in the tables and the ratings of the samples using IELCA mark scheme.

To investigate judges' consistency and agreement, Cronbach alpha and intra-class correlation for all IELCA exam sections were used. Table 15 indicates high alpha and

Table 11 Reading and Listening cut scores

Section	B1 cut score	B2 cut score	C1 cut score	C2 cut score
Reading General	12	23	36	40
Reading Academic	11	23	35	40
Listening General	14	24	35	40

Table 12 IELCA Writing and Speaking cut scores

Writing General		Writing Academic		Speaking	
IELCA score	**CEFR level**	**IELCA score**	**CEFR level**	**IELCA score**	**CEFR level**
40–50	B1	38–50	B1	40–59	B1
51–55	B2	51–64	B2	60–88	B2
56–74	C1	65–74	C1	89–97	C1
75–83	C2	75–83	C2	98–100	C2

Table 13 Writing General and Academic—conversion of CEFR levels to numbers

Samples	CEFR judgements			Marking within IELCA rubrics			
	Mean numeric	**CEFR level**	**SD[a]**	**Mean grade parts**	**SD**	**Mean grade overall**	**SD**
Writing General							
S1P1	1.8	A2+	.42	18.4	1.84	40.2	2.35
S1P2	2	B1		19.8	.42		
S2P1	4.7	B2	.48	26.4	.97	58.7	1.70
S2P2	5.8	B2+	.42	32.3	1.25		
S3P1	4.2	B2	.42	25.8	1.23	52.4	1.96
S3P2	4.2	B2	.42	26.6	.84		
S4P1	6	C1		31.7	.48	57.8	2.20
S4P2	4.6	B2	.52	26.1	1.85		
S5P1	7	C1		37.1	.32	76	.48
S5P2	7	C2		38.9	.32		
S6P1	2	B1		18.2	.42	35.1	.57
S6P2	2	B1		16.9	.57		
S7P1	4	B2		25.4	.70	51.3	.67
S7P2	4	B2		25.9	.32		
Writing Academic							
S1P1	**1.2**	**B1**	**.42**	**19.8**	**.79**	**39.7**	**1.16**
S1P2	1.1	B1	.32	19.9	.57		
S2P1	4.8	B2+	.42	31.7	1.77	62.7	3.65
S2P2	4.8	B2+	.42	31.8	1.81		
S3P1	3	B2		25.6	.52	52.1	2.60
S3P2	3.3	B2	.48	27.5	1.08		
S4P1	1	B1		18.9	.57	38.6	.70
S4P2	1	B1		19.7	.48		
S5P1	3	B2		24.7	.48	51.2	.79
S5P2	3	B2		26.5	.53		
S6P1	5	C1		32.4	.52	65.3	.48
S6P2	5	C1		32.7	.48		
S7P1	6	C2		38.1	.32		
S7P2	6	C2		39		77.1	.32

[a]Standard deviation

Table 14 Speaking conversion of CEFR levels to numbers

Samples	CEFR judgements			Marking within IELCA rubrics	
	Mean numeric	CEFR level	SD[a]	Mean grade overall	SD
S1	2.8	A2+	1.03	38.7	1.70
S2	5.1	B2+	.32	67.5	1.08
S3	6.9	B2+	.32	88.9	4.28
S4	3	B1		41.5	0.85
S5	5.7	B2+	.48	71.3	1.16
S6	7	C1		88.4	1.26
S7	8	C2		98.1	0.32

[a]Standard deviation

ICC values, thus suggesting the judgement process was highly reliable, with a complete agreement for the Writing, Speaking and Listening sections.

In the Validation stage, a decision consistency table was created using teachers' judgements and students' placement into CEFR levels based on the cut scores suggested at the end of the Standardisation stage. Table 16 describes the comparison between IELCA test results and teacher estimates for different sections of the testing with the results suggested a high agreement between the candidate classifications made by the test based on the cut scores determined at the end of the Standard-Setting stage in addition to the ones made by the teachers. The agreement percentages for Academic Reading were 74.4%, 83.6% for Listening and 98.68% and 98.29% for General Writing and Academic Speaking, respectively. The agreements for General Reading were 99.34% and 98.68% for General Listening, Writing and Speaking, respectively.

