PASAA Journal
Special Issue 2025/2026, 128–165

Aligning the CULI Test with CEFR Illustrative Scales

Sasithorn Limgomolvilas

Chulalongkorn University Language Institute, Bangkok, Thailand
Corresponding author: Sasithorn.li@chula.ac.th

Article information	
Abstract	The Chulalongkorn University Language Institute (CULI) Test was first developed in 2006, which was before Thailand required language tests to be aligned with the Common European Framework of Reference (CEFR). Aligning the CULI Test, a local test, to the CEFR, however, can offer stakeholders a more reliable indicator of test-takers' proficiency level. This study sought to align each item on the CULI Test with the CEFR communicative language competence scales through quantitative and qualitative methods utilizing eight expert judgments and Rasch measurements. The results showed that most of the items were between the B1 and B2 level; other items were at the A2 and C1 level. Qualitative analysis revealed that the items included 8 out of 42 specific activities and strategies highlighted in the CEFR illustrative scales. Reading for information and argument was the most common question type, accounting for 24 of 100 items, followed by overall written comprehension with 20 items, while the least common type was overall oral comprehension with five items. The results indicate that the CULI test is mainly suitable for differentiating test takers at two CEFR levels: B1 and B2. This raises two options for CULI Test developers to consider: adding additional items at the A2 and C1 levels to improve the test's capacity to

E-ISSN: 2287-0024

	differentiate among a wider range of CEFR levels or redesigning the test as a specific proficiency test of the B1 or B2 level.
Keywords	CEFR, standard setting, illustrative scales, large-scale assessment
APA citation:	Limgomolvilas, S. (2025/2026). Aligning the CULI test with CEFR illustrative scales [Special Issue]. *PASAA, 73*, 128–165.

1. Introduction

At present, as many English proficiency tests are readily available for test takers to choose from, various issues must be considered when choosing which type of test to take. Among the internationally recognized tests are IELTS (International English Language Testing System), TOEIC (Test of English for International Communication), and TOEFL (Test of English as a Foreign Language). It can undoubtedly be said that not every test will match a particular test taker's objective. Consideration over which test to take can include a test taker's objectives, test information, the availability of the test, and the criteria used by the test in determining language proficiency. The details and purposes for using the test scores are also necessary for a test taker to compare and decide, and thus, all test providers should ensure that all relevant information is made publicly available to ensure transparency. Furthermore, test takers need to consider the duration, practice time for the test, and scheduling of their selected test (Jabeen et al., 2025). In addition, it is important to consider whether the test is based on national or international standards. The sheer number of factors to take into account can make the selection of a suitable test a highly stressful task.

Meeting all the conditions above would be of great value to test takers in selecting an appropriate test for the right purpose with adequate preparation. One of the most well-known sets of standards for ascertaining language proficiency is the Common European Framework of Reference (CEFR), which has had a major influence on the field of second language education and assessment (Baharum et al., 2021; Papageorgiou et al., 2015). The CEFR is a framework for language

proficiency developed by the Council of Europe (2001), which can be adapted for all languages and many purposes. The CEFR is divided into 6 levels, ranging from the lowest level of language proficiency, A1 or beginner, all the way up to C2, which is the most fluent. These levels are regularly used as a reference for language tests to set cut scores (Eckes, 2015). Whether the test is IELTS, TOEIC, or TOEFL, these internationalized tests have mainly aligned their test takers' scores to CEFR levels within the range of B1-C1.

Following these standards, Thailand's Office of the Basic Education Commission has set the CEFR as the reference criteria for the language proficiency standards of Thai students since 2014 (Wudthayagorn, 2018). This CEFR standard has been applied not only to high school level but also to undergraduate level, as the Ministry of Education has aimed that all students should reach B1 and B2 level respectively by the time they graduate (Anantapol et al., 2018).

This research examined a test developed by Chulalongkorn University Language Institute (CULI), known as the CULI Test, which is intended to be used as a standardized indicator of English language proficiency for the workplace. Currently, CULI Test results, which include a general description of English proficiency level, can be submitted to some workplaces for consideration of promotions and salary increases. However, the CULI test is not administered widely, but solely at Chulalongkorn University. One of the explanations could stem from the fact that the test is not aligned with the CEFR, which can allow organizations to standardize employee assessment results. This research responds to the need identified by Anantapol et al. (2018) to align the competence of test takers to international standards such as the CEFR. Initially, the CULI Test was developed for local purposes as a proficiency test, which differentiated the test takers into different English proficiency levels based on TOEIC score comparisons of the test takers through concurrent validity. Nevertheless, the proficiency levels of the test can be defined in more direct and clearer detail to relate to international frameworks by mapping the levels to a global scale. Thus,

this study aimed to align the test to CEFR scales of communicative language competence by analyzing the item difficulties of the CULI Test against CEFR levels using expert judgements together with a Rasch measurement model.

