Mapping the CU-TEP to the Common European Framework of Reference (CEFR)

Jirada Wudthayagorn
Chulalongkorn University Language Institute, Bangkok, Thailand
wjirada@chula.ac.th

Abstract

The purpose of this study was to map the Chulalongkorn University Test of English Proficiency, or the CU-TEP, to the Common European Framework of Reference (CEFR) by employing a standard setting methodology. Thirteen experts judged 120 items of the CU-TEP using the Yes/No Angoff technique. The experts decided whether or not a borderline student at A2, B1, B2, and C1 levels would correctly answer each item. They judged the items for three rounds. Expert judgment from the third round shows that the CU-TEP cut-off scores for A2, B1, B2, and C1 levels are 14, 35, 70, and 99, respectively, out of the total of 120 points. The standard deviations of A2, B1, B2, and C1 levels are 4.75, 10.68, 19.57, and 10.11, respectively. The standard errors of judgment are 1.32, 2.96, 5.42, and 2.80, respectively. Once mapped with the CEFR, the CU-TEP scores are now meaningful in that, first, score users would know which CU-TEP score range falls into which particular CEFR level, and, second, score users would also know what test takers can do with the English language with respect to a particular CEFR level. Discussion, recommendation, and limitations of the study will also be presented in this article.

Keywords: Mapping, CU-TEP, CEFR

Problem and Motivation of the Study

In Thailand, although English has a foreign language status, it is not foreign to policy makers, administrators, employers, employees, parents, teachers, and students. English is one of the most important indicators of social, academic, and professional advancement and success. As such, English is a core subject in all levels of curriculum from primary education to higher education. Paradoxically, the desired English language proficiency level of Thai citizens has never been met. For example, the National Institute of Educational Testing Service (NIETS) (2018) reported that the average scores of the English subject in the Ordinary National Educational Test (O-NET) across all levels of basic education have remained low, with students achieving only 30–40 percent of the total test score. By the same token, Wichaiyutphong (2011) mentioned that Thai employees working in an international organization admitted that they could not speak English fluently because of their limited vocabulary, that they had difficulty comprehending unfamiliar accents and pronunciation of their foreign colleagues, and that they encountered communication challenges due to cultural differences.

English education reforms from primary to secondary to higher education are being implemented. The Office of the Basic Education Commission (2014), a department under the Ministry of Education, introduced the Common European Framework of Reference (CEFR) to the basic education system, suggesting that the CEFR be used as a framework for English learning, teaching, and assessment. The important aim of implementing this CEFR policy is to set an achievement benchmark for Thai students, indicating that students graduating from grade 6 should achieve an English proficiency level of at least A1, grade 9 of A2, and grade 12 and

vocational college of B1. In the following year, the Office of the Higher Education Commission, also a department under the Ministry of Education, adopted the CEFR into higher education, suggesting that every higher education institution assess the students' English language proficiency upon graduation, and that such proficiency be aligned with the CEFR or other similar standards (Office of the Higher Education Commission, 2015).

Since then, many questions have arisen, but the most frequently asked is "Which English test is a good test to align results with the CEFR?" The Office of the Basic Education Commission (2014, p. 12) suggests using such standardized tests as the TOEFL iBT, the TOEIC, the IELTS, and the CU-TEP. Among these standardized tests, the CU-TEP is the only locally-developed test that is recommended. While international standardized tests have been mapped to the CEFR—for example, at Educational Testing Service (ETS), Tannenbaum and Wylie (2008) mapped the TOEFL iBT to the CEFR and Tannenbaum and Baron (2011) mapped the TOEFL ITP to the CEFR—to the best of my knowledge, the CU-TEP has never been mapped to any language standards, including the CEFR. Thus, mapping the CU-TEP to the CEFR is considered timely and essential so that CU-TEP scores can be interpreted meaningfully with respect to the CEFR.

Purpose and Scope of the Study

The purpose of this study was to map the CU-TEP to the CEFR. Standard setting was the methodology used in this study. The only CEFR levels included in this study's standard setting were A2, B1, B2, and C1 because the CU-TEP was originally designed for university students whose English proficiency level is expected to range from A2 to C1. Put differently, A1 and C2 levels were not included because A1 is too technically low while C2 is too technically high for the CU-TEP to reflect the English proficiency of university students.

Significance of the Study

As the CEFR is becoming more well-known and more widely used in both basic and higher education in Thailand, the result of this study will be useful for CU-TEP score users, such as administrators, instructors, and students, to understand the meaning of the CU-TEP scores with respect to the CEFR levels.

Literature Review

The Common European Framework of Reference or CEFR (or CEF or CEFRL)

The Common European Framework of Reference is a guideline or reference used in language learning, teaching, and assessment. De Jong (2016) mentioned in his talk at the LTRC 2016 Conference that the CEFR is based on Brian North's thesis in 1986, which focused on activities for continuous communicative assessment. Later on, in 2001, the Council of Europe adopted North's idea and fully developed it with the aim

> "to provide a transparent, coherent and comprehensive basis for the elaboration of language syllabuses and curriculum guidelines, the design of teaching and learning materials, and the assessment of foreign language proficiency. It is used in Europe but also in other continents and is now available in 40 languages."
>
> (http://www.coe.int/t/dg4/linguistic/cadre1_en.asp)

The CEFR has three major levels, which are A (Basic User), B (Independent User), and C (Proficient User). Within each major level are two sub-levels, resulting in a total of six levels,

which are A1, A2, B1, B2, C1, and C2. These six levels constitute the CEFR global scale whose descriptors can be referred to in the CEFR formal publication[1].