Discussion

The current linking process was established to (a) relate IELCA alignment to the CEFR descriptors, thus laying the foundation for the mark schemes (Speaking and Writing) and item construct for all components of each subtest; (b) offer evidence in support of the claims made on rational and accurate decisions and judgements that emanate from the five linking stages; (c) enable the test development team, i.e. item writers, reviewers and raters, to acquire more knowledge and be more familiarised with the CEFR descriptors; and (d) develop systems to maintain a quality approach to the CEFR level benchmarking and future standardisations. The observations gleaned were about the accuracy in the claims on the CEFR descriptors. For instance, due to some lack of C2 items in IELCA subtests of Reading and Listening, the cut score for C2 was higher, which was also confirmed by teacher judgements' cut scores. In general, a growing

Table 15 Alpha and ICC values for the general and academic modules

Section	No. of items/samples	Alpha	ICC
Reading General	40	.98	.88
Reading Academic	40	.98	.86
Listening	40	.98	.85
Writing Academic	8	.99	.97
Writing General	8	.99	.95
Speaking	8	.99	.93

Table 16 Agreement between test results and teacher estimates

		Teacher estimates				Total
Academic Reading (74.4%)		B1	B2	C1	C2	
Test results	B1	74				
	B2		170			
	C1			33		
	C2				16	
Total		74	170	33	16	293
Academic Listening (83.6%)		B1	B2	C1	C2	
Test results	B1	10	5			
	B2		184	43		
	C1			51		
	C2					
Total		10	189	94		293
Academic Writing (74.4%)		B1	B2	C1	C2	
Test results	B1	85				
	B2		177			
	C1			24		
	C2				8	
Total		85	177	24	8	293
General Writing (98.68%)		B1	B2	C1	C2	
Test results	B1	95				
	B2		181			
	C1			21		
	C2				7	
Total		95	181	21	7	304
Academic Speaking (98.29%)		B1	B2	C1	C2	
Test results	B1	88				
	B2		184			
	C1			26		
	C2	88			6	
Total		88	184	26	6	304
General Reading (99.34%)		B1	B2	C1	C2	
Test results	B1	103				
	B2		174			
	C1			17		
	C2				10	
Total		103	174	17	10	304
General Listening (98.68%)		B1	B2	C1	C2	
Test results	B1	76				
	B2		183			
	C1			24		
	C2				9	
Total		76	183	24	9	292[a]
General Speaking (98.68%)		B1	B2	C1	C2	
Test results	B1	94				

Table 16 Agreement between test results and teacher estimates *(Continued)*

		Teacher estimates				Total
	B2		189			
	C1			15	6	
	C2					
Total		94	189	15	6	304

[a]1 candidate below B1 according to the test was not included in the data

consensus was not an overly difficult outcome to achieve, and panel members commented that the process of referring back to sections of utterances, in addition to relating to the deeper profiles in Table C2 of the CEFR, was valuable, thus playing a pivotal role throughout the Familiarisation stage. Despite some disagreements among panellists, consistency remained high, and it was felt that closer observations on the descriptors bore useful discussion, which raised a deeper awareness of the CEFR and fostered confidence in moving to the next stages of the process.

Like other studies (e.g. Alderson et al., 2006; Ferrara et al., 2014; Figueras et al., 2005 and Fleckenstein et al., 2018), the use of the CEFR descriptors on the part of teachers to map the IELCA levels was increasingly productive as this linking reflected their education training that can ultimately be used in innovative ways to address examination or curriculum issues. Evidence of validity arguments of other examinations has been tackled in many studies; however, most of these studies (e.g. Bechger et al., 2009; Green, 2018; Lim et al., 2013 and Papageorgiou, 2010) addressed such arguments from one perspective, that of Standard Setting. Unlike these studies, this study indicated that validity of IELCA examinations could only be achieved through the five linking stages. Also, this study meets with other studies (e.g. Hambleton et al., 2000; Plake & Hambleton, 2001) in signposting the panellists' inability to get the mapping done from the first trial. For instance, in the Familiarisation stage, it was obvious that the panellists pitched some tasks at the wrong CEFR level; however, it was unclear whether such indecisiveness was due to the item per se or to the nature of understanding the CEFR descriptors.