2. Literature Review

Many international standardized tests are aligned with the CEFR, such as the IELTS, TOEIC, Cambridge English exams, and Pearson Test of English (PTE), as shown in Table 1. Some tests are specific in defining the level according to the skill being measured, such as Speaking, Writing, Listening, and Reading. With the exception of the PTE, most of these tests do not claim to evaluate test takers' performance with reference to the full range of the CEFR levels. However, all tests claim to determine proficiency within, at least, the B1-C1 range. This is important, as many academic institutions require at least a B2 level of proficiency for admission (Green, 2017; Papageorgiou et al., 2015). For example, a study by Tangsakul and Poonpon (2024) found that their university's Academic English Language Test for undergraduate students included a majority of test items between the B1-B2 levels. Accordingly, this study did not consider it necessary for the CULI Test to differentiate performance at the A1 level, as test-takers with this proficiency level are not the target group of stakeholders.

Table 1

Tests Compared to CEFR Levels

Test	CEFR levels	Objectives
TOEIC	A1 – C1	workplace English proficiency
IELTS	A2 – C2	academic and general English
PTE	A1 – C2	academic English and migration
TOEFL	B1 – C2	academic English

2.1 CULI Test

The CULI Test was developed in 2006 to test English proficiency for occupational and international communication purposes. Unlike the Chulalongkorn

University Test of English Proficiency (CU-TEP), which tests general and academic English, the CULI Test is geared toward more specific usage of English for work and international communication. A total of 100 test items are included, with 50 items testing listening skills and 50 items testing reading and writing skills, with each section divided into subsections as shown in Table 2. The time allowed for the test is 100 minutes, with 40 minutes allocated to the first part and 60 minutes to the second.

Table 2

CULI Test Parts and Items

Part	Subsection	No. of item
Listening	Photographs	10
	Question - Responses	10
	News and Announcements	10
	Short Conversations	10
	Short Talks	10
Writing and Reading	Sentence Completion	20
	Text Completion	6
	Single Passage	14
	Double Passages	10

Although the CULI Test has been used for 19 years with some minor changes, while retaining the number of items at 100, examining the test format through comparison with CEFR should help to modernize the test to an international standard. Some standardized international tests, like TOEFL, have made some changes to their test from paper-based test to computer based, with some changes to the test format. Additionally, when the CEFR was announced, further studies (Education Testing Service, 2020) were conducted to align the test to its guidelines.

In a similar vein, the CULI Test warrants a thorough analysis to evaluate its alignment with internationally recognized standards, thereby ensuring its validity and relevance in a global context. Aligning the test to a widely recognized scale can enhance its applicability to the can-do statements found in the CEFR descriptors and can facilitate the process of test improvement and development.

2.2 Mapping Tests with CEFR

Martyniuk (2010) recommends four stages in mapping a test to the CEFR, which are familiarization, specification, standardization, and empirical validation. First, in the familiarization stage, test developers are familiarized with the CEFR to ensure that they have a comprehensive understanding of its levels and descriptors. Second, in the specification stage, the content of the test is compared to the levels mentioned in the CEFR. Next, in the standardization stage, the test undergoes expert judgment to determine the extent to which it aligns with CEFR levels. The final stage is empirical validation, in which the data and the ratings are collected and analyzed to prove that the test and the relationship to CEFR levels are reliable.

Although it is best practice to conduct all four of the above stages, this may not always be practical or necessary, especially if the test is relatively low stakes. In such cases, it is deemed acceptable to emphasize two stages: specification and standardization (Martyniuk, 2010). When considering these two procedures, it is worth comparing them to a simpler and faster method of standardization through expert judgement with support of Many-Facet Rasch Measurement. This method was claimed by Eckes (2015) to achieve similar goals of setting standards for several types of assessments. This method primarily relies on a group of experts rating the level of performance required to achieve each item on a provided proficiency scale (Eckes, 2015); thus, recruiting the right experts is a crucial step. These experts need to provide an evaluation of an “absolute” performance level based on their knowledge or opinion about the overall skills or ability levels of the group being examined (Cizek, 2012). In addition, selecting raters who can provide

accurate evaluations based on the assessment standards is essential in validating an assessment (Cizek & Bunch, 2007).