The CEFR includes descriptive can-do statements, called "descriptors," that exemplify what a foreign language learner can do at each proficiency level. At the Basic level, learners can use a foreign language to handle everyday activities with "here and now" topics. At the Independent level, learners can expand their territory of foreign language use, for example, they can survive in an area where such foreign language is spoken and can deal with both concrete and abstract topics. At the Proficient level, learners are considered fluent and can use a foreign language flexibly in social, academic, and professional situations.

The CEFR also offers detailed descriptors for the four language skills[2]. For example, at C2 reading, learners "can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning" (Council of Europe, 2001, p. 10). At C1 listening, learners "can understand enough to follow extended speech on abstract and complex topics beyond [their] own field, though [they] may need to confirm occasional details, especially if the accent is unfamiliar" (Council of Europe, 2001, p. 8). At B1 writing, learners "can write personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point [they feel] to be important" (Council of Europe, 2001, p. 19). It must be noted that the *Manual for Language Test Development and Examining* published by the Council of Europe (2011) suggests that the can-do statements are "illustrative" and not meant to be "exhaustive, prescriptive, a definition, a curriculum, [or] a checklist" (p. 13). Proficiency levels, therefore, could and should be recognized by related stakeholders based on the can-do statements.

The CEFR and Thai Education

The CEFR was officially introduced into the Thai education system at basic education level in 2014 and at higher education level in 2015. A great number of movements have been observed since then. For example, in 2014–2015 academic year, approximately 46,000 Thai teachers of English at the basic education level sat for an English examination. It was hoped that this examination would provide baseline data of teachers' English proficiency so that follow-up plans to improve their proficiency level could be implemented. Many universities also have developed their own English tests to assess students' English ability both for admission and graduation purposes.

At Chulalongkorn University, the Chulalongkorn University Test of English Proficiency, or the CU-TEP, was developed as an English proficiency test "to measure [students'] ability to use English for academic purposes in listening, reading, and writing. The test is appropriate for academic admissions both at undergraduate and graduate levels" (Chulalongkorn University Academic Testing Center, 2007).

Standard Setting

Standard setting is a methodology used to map levels of language proficiency with cut-off scores that correspond to the respective proficiency levels (Bejar, 2008). Livingston and Zieky (1982) suggested five essential steps in standard setting—that is, selecting experts, defining borderline knowledge and skills, familiarizing and training experts with the technique chosen for the standard setting activity, collecting expert judgment, and calculating expert judgment to identify cut-off scores.

There are two key issues in the standard setting methodology that deserve further elaboration—experts and techniques.

With regard to experts, the number of experts is directly related to the degree of error of judgment. The greater number of experts, the less error of judgment. However, when taking time and budget into consideration, involving a large number of experts may prove the study financially infeasible. Cizek (2012) suggested that 10 to 15 experts be involved in a standard setting activity, and previous studies mapping standardized tests to the CEFR or similar language standard have involved 13 to 23 experts. For example, as will also be discussed in later sections, Tannenbaum and Wylie (2008) recruited a group of 23 experts to map the TOEFL iBT to the CEFR and another group of 22 experts to map the TOEIC and the TOEIC Bridge to the CEFR; Tannenbaum and Baron (2011) recruited 18 experts to map the TOEFL ITP to the CEFR; Tannenbaum and Baron (2015) worked with 13 experts to map the TOEIC with the Vietnamese National Standard; and, Ativorakun and Wudthayagorn (2018) involved 14 experts in mapping the Srinakharinwirot University – Standardized English Test with the CEFR. Thus, based on expert recommendation and previous studies, the number of experts ranging from 10 to 20 could be considered ideal for a standard setting activity.

In respect of techniques, according to Cizek (2012), there are over 60 techniques that can be used to designate cut-off scores in a standard setting methodology. However, the majority of the standard setting studies mentioned above employed Angoff methods, either modified Angoff or Yes/No Angoff. Angoff methods require the experts to focus on test takers' ability with respect to test items, notably those of multiple-choice format. Therefore, Angoff methods are appropriate for experts who are familiar with characteristics of test takers as well as with test items, especially when the items are of multiple-choice in nature. Other methods, such as Bookmark method, Basket method, and Body of Work method, can also be used depending on type of the test, purpose of the test, and experience of the experts, among other factors. Yet, most importantly, Shin and Lidster (2017) discovered that different standard setting methods yield different results. Therefore, selecting an appropriate technique, while a challenging task, is highly imperative.

Mapping of Major Standardized Tests to the CEFR

The TOEFL iBT, the TOEFL ITP, the TOEIC, the TOEIC Bridge, and the IELTS are major standardized tests used around the world to assess English proficiency of non-native speakers—i.e., speakers whose English is not first language—for academic and professional purposes. These tests were developed many decades before the arrival of the CEFR. For instance, the TOEFL paper-based, which is now the TOEFL ITP, was developed in the 1960s. Later on, when the CEFR became globally prominent, these standardized tests were mapped to the CEFR so that their scores can be recognized meaningfully, and scores across tests can be compared accordingly.

The TOEFL iBT, the TOEFL ITP, the TOEIC, and the TOEIC Bridge tests are developed by Educational Testing Service (ETS). While the TOEFL iBT and the TOEFL ITP focus on academic English found in North American contexts, the TOEIC focuses on English used in the global workplace. The TOEIC Bridge concentrates on English in the global workplace as well, but at a more basic level than that of the TOEIC.