Despite its widespread in many geographical areas, the CEFR needs to be critically evaluated as to its transparent nature and role in defining constructs comprehensively and supporting the panellists, experts, judges, or teachers to carry out the mapping task. For instance, carrying out the Familiarisation and Specification phases effectively was fraught with some challenges the first of which was the panellists' familiarity degree with the use of the different Standard-Setting methods. The second challenge was supplemented by a problematic compendium of a useful selection of the most effective method for the mapping task. Also, the used Standard-Setting methods were perceived to be an amalgam of objective and subjective evaluations that might pose serious problems to the panellists in that reverting to teachers or judges in stage five as part of the external validation task might lead to detrimental effects especially when these experts lack the proper knowledge of the CEFR descriptors or when they are not equipped with the right experience tools to conduct this task without any arbitrariness or bias. The panellists' views of the CEFR mapping were couched in their views of teaching, language learning and assessment, and what was scaled in their mapping task was their

conceptions of the notion of 'threshold'. This is what North (2000) addressed two decades ago. Since then, the approach has not undergone any relevant changes.

Panellists were made aware of the fact that aiming for a perfect consistency was an impossible task to achieve; rather, the alternative approach was to reduce their rating inconsistency (McNamara, 1996). Rater indecisiveness was coupled with some rating inconsistency which was due to other factors such as central tendency or halo effect. Other studies addressed this similar finding on rater (in)consistency (e.g. Engelhard, 2009; McNamara et al., 2019). Raters with fit values greater than 1 indicated more variation than expected, and data provided by raters tended to misfit the model, but raters with fit values less than 1 showed less variation than expected. Also, data provided by these raters tended to overfit the model. As a rule of thumb, Linacre (2002) suggested using .50 as a lower control limit and 1.50 as an upper control limit for infit and outfit mean-square statistics. Other researchers strictly suggested a control of .70 (or .75) as lower limit and 1.30 as an upper limit (Bond & Fox, 2001; McNamara, 1996). A further analysis of rater facet showed that rater severity varied signifcantly, though not large, among the fourteen raters in rating B1-C2 performances. The results also showed other rater effects such as central tendency and halo effect, which suggested the need for further rater familiarisation from B1 to C2 to enhance rater consistency and control central tendency or halo effect.

The many-facet Rasch measurement addressed two sources of variability, rater and criteria domains, in Writing and Speaking scores. The global fit results indicated that the Writing and Speaking tasks fit the three-facet Rasch model of examinee, rater and criteria. The investigation into rater facet showed that rater severity was maintained to the appropriate extent among raters in grading Writing and Speaking. This result seems set to confirm findings in L2 performance rating research that rater training can steadily help increase inter-rater and intra-rater reliability (McNamara, 1996; McNamara et al., 2019). As there were only 293 students in the Academic Reading dataset and 304 students in the General Reading dataset (Table 16), unidimensional IRT rather than the sophisticated bifactor-MIRT was applied to validate the Reading subtest. Particular validating analyses included computing descriptive and reliability statistics, forcing a 1PL-IRT structure to detect local dependence within each section and performing 1PL-IRT to calibrate each section of the six sections. Results of 1PL-IRT showed that, across the six Reading sections, all item discriminations, forced to be equal within each section, and the majority of item difficulty estimates were within the typical range, indicating that, in general, the Reading test sections were measuring the underlying trait of the reading ability adequately.

The validity of the IELCA Listening test was examined using bifactor-multidimensional IRT modelling whose application was meant to ensure that the Listening test was valid in measuring the primary factor of the general Listening ability, regardless of what test method, response format, or test content was involved. Analytical steps involved computing descriptive and reliability statistics, assessing the appropriateness of using bifactor-MIRT (i.e. dimensionality assessment and local dependence detection) and bifactor-MIRT calibrating. The results of reliability statistics showed high reliability in the Listening test as a whole (α = .83). The results of bifactor-MIRT with one general Listening ability plus one or several hypothesised response format factors performed on each section showed that all three sections tapped into the general

Listening factor well. The examinations' items indicated high reliability in reflecting the general Listening ability, which could act as a strong evidence for test validity.

Like the case with other studies, in this study, the judges' familiarity with grading and students' performances with a given benchmark helped teachers reach more objective judgements on the cut-off scores (Figueras et al., 2009; Jones, 2009). The judges in this study used the Standard-Setting qualification to select answers and decisions by following documented steps to align test scores to standards to eliminate any arbitrary and biassed judgement. This result is echoed in other studies (e.g. Manias & McNamara, 2016). However, unlike other studies (e.g. Florez, 2012), having an overall agreement among the panellists on the cut-off scores seems to be debatable. Not only did teacher judgements serve as a means for empirical validation for IELCA, but also as a second, examiner-centred standard-setting method. Findings of teacher estimates confirmed the cut scores suggested as a result of the CEFR Standardisation stage.