Although expert judgement is applied in most methods of standard setting for language assessment, some issues have been observed in its application. Two matters are of concern: the quality of the expert and the number of experts employed. The exact number of experts involved in the judgment process is of less importance. According to Flowerdew (2002), “as long as the participants are representative of the group and its culture… then there is no fixed criterion for the number of participants” (p. 283). As a matter of fact, some studies have involved very small numbers of experts. A study by Bramley and Wilson (2016) on the General Certificate of Secondary Education (GCSE), a high-stakes test, involved only three experts, as the researchers determined that this was sufficient to achieve an acceptable level of reliability when comparing expert evaluations. It is thus justifiable for this study to employ a small number of valuable experts.

Nevertheless, there have been attempts to offer guidelines for the number of expert judgments. Cizek and Bunch (2007) have suggested that 20-25 participants should be employed for high-stakes assessment, while lower-stakes assessment content can involve panels of eight to ten members (Cizek & Bunch, 2007). In this study, adopting Cizek and Bunch’s (2007) guideline seems appropriate as the CULI Test is relatively low stakes, and recruitment of the required number of experts is feasible for the researcher. Apart from the concern on the number of the experts, the Rasch model measurement can be used to show the severity of the experts. Allowing different facets to be analyzed into one scale, the Rasch model can offer their calculation result to prove the validity of the expert judgment on the test (Lee et al., 2022).

As well as ensuring that there is an adequate number of qualified experts, it is also important that the scale used in the process of standard setting be selected with caution. In standard setting through benchmarking, Humphry et al.

(2014) has suggested caution when consulting experts as they tend to overvalue the test taker's ability on the easy items and undervalue it on the hard items. Utilizing Angoff to estimate the item difficulty and the cut scores can produce an invalid result (Impara & Plake, 1997) as Athiworakun and Wudthayagorn (2018) have specified that consideration is needed when engaging more than three points of cut score. Even if experts are well trained, the scale of reference in benchmarking could be a source of exaggeration. Most of the studies conducting standard setting (Humphry et al., 2014) were conducted with the yes-no Angoff method, which requires the expert to give either a positive or negative answer, resulting in more consistent responses. To limit the overgeneralization in this study, the experts were asked to answer yes or no to each item on 4 levels of the CEFR scale: A2, B1, B2, and C1.

2.3 CEFR Illustrative Descriptor Scales

The CEFR provides a descriptive scheme of English proficiency level in terms of can-do statements. A table of general can-do statements for all levels is widely available to describe the ability of learners from A1 to C2 (Council of Europe, 2001). Furthermore, a companion volume of CEFR is also available to provide further guidance on how to implement this scale in learning, teaching, and assessment. The most recent version of this companion volume (Council of Europe, 2020) includes illustrative descriptor scales that define what needs to be tested and how this testing should be conducted, following the notion of communicative language competence. CEFR illustrative scales are categorized to include most of the situation that needs to be tested. In fact, previous studies by Azman et al. (2021), Hidri (2021), and Shak and Read (2021) have adopted this scale to establish their educational aims and validate their specific assessment procedures. The CEFR illustrative scales have been shown to be useful in these studies in terms of increasing validity and allowing for more direct matches between test results and the CEFR can-do statements.

The CEFR framework was developed based on models of communicative language competence which emerged in the applied linguistics field in the 1980s (Council of Europe, 2020). Originally, such models involved four facets: strategic competence, linguistic competence, pragmatic competence, and sociocultural competence. However, as strategic competence pertains to activities that are not always associated with language, this facet is omitted from CEFR-based descriptors (Council of Europe, 2020). The three factors and their subskills are shown in Table 3. An important step in CEFR alignment is categorizing what language is to be tested according to these subskills.

Table 3

Aspects of Communicative Language Competence (Council of Europe, 2020)

Linguistic competence	**Sociolinguistic competence**	**Pragmatic competence**
General linguistic range	Sociolinguistic appropriateness	Flexibility
Vocabulary range		Turn taking
Grammatical accuracy		Thematic development
Vocabulary control		Coherence and cohesion
Phonological control		Propositional precision
Orthographic control		Fluency

After this categorization is established, the next step is to determine how the can-do proficiency statements in the CEFR scale can be tested under these aspects. As the four skills of listening, speaking, reading, and writing cannot encompass the intricacy of communication in real life, another four modes of communication are designated in the CEFR scale, which are reception, production, interaction, and mediation. In this study, three of these modes were employed when aligning the CULI Test with the CEFR, with the exception being mediation. The act of mediation is described as “the user/learner acts as a social agent who