As for skill coverage, the TOEFL iBT covers all four language skills in both separate and integrated manners and has a total score range of 0 to 120. The TOEFL ITP focuses on listening, structure and written expression, and reading. The TOEFL ITP score ranges from 310-677. As

for the TOEIC, the test contains only listening and reading sections, with a score range of 5 to 495 for each section. Two optional TOEIC modules—speaking and writing—were more recently added, and each module has a score range of 0 to 200. Finally, the TOEIC Bridge covers listening and reading skills, with a score range of 10 to 90 for each section.

In terms of mapping the abovementioned standardized tests to the CEFR, Tannenbaum and Wylie (2008) recruited two groups of experts for a standard setting activity. The first group consisted of 23 experts from 16 countries who judged the TOEFL iBT test items and tasks. Another group of experts for the TOEIC and the TOEIC Bridge standard setting activity consisted of 22 experts from 10 countries. For both standard setting activities, the modified Angoff technique was used. The experts judged test items and tasks for multiple rounds. In each round, judgment of experts was discussed and impact data provided so that subjective agreement among experts could be achieved. The final cut-off scores for the TOEFL iBT, the TOEIC, and the TOEIC Bridge with respect to the CEFR levels can be seen in Tables 1, 2, and 3 (adapted from Tannenbaum & Wylie, 2008, p. 37).

Table 1: The TOEFL iBT cut-off scores according to the CEFR levels

CEFR levels	TOEFL iBT sections			
	Writing (max. 30 points)	Speaking (max. 30 points)	Listening (max. 30 points)	Reading (max. 30 points)
A1	-	8	-	-
A2	11	13	-	-
B1	17	19	13	8
B2	21	23	21	22
C1	28	28	26	28
C2	-	-	-	29

As illustrated in Table 1, the TOEFL iBT cut-off scores show imbalanced alignment across language skills. That is, the four skills do not begin and end at the same CEFR level. Writing tasks can identify A2 to C1 levels. Speaking tasks can identify A1 to C1 levels. Listening tasks can capture the narrowest range of proficiency, only from B1 to C1 levels. Finally, reading tasks capture the highest levels of proficiency, starting at B1 and reaching up to C2, and are the only tasks that can identify this highest level of the CEFR. This imbalance of alignment lies on the fact that the TOEFL iBT was not originally designed in parallel to the CEFR. The construct of the TOEFL iBT was based on the original TOEFL framework document written by Jamieson, Jones, Kirsch, Mosenthal, and Taylor (2000), who stated that:

> "The purpose of the TOEFL…test will be to measure the communicative language ability of people whose first language is not English. It will measure examinees' English-language proficiency in situations and tasks reflective of university life…where instruction is conducted in English." (p. 10)

Put differently, test items and tasks in the TOEFL iBT were written independent of the CEFR can-do statements. The process of mapping the TOEFL iBT test items and tasks to the CEFR was carried out later using a standard setting methodology. This is why a balanced mapping from A1 to C2 cannot sensibly be accomplished across language skills.

As for the TOEIC, the original test contains only listening and reading sections. Later on, a separate test covering writing and speaking skills was added as optional modules to test takers. Therefore, in order to measure all four language skills, test takers would need to take two

separate TOEIC tests. The TOEIC cut-off scores in relation to the CEFR levels are illustrated in Table 2.

Table 2: The TOEIC cut-off scores according to the CEFR levels

CEFR levels	TOEIC sections			
	Writing (max. 200 points)	Speaking (max. 200 points)	Listening (max. 495 points)	Reading (max. 495 points)
A1	30	50	60	60
A2	70	90	110	115
B1	120	120	275	275
B2	150	160	400	385
C1	200	200	490	-
C2	-	-	-	-

As seen in Table 2, the writing, speaking, and listening sections of the TOEIC capture test takers' proficiency levels from A1 to C1, while reading seems less difficult, capturing the proficiency levels from A1 to only B2. It is interesting to note that in order to reach C1 in writing and speaking, test takers must obtain full scores of the test, which is a very ambitious feat. On the other hand, the TOEIC test is unable to identify C2 level. Therefore, while some test takers may possess an English proficiency level of C2, their high level of proficiency would not be captured by the TOEIC test.

Similar to the original TOEIC, the TOEIC Bridge consists of listening and reading sections. As the TOEIC Bridge is shorter and cheaper than the TOEIC, test takers can use the TOEIC Bridge as a proxy to recognize their English proficiency prior to taking the full-option TOEIC. The TOEIC Bridge cut-off scores in relation to the CEFR levels are illustrated in Table 3.

Table 3: The TOEIC Bridge cut-off scores according to the CEFR levels

CEFR levels	TOEIC Bridge sections	
	Listening (max. 90 points)	Reading (max. 90 points)
A1	46	46
A2	70	64
B1	86	84

As demonstrated in Table 3, in order to achieve an A1 level, test takers must be able to earn at least 46 out of 90 points for each section, which is about 50 percent of the test paper. Also, in order to reach a B1 level, test takers must be able to earn at least 86 points for listening and 84 points for reading, which is greater than 90 percent of the test paper.

In 2011, Tannenbaum and Baron mapped the TOEFL ITP to the CEFR based on judgment of 18 experts from 14 countries around the world (Tannenbaum & Baron, 2011). The standard setting activity consisted of three rounds of judgment, also employing the modified Angoff technique. Cut-off scores recommended by the experts were first reported in raw scores. These raw scores were converted to scaled scores which were then combined to arrive at the total cut-off scores. As seen in Table 4, the results show that, in total scaled scores, A2 level starts at 337, B1 at 460, B2 at 543, and C1 at 627.