Implications and limitations

This study had direct pedagogical, methodological and research implications. The pedagogical implications addressed teachers' effective use of the CEFR descriptors to map IELCA examinations. The proliferation of this strategy use can empower teachers to implement in-class curriculum activities in transparent and coherent ways. Additionally, this mapping task inferred different ways to involve learners and include their needs in classroom-based learning and assessment activities. The research implication accentuated the importance of continuing to use the CEFR to map examinations by addressing all the different mapping stages (CoE, 2009). For instance, the use of the examinee paper selection method in the Standard-Setting stage had its merit in the linking process. Many studies have carried out popular Standard-Setting methods; however, not all of them have addressed the examinee paper selection method despite its validity and reliability value in the linking process. Generally, this method is practical and heuristic in nature and in deciding the test-takers' actual performance since it requires the inclusion of a large number of examinees to have a good representative sample of the distribution of test papers. However, the straightjacket nature of this method is limited since each question is weighted on equally with other items in determining the test-takers' performance. For the IDM method, the challenge rests on showing the test-takers' performance on a scale in the final decision of the cut-off score as the test-takers' chance of success might vary from one item to another. This method is time-consuming for the panellists since they have to link each item to the CEFR descriptors, which may cause some fatigue and adverse and gradual loss of motivation (see Alderson et al., 2006). In addition, the panellists might not have a clue about item difficulty since this method addresses a hypothetical situation of test-takers performing at a given level and getting a score that reflects this ability level (Shepard, Glaser, Linn, & Bohrnstedt, 1993). Ordering the test items in the IDM method as to their ascending difficulty level might impact the target cut-off score for panellists to check the rubric and define the content knowledge and skills for each item ab initio.

Maintaining judges' training and up-to-date familiarisation with the CEFR needs to be attended to continuously since judges sometimes do not share the same rating experience. Unfortunately, the manual does not state a clear experience requirement of the panellists' or judges' rating experience or expertise, nor does it lay the foundation

for the multilingual dimension of some raters when they are involved in selecting the cut-off scores. Therefore, predicting the cut-off scores where the test-takers are expected to get an item correct remains one of the decisive factors that judges might struggle with no matter how experienced and familiar they are with benchmarks.

One of the aims of conducting any mapping task is to define the cut-off score in the Standard-Setting stage (Hambleton et al., 2000) since the candidates 'with scores above the cut score have generally achieved an appropriate performance level and that those with scores below the cut score have not achieved the performance level' (Kane, 2017, p. 13). The Standard-Setting outcomes could be biassed and subjective, especially when they place candidates at the wrong level or band. However, setting up a threshold for the candidates' performance will continue to remain a challenging task for the people carrying out standard setting in particular and the mapping task in general since the performance level that the test-takers are assigned must reflect their performance level in terms of what they can do. This threshold stands at the heart of test validity where interpretations and uses of test scores become decisive for the life of test-takers and other stakeholders (McNamara, 1996).

Recommendations for future research

As part of LRN's commitment to quality assurance, further standard setting and mapping stages were required, using more samples for all subtests, which are done every 3 years. The need for benchmarking against other exam boards, offering similar qualifications with similar test constructs, if any, was also proposed, thus feeding into the test development cycle and qualification review. Despite all these heuristic standard procedures, some research ideas need to be triggered. For instance, in using the IDM method, panellists could verbalise their thoughts of the examination contents and mapping so that comprehensive evaluations can take place, and perhaps discussing the test contents could be an additional step towards exam validation. To counterbalance the situation of selecting the appropriate Standard-Setting method, there should be a selection of a pool of judges with different educational, professional experiences and adequate training in how to grade tasks and a good sample of test items that represent the measured construct, as well as representativeness of the test-takers.