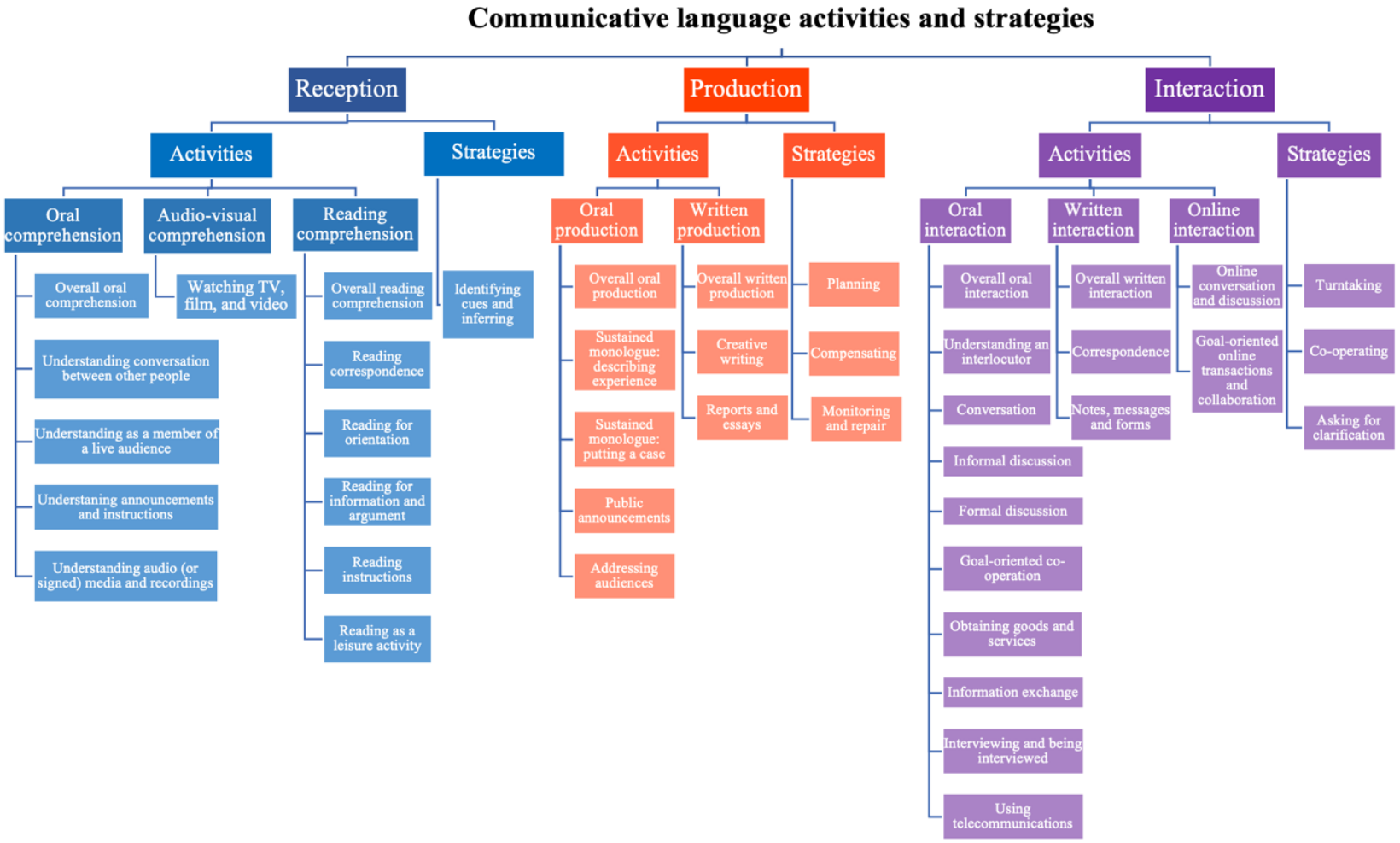
creates bridges and helps to construct or convey meaning, sometimes within the same language, sometimes across modalities and sometimes from one language to another" (Council of Europe, 2020, p. 90). As the CULI Test is multiple-choice and aims at assessing learners' proficiency, —not assessing learners' performance in various activities, —mediation is beyond the scope of this test.

Figure 1 presents communicative language activities and strategies under three domains: reception, production, and interaction. For each domain, relevant activities and strategies are presented. Activities are further specified according to their modes. Reception activities are divided into oral comprehension, audio-visual comprehension, and reading comprehension. Production activities encompass oral production and written production. Interaction activities include oral interaction, written interaction, and online interaction. After this categorization, 42 specific situations were designated with scales ranging from pre-A1 to C2. However, not all situations can include pre-A1 or C2 descriptors. Some activities appear over the competence of the pre-A1 level while the difficulty level in some activities can be defined up to C1.