Table 4: The TOEFL ITP cut-off scores according to the CEFR levels

CEFR levels	Total cut-off scores (max. 677 scores)	TOEFL ITP sections		
		Listening (max. 68 points)	Structure and Written Expression (max. 68 points)	Reading (max. 67 points)
A2	337	38	32	31
B1	460	47	43	48
B2	543	54	53	56
C1	627	64	64	63

It is worth noting here that the structure of the CU-TEP is quite similar to that of the TOEFL ITP in that it also includes listening, structure, written expression, and reading. However, the task types in the CU-TEP are not exactly the same as those in the TOEFL ITP. While the TOEFL ITP consists only of multiple-choice items, the CU-TEP contains both multiple-choice items and a cloze test (with multiple choices offered). In the past, CU-TEP scores were equated with TOEFL ITP scores. However, the CU-TEP only reports raw scores, not scaled scores as does the TOEFL ITP.

Another major standardized test, besides those developed by the ETS, that deserves elaboration is the IELTS. The IELTS is jointly owned by the British Council, the International Development Program (IDP Education, Australia), and Cambridge Assessment English. It aims to measure English language proficiency of those who want to study or work in locations where English is used as a medium of communication. The IELTS's scaled scores are reported in nine bands, from Band 1 (Non-user) to Band 9 (Expert user). The IELTS bands also show an estimate reference to the CEFR levels, but no clear cut-off scores are given. The rationale for not setting definite cut-off scores could stem from the viewpoint of IELTS test developers, whose grounds are clearly stated in the official IELTS website that

> "the CEFR was not designed to provide the basis for precise equating, nor was it intended to be a prescriptive tool to impose standardised solutions. Rather it was designed as a common framework of reference, primarily intended as 'a tool for reflection, communications and empowerment', as described by John Trim, its co-ordinating author."
>
> (https://www.ielts.org/ielts-for-organisations/ common-european-framework)

As such, the IELTS bands in relation to the CEFR levels are an approximation, with bands 4 to 9 relatively aligned to CEFR's A2 to C2 levels, as demonstrated in Figure 1.

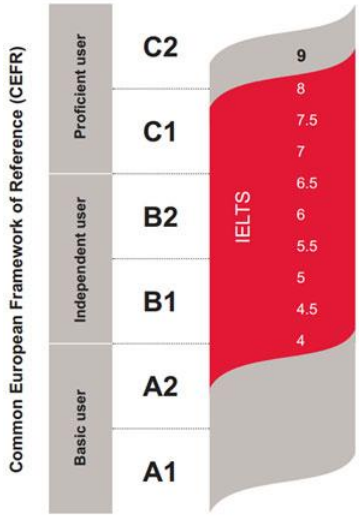

Figure 1: The IELTS bands with respect to the CEFR levels
Source: https://www.ielts.org/ielts-for-organisations/common-european-framework

Lastly, an English standardized test developed by a Thai university—called the Srinakharinwirot University – Standardized English Test (SWU-SET)—has recently been mapped to the CEFR by Ativorakun and Wudthayagorn (2018). The SWU-SET was developed by the Language Centre, International College for Sustainability Studies, Srinakharinwirot University, with an aim to assess university students' English language proficiency. For its standard setting activity, 14 lecturers of English at Srinakharinwirot University acted as experts and judged 100 multiple-choice test items against the CEFR's A2, B1, and B2 levels for three rounds. The modified Angoff technique was used to determine the cut-off scores of the three CEFR levels. The results show that test takers need to earn, out of 100, at least 22 points in order to achieve A2 level, at least 50 points for B1 level, and at least 78 points for B2 level.

In sum, it can be seen that the aforementioned standardized tests are similar in that they aim to measure language proficiency of non-native English speakers who intend or have the need to use English to communicate with both native and non-native English speakers in various academic and professional settings. With the exception of the SWU-SET, these standardized tests were developed and have been widely used long before the development and prominence of the CEFR. However, after the CEFR became well-recognized and its use widespread, researchers in various test-developing institutions made efforts to map their standardized tests to the CEFR, using standard setting, to establish cut-off scores or associate test score bands to the CEFR levels. As a result, test takers' English proficiency reported across standardized tests can be consistently described, understood, and interpreted using the same framework of reference, which is the CEFR.

As the CEFR has been set as part of an English language policy in Thailand, in both basic and higher education levels, mapping a well-established local standardized test such as the CU-TEP to the CEFR is an essential task. This timely issue is crucial because the result of the mapping will be useful for all stakeholders who use CU-TEP scores to make high-stakes decisions, such as placement, admission, and graduation.

Research Methodology

The CU-TEP

Upon submission and approval of formal request, one CU-TEP test form was selected and delivered to the researcher of this study by Chulalongkorn University Academic Testing Center. Prior to the mapping process, some items in this test form were revised by CU-TEP item writers, using Classical Theory to perform item analysis before revision, so as to ensure that all items in the form were suitable for assessing test takers' English language ability. Nonetheless, the structure of the test remained the same, comprising three sections with 120 items in total: 30 items in the listening section, 60 items in the reading section, and 30 items in the writing section. All items are in a 4-option multiple-choice format. The revised test was then piloted, which is an important validation process to ensure that all items were qualified for the mapping purpose[3].