Placing test-takers at a given performance level is challenging for the panellists who should take the necessary action and justify the interpretability and intended uses of the test scores, regardless of the nature of standard setting, be it procedural, internal, or criterion-based. In addition, attaining the internal consistency requires maintaining dependable scoring over different rounds, a heuristic process for examinations that can ultimately avoid construct fuzziness. To attain rater consistency, teachers, as well as internal and external panellists, will always need careful and continuous training workshops in mapping examinations with the appropriate CEFR descriptors, and achieving this goal can ultimately help the panellists revisit the actual sources of indecisiveness. Research should consider the comprehensive nature of the CEFR (Weir, 2005) in checking whether the way panellists understand and use the CEFR descriptors is related to the CEFR per se, the panellists' experience as raters, examiners and judges or the constructs of the different examinations. The CEFR framework will continue to impact users to align examinations and curricula with international benchmarks; however, CEFR users, such as teachers, are called upon to use the framework in a malleable way

so that it leads to objective mapping. Despite these promising results, still there remain some challenges that examination providers should address the first of which is the accuracy and exact matching of the CEFR levels to all examinations. The second challenge hinges upon targeting future studies on how panellists' judgements are made when they are engaged in discussions to select the appropriate cut-off scores for examinations in increasingly changing socio-political geographical areas that are marked by cultural and linguistic diversities.

Conclusion

As noted above, the study addressed the linking stages between the IELCA examinations and the CEFR and how these stages led to efficient, reliable and valid linking results. The linking process, in general, was arduous since it required extensive use of resources over a period of 6 months; nevertheless, it not only up-skilled LRN's team but also served as an integral part of the IELCA test development process. Carrying out the manual's linking stages will continue to remain a challenging and highly contentious task for many stakeholders, and its considerable challenge is tied up with the nature of variance that panellists exemplify when they are engaged in the task, as well as relevance and efficiency of the different mapping stages. However, inconsistencies in the cut-off scores that panellists produced should be seen as a healthy step towards implementing more linking rounds to arrive at a clearer and more general consensus in relation to the requirements of the manual (2009). Through the involvement of examinations, panellists, teachers and test-takers' abilities, this study tried to show that the CEFR, along with its initial linking role, could always be re-hailed for its use in a well-defined triadic framework of teaching, learning and assessment.

Abbreviations

ALTE: Association of Language Testers in Europe; CEFR: Common European Framework of Reference; CoE: Council of Europe; IDM: Item-descriptor matching; IELCA: International English Language Competency Assessment; IELTS: International English Language Test System; IRT: Item Response Theory; LRN: Learning Resource Network

Acknowledgements

I would like to thank the Learning Resource Network for providing data of the study. Special thanks go to the reviewers who edited this manuscript.

Author's contributions

The author contributed to this study. The author(s) read and approved the final manuscript.

Funding

This study received no funds.

Availability of data and materials

Data will not be shared for the confidentiality of the panellists and examinees.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declares no competing interests.

Received: 11 December 2020 Accepted: 5 May 2021
Published online: 14 June 2021

References

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.

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

Athanasiou, A., Constantinou, E. K., Neophytou, M., Nicolaou, A., Sophocleous, S. P., & Yerou, C. (2016). Aligning ESP courses with the common European Framework of Reference for Languages. *Language Learning in Higher Education*, *6*(2), 297–316.

Bechger, T. M., Kuijper, H., & Maris, G. (2009). Standard setting in relation to the Common European Framework of Reference for Languages: The case of the state examination of Dutch as a second language. *Language Assessment Quarterly*, *6*(2), 126–150. https://doi.org/10.1080/15434300802457521.

Blömeke, S., & Gustafsson, J. (2017). Introduction. In S. Blömeke, & J. Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective*, (pp. 1–10). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-50856-6_1.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9781410600127.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, *30*(2), 93–106. https://doi.org/10.1111/j.1745-3984.1993.tb01068.x.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications. https://doi.org/10.4135/9781412985918.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Press Syndicate.

Council of Europe. (2009). Relating language examinations to the common european framework of reference for languages: Learning, teaching, assessment (CEFR).

Engelhard, G. (2009). Evaluating the judgements of standard-setting panellists using Rasch measurement theory. In E. V. Smith, & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models*, (pp. 312–346). JAM Press.

Ferrara, S., Perie, M., & Johnson, E. (2014). Matching the judgmental task with standard setting panellist expertise: The item-descriptor matching method. *Journal of Applied Testing Technology*, *9*(1), 1–20.

Figueras, N., Kaftandjieva, F., & Takala, S. (2013). Relating a reading comprehension test to the CEFR levels: A case of standard setting in practice with focus on judges and items. *Canadian Modern Language Review*, *69*(4), 359–385. https://doi.org/10.3138/cmlr.1723.359.