Experts

Experts participated in this study were 13 experts, who were Thai lecturers of English (12 females, 1 male), working as full-time faculty members at Chulalongkorn University Language Institute (CULI). At the time of the study, the experts held at least a master's degree in the field related to English language teaching, and their teaching experience ranged from seven to 40 years, with a mean of 17 years and a standard deviation of 8.544 years. All experts had taught both undergraduate and graduate students at Chulalongkorn University and had been involved in English learning activities for students at CULI's Self-Access Learning Center. Thus, they were all familiar with Chulalongkorn University students.

Yes/No Angoff Technique

Yes/No Angoff was the standard setting technique used in this study. This technique is practical for standard setting that involves many cut-off score levels (Tannenbaum & Baron 2015). Therefore, it was deemed an appropriate technique for this study, as the experts were required to establish four cut-off scores for the CU-TEP, at the levels of A2, B1, B2, and C1.

In the Yes/No Angoff technique, each expert decides whether a single borderline student would (Yes) or would not (No) answer a particular item correctly (Mellone & Faben 2014; Tannenbaum & Baron 2015). Each expert in this study was thus asked to think about a single borderline student based on his or her teaching experience and judge whether or not that borderline student would or would not correctly answer each CU-TEP test item. The questions asked of each expert were:

- Would or would not an A2 borderline student correctly answer item X?
- Would or would not a B1 borderline student correctly answer item X?
- Would or would not a B2 borderline student correctly answer item X?
- Would or would not a C1 borderline student correctly answer item X?

The experts would then engage in such judgment for all 120 items of the CU-TEP.

Pre-Meeting Activity The revised and validated CU-TEP test form and the *Structured overview of all CEFR scales* document (Council of Europe, 2001) were distributed to each expert prior to the date of standard setting activity so that the experts would have sufficient time to review the CU-TEP test items and familiarize themselves with the CEFR scales. Referring to the *Structured overview of all CEFR scales* document (Council of Europe, 2001), the relevant scales that were used in this study included:

- Global scale (p. 5);
- Overall listening comprehension scale (p. 8);
- Understanding interaction between native speaker scale (p. 8);
- Listening as a member of a live audience scale (p. 9);
- Overall reading comprehension scale (p. 10);
- Reading for orientation scale (p. 11);
- Reading for information & argument scale (p. 11);
- General linguistic range scale (p. 27);
- Vocabulary range scale (p. 27);
- Grammatical accuracy scale (p. 28); and,
- Vocabulary control scale (p. 28).

Meeting Activity

The meeting activity for this study's standard setting took three days. On the first day, experts were given an orientation on the importance of the study, the standard setting procedure, and the Yes/No Angoff technique. Then, the experts were trained to use the Yes/No Angoff technique to judge samples of non-CU-TEP test items so as to familiarize them with the standard setting process as well as the Yes/No Angoff technique. After the experts were comfortable with both the process and the technique, the first round of expert judgment on the CU-TEP test items began. The second round of judgment was conducted on the following day. Then, the third, which was final, round of judgment was conducted one week later.

During the activity, each expert was given a form to record their judgment. After considering whether or not, for example, an A2 borderline student would answer Item 1 correctly, the expert would fill in the form by writing number 1 for "Yes" or number 0 for "No" in the box for Item 1. They would ask themselves such question and make such judgment for all 120 items of the CU-TEP. Once finished, they would repeat the same questioning and judging process for a B1 borderline student, then a B2, and then a C1, respectively. This was then considered a completion of Round 1 activity (day 1). The whole process was repeated for Round 2 (day 2) and Round 3 (day 3). A sample compilation of expert judgment based on an A2 borderline student is given in Figure 2.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item …	Item 118	Item 119	Item 120
Expert 1	0	0	0	1	1	0	0	0	0	1
Expert 2	0	0	0	0	0	0	0	0	0	1
Expert 3	0	0	0	0	1	0	0	0	0	0
Expert 4	0	0	0	1	0	0	0	0	0	0
Expert 5	0	0	0	1	0	0	0	0	0	1

Figure 2: A sample compilation of expert judgment based on an A2 borderline student

At the end of each round, results from all expert judgment were calculated using descriptive statistics—that is, minimum (min), maximum (max), average (mean), and standard deviation (SD). The standard error of judgment (SEJ) was also calculated using the Central Limit Theorem based on MacCann and Stanley (2004).

Post-Meeting Activity

To ensure validity of the standard setting process, five evaluation forms were given to the participating experts to complete. The evaluation forms were adapted from Kollias (2013) and were distributed at different times during the three-day activity, namely, at orientation, after the training session, at end of Round 1, at end of Round 2, and at end of Round 3 for final evaluation. In overall, the experts felt confident with their judgment and were satisfied with the standard setting activity.

Results

This section discusses the CU-TEP cut-off scores with respect to CEFR's A2, B1, B2, and C1 levels from the three rounds of participating experts' judgment.

Round 1

Table 5: Descriptive statistics of Round 1 judgment

	CU-TEP statistics for each CEFR level			
	A2	B1	B2	C1
Min	3	22	49	90
Max	19	53	99	112
Mean	12.46	39.62	80.38	102.92
SD	4.52	10.06	15.89	6.46
SEJ	1.25	2.79	4.40	1.79

Table 5 presents the descriptive statistics of expert judgment in Round 1. The mean cut-off scores for A2, B1, B2, and C1 are rounded to 12, 40, 80, and 103, respectively. The smallest standard deviation is at A2, which means that the experts seemed to highly agree on the cut-off score for this level. On the other hand, the largest standard deviation is at B2, indicating that the experts seemed to have different opinions for B2 cut-off score. The standard error of judgment (SEJ) also shares the same pattern as the standard deviation, that is, the smallest SEJ is at A2 and the largest at B2. Note that, in Round 1, the experts judged each item based on their own experience of encounters with their own students. No impact data—such as the difficulty index of test items—were given. Nonetheless, based on the response in the evaluation form of all experts, the majority agreed with Round 1 judgment.

Round 2

Table 6: Descriptive statistics of Round 2 judgment

	CU-TEP statistics for each CEFR level			
	A2	B1	B2	C1
Min	3	13	38	68
Max	19	50	87	109
Mean	12.46	36.76	70.84	99
SD	4.52	12.97	16.72	11.71
SEJ	1.25	3.59	4.63	3.24

The descriptive statistics of Round 2 judgment is shown in Table 6. The mean cut-off scores for A2, B1, B2, and C1 are rounded to 12, 37, 71, and 99, respectively. In Round 2, cut-off scores of B1, B2, and C1 decreased, while cut-off score of A2 remained the same. However, it is interesting to note that the standard deviations of B1, B2, and C1 increased. Likewise, the

SEJs of these levels also increased. In general, the experts lowered cut-off scores for Round 2, but larger standard deviations in this round, compared to Round 1, indicate more disagreements among experts.

After Round 2, the impact data—that is, the difficulty index of test items—were given to the experts. They then discussed each test item based on the impact data. For example, if a particular item appeared easy, the experts discussed to reconsider if a borderline student at a particular CEFR level would or would not correctly answer that item. They were allowed to change their decision based on the discussion, which then led to Round 3 judgment.

Round 3

Table 7: Descriptive statistics of Round 3 judgment

	CU-TEP statistics for each CEFR level			
	A2	B1	B2	C1
Min	8	18	36	74
Max	28	45	105	116
Mean	13.62	34.54	70.49	98.74
SD	4.75	10.68	19.57	10.11
SEJ	1.32	2.96	5.42	2.80

Table 7 shows the descriptive statistics of Round 3 judgment. The mean cut-off scores for A2, B1, B2, and C1 are rounded to 14, 35, 70, and 99, respectively. In Round 3, which is the final round of judgment, A2 cut-off score is greater than that of Round 2. In contrast, B1 and B2 cut-off scores are lower than those of Round 2, while C1 cut-off score remains the same. The highest standard deviation is at B2, as is the highest SEJ. The CU-TEP cut-off scores formally reported in this study are based on the experts' judgment in this final round.

Discussion and Recommendations

Range of Cut-Off Scores and CEFR Descriptors

In this study, a standard setting activity involving 13 experts was carried out to map the CU-TEP to the CEFR. The Yes/No Angoff technique was used to ask each expert whether a single borderline student at different CEFR levels would or would not correctly answer each CU-TEP test item. They made such judgment item-by-item for each of the CEFR level relevant in this study—which are A2, B1, B2, and C1 of the CEFR global scale—for a total of three rounds. The range of CU-TEP cut-off scores with respect to the CEFR levels, which are based on rounded mean scores obtained from the final round of expert judgment, is shown in Table 8.

Table 8: The CU-TEP cut-off score ranges with respect to the CEFR levels

CU-TEP cut-off score ranges (max. 120 points)	CEFR levels
14 – 34	A2
35 – 69	B1
70 – 98	B2
99 – 120	C1

As seen in Table 8, the widest range of CU-TEP scores in relation to the CEFR level is at B1 (35 points), followed by B2 (29 points), C1 (22 points), and A2 (21 points). This inconsistent range suggests that in order for test takers to move from one level to the next, they need to

expend varying degrees of efforts and, most probably as a result, time. That is, in order to move from A2 to B1, test takers need to accomplish at most 21 points. In contrast, in order to move from B1 to B2, test takers need to overcome up to 35 points, which could be a challenging hurdle as that number of points constitute over one-third of the total test scores. Future research may need to focus on specific proficiency level(s) and investigate how to move test takers from one level to the next, such as documenting hours of test preparation at each level.

Using CU-TEP Cut-Off Scores

After being mapped, the cut-off scores of the CU-TEP now carry meaning with respect to the CEFR levels. Stakeholders who will be using CU-TEP scores can now interpret that, for example, if a student has a CU-TEP score of 60, this student is considered to have an English proficiency equivalent to the CEFR level of B1, and, based on the CEFR global scale, this student can

- *understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.;*
- *deal with most situations likely to arise whilst travelling in an area where the language is spoken;*
- *produce simple connected text on topics which are familiar or of personal interest; and,*
- *describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.* (Council of Europe, 2001, p. 5)

Other CU-TEP score users, such as school administrators or teachers, can also use this information on cut-off scores along with related CEFR levels and descriptors for such matters as setting admission or graduation policy, designing or revising English language curriculum, or planning classroom lessons and activities. This is because once they can identify students' current English proficiency level, they can make informed decisions based on the status quo or head into the direction toward improving students' language ability. To interpret the meaning of CU-TEP scores based on the CEFR levels, score users can refer to the CEFR global scale descriptors found in the formal CEFR publication by the Council of Europe[4].

Dealing with False Positive and False Negative Results

By using the cut-off scores, two possibilities can be observed. First, a test taker who belongs to a lower proficiency level may obtain a score above the cut-off score of his or her actual level, resulting in a false positive of reported higher proficiency than that of reality. Another possibility is that, in contrast to the first possibility, a test taker who belongs to a higher proficiency level may obtain a score below the cut-off score of his or her actual level, resulting a false negative of reported lower proficiency than that of reality. For instance, B1 test taker may be identified by the test as a B2 (false positive), or a B2 test taker may be identified as a B1 (false negative). These errors may not be critical under low-stakes circumstances, such as placement of students in different class sections. However, when it comes to high-stakes decisions, such as granting of degrees to graduating students or admission of new company recruits, the consequences can be serious and even damaging to both the test takers and the decision makers.