Figueras, N., North, B., Takala, S., Verhelst, N., & Piet, V. A. (2005). Relating examinations to the common european framework: A manual. *Language Testing*, *22*(3), 261–279. https://doi.org/10.1191/0265532205lt308oa.

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

Fleckenstein, J., Leucht, M., & Köller, O. (2018). Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Language Assessment Quarterly*, *15*(1), 90–101. https://doi.org/10.1080/15434303.2017.1421956.

Florez, I. R. (2012). Examining the validity of the Arizona English Language Learners Assessment cut scores. *Language Policy*, *11*(1), 33–45. https://doi.org/10.1007/s10993-011-9225-4.

Fulcher, G. (2008). Testing times ahead? *Liaison Magazine*, *1*, 20–24.

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, *15*(1), 59–74. https://doi.org/10.1080/15434303.2017.1350685.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*(4), 355–366. https://doi.org/10.1177/01466210022031804.

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.

Hein, S. F., & Skaggs, G. E. (2010). Conceptualizing the classroom of target students: A qualitative investigation of panelists' experiences during standard setting. *Educational Measurement: Issues and Practice*, *29*(2), 36–44. https://doi.org/10.1111/j.1745-3992.2010.00174.x.

Hidri, S. (2020). The IELCA and IELTS exams: A benchmark report. *Journal of Asia TEFL*, *17*(2), 742–749. https://doi.org/10.18823/asiatefl.2020.17.2.33.742.

IELCA (2019). https://www.lrnglobal.org/international-english-language-competency-assessment-ielca/

Jones, N. (2009). A comparative approach to constructing a multilingual proficiency framework: Constraining the role of standard setting. In N. Figueras, & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives*, (pp. 35–43). CITO.

Kane, M. T. (2017). Using empirical results to validate standards. In S. Blömeke, & J. Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective*, (pp. 11–29). Springer. https://doi.org/10.1007/978-3-319-50856-6_2.

Kim, M., & Crossley, S. A. (2020). Exploring the construct validity of the ECCE: Latent structure of a CEFR-based high-intermediate level English language proficiency test. *Language Assessment Quarterly*, *17*(4), 434–457. https://doi.org/10.1080/15434303.2020.1775234.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*, *13*(1), 32–49. https://doi.org/10.1080/15305058.2012.678526.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardised mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2013a). *Facets computer program for many-facet Rasch measurement, version 3.71. 0*. www.winsteps.com.

Linacre, J. M. (2013b). *A user's guide to FACETS Rasch-model computer programs*.

Manias, E., & McNamara, T. (2016). Standard setting in specific-purpose language testing: What can a qualitative study add? *Language Testing*, *33*(2), 235–249. https://doi.org/10.1177/0265532215608411.
McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: The role of measurement*. Oxford University Press.
McNamara, T. F. (1996). *Measuring second language performance*. Longman.
Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives*, (pp. 249–281). Erlbaum.
Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.
Nichols, P., Twing, J., Mueller, C. D., & O'Malley, K. (2010). Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice*, *29*(1), 14–24. https://doi.org/10.1111/j.1745-3992.2009.00166.x.
North, B. (2000). *The development of a common framework scale of language proficiency*. Peter Lang. https://doi.org/10.3726/978-1-4539-1059-7.
North, B. (2004). Relating assessments. In *Insights from the common European framework*, (vol. 77).
North, B. (2014). *The CEFR in practice*. Cambridge University Press.
North, B., Figueras, N., Takala, S., Van Avermaet, P., & Verhelst, N. (2009). Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). In *A Manual. Strasbourg: Council of Europe.*
Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261–282. https://doi.org/10.1177/0265532209349472.
Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. Language Assessment Quarterly, 13(2), 109-123. https://doi.org/10.1080/15434303.2016.1149857
Plake, B. S., & Hambleton, R. K. (2001). The analytic judgement method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 283–312). Lawrence Erlbaum Associates, Inc..
Savski, K. (2020). Local problems and a global solution: examining the recontextualization of CEFR in Thai and Malaysian language policies. Language Policy, 1-21. https://doi.org/10.1007/s10993-019-09539-8
Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. National Academy of Education.
The Association of Language Testers in Europe. (n.d.) The ALTE 'CanDo' project 1992-2002 http://www.alte.org/attachments/files/alte_cando.pdf
Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language testing*, *22*(3), 281–300. https://doi.org/10.1191/0265532205lt309oa.

Publisher's Note