Livingston and Zieky (1982) explained that no test is ever completely perfect to measure what it is aimed to measure, thus, for the majority of tests, it is not possible to arrive at cut-off

scores that are completely free of error of judgment. As such, a certain extent of error of judgment is always present and can lead to the two aforementioned possibilities. What is needed for further research, then, is the documentation of misplaced test takers. This can be done by triangulating data related to such test takers obtained from various sources, such as interviewing teachers or supervisors about the test takers' English proficiency level, asking the test takers to do a self-assessment, or reviewing the test takers' academic records. Furthermore, continuous improvement of the test, such as the CU-TEP, is crucial so as to minimize the error of judgment. This can be done through making the construct of the CU-TEP more representative of the target language use constructs and validating the test items before actual test administration.

Standard Errors of Judgment in Standard Setting

Standard setting calls for, and involves, subjective agreement among experts regarding cut-off scores (Cizek, 2012). In fact, it is important to note that objective agreement is impossible to reach because each expert brings his or her own experience of encounters with borderline students into the discussion. Yet, expert agreement can be gleaned from the size of standard errors of judgment of cut-off scores. For all judgment rounds in this study, the standard errors of judgment of A2 and C1 levels are the smallest and the second smallest, respectively.

From this, it can be interpreted that the experts in this study seemed to agree upon whether a borderline A2 student and a borderline C1 student would or would not correctly answer the test items. Larger standard errors of judgment can be observed at B1 and B2 levels, which means the experts had different perceptions on the ability of a borderline student at B1 and B2 levels. Yet, in overall, it can be observed that standard error of judgment of cut-off scores are relatively small in this study, signifying a relatively high level of agreement among participating experts.

Nonetheless, the CEFR is not designed to provide a clear-cut boundary of each proficiency level. While B1–B2 levels are in the middle, A1–A2 and C1–C2 levels are at the far end of the spectrum, albeit on a different side. This means that, Basic and Proficient language users can be easily identified, as their proficiency falls on a definite extreme of the spectrum. However, identifying Independent language users whose proficiency falls along the range of the spectrum is not a straightforward task, hence expert judgment can deviate. Therefore, decisions made for cut-off scores of B1 and B2 levels may not be consistent across experts, and this can be observed through the standards error of judgment. This circumstance is also evident in the current study, as the standards error of judgment for B2 are the highest, and B1 the second highest, for all three rounds.

Familiarization with Standard Setting Process and CEFR Descriptors

It is suggested that familiarization with the standard setting process as well as with the language standard used for mapping reference—in this case, the CEFR descriptors—could be a critical factor to help minimize errors of judgment (e.g., Cizek 2012; Takala & Kollias 2015). As experts in this study had never mapped a test to the CEFR, these experts, through comments given in the evaluation forms, stated that pre-meeting and training activities were proved useful. In the pre-meeting activity, they were assigned to study the CEFR and the CU-TEP. Before mapping, the experts were trained to do so. It was done first on the first day in the meeting activity. Then, they evaluated themselves if they were ready to move on. Future research is needed at this point to investigate the nature of pre-meeting and training activities that help experts to be better familiar with standard setting process and CEFR descriptors.

Limitations of the Study

For this particular standard setting study, two main limitations emerged as follows:

First, the construct of the CU-TEP is underrepresented, as it contains items that assess only receptive skills. Even though there is a section on writing, the test items in such section come in a form of error identification. Thus, test takers read and select the answers based mostly on their linguistic competence (i.e., grammar and vocabulary knowledge). This is not considered a direct measure of writing skills, as test takers are not asked to provide actual writing samples. This also means that, when the experts in this study had to map the writing test items to the CEFR, they had to base their judgment on other "proxy" scales, such as the overall reading comprehension scale and the general linguistic range scale, among others, as opposed to on actual writing-related scales.

Second, the CEFR descriptors are illustrative, not definitive, meaning that it can be interpreted differently by different experts. For example, the descriptors of the A1 level state that a language user at this level *"can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type"* (Council of Europe, 2001, p. 5). In reading such description, different interpretations may arise, such as what types of phrases are considered "very basic." Similarly, at B1 level, the descriptors state that a language user *"can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc"* (Council of Europe, 2001, p. 5). The interpretation of "familiar matters" based on each expert's experience may vary. Last but not least, at the C1 level, the descriptors state that a language user *"can understand a wide range of demanding, longer texts"* (Council of Europe, 2001, p. 5). Again, the interpretation of "longer" is unclear and may vary among experts as well. Undoubtedly, different interpretations can lead to deviation in judgment of test items and cut-off scores. Thus, discussion among experts in the standard setting process is much encouraged so that understanding and interpretation of the CEFR descriptors would be consistent, leading to a more valid and less error-prone judgment.

Acknowledgements

This study was fully funded by the Learning Innovation Center of Chulalongkorn University. I would like to thank the former Director of the Learning Innovation Center, Mrs. Prapaipis Mongkolratana, who truly understood the importance of this study and made it possible through generous time and financial resources. I would also like to thank the 13 experts who made this study possible—Boonsiri Anantaset, Samertip Kanchanachari, Chatraporn Piamsai, Sutthirak Sapsirin, Pimpan Syamananda, Tanyaporn Arya, Chuloporn Kongkeo, Pajaree Nipaspong, Woralan Kongpolphrom, Parima Kampookaew, Boonyakorn Siengsanoh, Narisa Jitpraneechai, and Wutthipong Laoriandee. All mistakes, however, are mine.

Notes

[1] Visit https://rm.coe.int/1680459f97 for more detail of the CEFR global scale.
[2] See the *Structured overview of all CEFR scales* document at https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168045b15e for detailed descriptors of the four language skills.
[3] The CU-TEP item revision and the standard setting activity were conducted in two different phases. Piamsai (2016) was in charge of the CU-TEP item revision phase. The revised CU-TEP form was then used to map with the CEFR in the subsequent standard setting phase, conducted in 2016 by Wudthayagorn. The standard setting results were presented as research report for internal use.
[4] Visit https://rm.coe.int/1680459f97 for more detail of the CEFR global scale.

About the Author

Jirada Wudthayagorn received a full scholarship from the Royal Thai Government to complete her Ph.D. in Applied Linguistics from The University of Pittsburgh. She is now a full time lecturer of English at Chulalongkorn University Language Institute teaching Experiential English, ESP, and language assessment courses. She is also the First Vice President of Asian Association of Language Assessment (AALA). Her research interests cover language assessment, language policy, and quantitative analysis. She can be reached at jirada.w@chula.ac.th

References

Ativorakun, C. & Wudthayagorn, J. (2018). Mapping Srinakharinwirot University – Standardized English Test (SWU-SET) onto the Common European Framework of Reference (CEFR). *Suranaree Journal of Social Science, 12*(2), 69-84.

Bejar, I. (2008). *Standard setting: What is it? Why is it important?* R&D Connections. Princeton, NJ: ETS.

Chulalongkorn University Academic Testing Center. (2007). *CU-TEP*. Available at http://www.atc.chula.ac.th/en_html/en_tep.html.

Cizek, G. J. (2012). *The forms and functions of evaluations in the standard setting process*. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations*. (pp. 165-178). New York: Routledge.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Structured overview of all CEFR scales*. Available at https://rm.coe.int/168045b15e.

Council of Europe. (2011). *Manual for language test development and examining: For use with the CEFR.* Available at https://rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-the-ce/1680667a2b.

De Jong, J. H. A. L. (2016). Modelling language competence into a global framework: Taking the past into the future. *Talk presented at the LTRC 2016 Conference, Palermo, Italy*.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). TOEFL 2000 Framework: A working paper. *TOEFL monograph series*. Princeton, NJ: ETS.

Kollias, C. (2013). *Collecting procedural evidence through comprehensive evaluation survey forms of panelists' impressions*. Hellenic American University, PowerPoint presentation at Pre-conference Workshop at the EALTA 2015 Conference, Copenhagen, Denmark.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Manuscript. Princeton, NJ: ETS. Available at https://www.ets.org/Media/Research/pdf/passing_scores.pdf.

MacCann, R. G. & Stanley, G. (2004). Estimating the standard error of the judging in a modified-Angoff standards setting procedure. *Practical Assessment, Research & Evaluation, 9*(5), 1-9.

Mellone, I. & Faben, C. (2014). *Are they mission ready? Using the modified Angoff method to set cut scores.* Camber Corporation Orlando, Florida.

National Institute of Educational Testing Service (NIETS). (2018). *O-NET score summary report of grade 12 students in academic year 2017*. Available at http://www.newonetresult.niets.or.th/AnnouncementWeb/PDF/SummaryONETM6_2560.pdf.

North, B. (1986). *Activities for continuous communicative assessment*. (unpublished MA thesis). University of Birmingham.

Office of the Basic Education Commission. (2014). *Guidelines for practices by Ministry of Education: English education policy reform*. Office of the Basic Education Commission, Ministry of Education. Available at http://old.drs.ac.th/ext/tch_data/tch_ 02.pdf.

Office of the Higher Education Commission. (2015). *Policy of upgrading English education standards of higher education institutions.* Office of the Higher Education Commission, Ministry of Education. Available at http://www.mua.go.th/users/bhes/front_home/Data%20Bhes_2559/04052559.pdf.

Piamsai, C. (2016). *Mapping the CU-TEP to the CEFR: A research report phase 1.* Learning Innovation Center, Chulalongkorn University. (in Thai)

Shin, S.-Y. & Lidster, R. (2014). Evaluating different standard-setting methods in an ESL placement test context. *Language Testing, 34*(3), 357-386.

Takala, S. & Kollias, C. (2015). *Standard setting – how to implement good practice*. Pre-conference Workshop at the EALTA 2015 Conference, Copenhagen, Denmark.

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology. *TOEFL iBT research report.* Princeton, NJ: ETS.

Tannenbaum, R. J. & Baron, P. A. (2011). Mapping TOEFL® ITP scores onto the Common European Framework of Reference. *Research memorandum*. Princeton, NJ: ETS.

Tannenbaum, R. J., & Baron, P. A. (2015). Mapping TOEIC® scores to the Vietnamese National Standard: A study to recommend English language requirements for admissions into and graduation from Vietnamese universities. *Research memorandum.* Princeton, NJ: ETS.

The Nation. (2015). *English teachers face test*. Available at http://www.nationmultimedia.com/national/English-teachers-face-test-30255137.html.

Wichaiyutphong, K. (2011). *English barriers for Thai employees in an international setting: A study at Thomson Reuters Company in Thailand*. Research paper. Language Institute, Thammasat University.