E-ISSN: 2287-0024

Figure 1

Communicative Language Activities and Strategies (Adapted from CEFR Companion Volume, 2020)



¥

Although the scales to test communicative language competence include 35 activities and seven strategies, the Council of Europe (2001) does not propose that all scales should be applied in one test. As a matter of fact, a practical test should not exceed seven categories as doing so can overwhelm the test taker's cognitive load (Council of Europe, 2001). Accordingly, test developers should consider carefully which categories should be assessed, and design appropriate tasks that can assess multiple categories. For example, Azman et al. (2021) have demonstrated how a monologue speaking test aligned to CEFR employs the test taker's linear cognitive process in six categories.

While the *what* is delineated in the communicative language competences and the *how* is articulated through the communicative language activities and strategies, the two dimensions are mutually reinforcing in the comprehension and implementation of assessment tasks. It is therefore essential to identify and integrate the most relevant elements to construct test items that validly and reliably measure learners' competences in alignment with the general CEFR scale, as evidenced through comparison with the CEFR illustrative descriptors.

3. Methodology

The aim of this research was to align the CULI Test to the CEFR scales of communicative language competence by analyzing the item difficulties of the CULI Test in terms of their CEFR level using expert judgements alongside Rasch measurements. Starting with quantitative methods, eight experts, who hold at least two years' experience in language assessment, were recruited, with one expert acting as a coordinator to lead a review session on the CEFR scale so that all experts held an adequate understanding of the CEFR. This familiarization training was held in the Thai language to avoid any confusion that may arise from misunderstanding (Hulešová & Vodičková, 2023). A three-hour orientation to the CEFR and the CULI Test was conducted, which included training and practice activities. During this session, the CEFR framework was described, with samples given of language users at each level according to the criteria set out in the CEFR.

Experts were encouraged to ask when they were unsure about the level of competence of the test takers in accordance with the CEFR standards. Additionally, experts were given sample test items and asked to evaluate the CEFR level to which each item was most appropriately aligned. The coordinator provided feedback and discussed any disagreements that occurred among the experts to ensure understanding of the test items in accordance with the CEFR framework and to build familiarity with the test item evaluation process.

After the training session, the experts were provided with a set of 100 test items in hard copies and asked to assess at which CEFR level learners would need to be at to answer each item correctly. The range of the CEFR levels in this study was set between A2 and C1. Each expert recorded their answers for the CEFR level of each item on an individual Excel spreadsheet. Their answers were coded to numbers for analysis with the Rasch Model using Conquest software (Adams et al., 2020). A Wright map, which showed the relationship between judges' severity and test item difficulty, was generated in ConQuest as evidence to support the consistency measurements.

In addition to the results from experts' judgements in the form of a Wright map and comments, a qualitative method was conducted by analyzing experts' notes that they had taken on the hard copies of the exam in the previous stage. Each item was analyzed by the researcher with reference to two factors: the communicative language competence and the communicative language activities and strategies, as outlined in Table 3 and Figure 1. After the analysis, the researcher then conferred with experts to reach an agreement on the level of each test item.

3.1 Experts

The eight experts were Thai full-time lecturers at Chulalongkorn University Language Institute (CULI). All experts had taught undergraduate students at Chulalongkorn University for at least two years. In addition, the experts were

required to have been involved with language assessment at an undergraduate level for at least one year or to have developed a minimum of two courses of classroom language assessment.

4. Results

4.1 Examining Test Items on the Wright Map

Figure 2 shows a Wright map of all CULI Test items assigned by experts to a scale of B1 to C1. Each X sign represents the 0.7 cases or the expected test takers who could perform the item. Ranking from the bottom on each level are the easiest test items for which most test takers at each CEFR level would have more than a 50% chance to answer correctly. On the other hand, the items appearing on the top part of the Wright map are the questions that test takers in each level would have less than 50% chance of answering correctly. Since the number of X means the number of expected test takers, most of the test takers have more than a 50% chance to correctly answer 40% of the test items. This means that for the level that the experts rated, this test may be too hard for the targeted test takers to achieve. To understand the causes of these difficulties, the next section examines the items in more detail.

Figure 2

Wright Map from Many-facets Rating Scale Analysis

4.2 Comparison of CULI Test to CEFR Illustrative Descriptor Scales

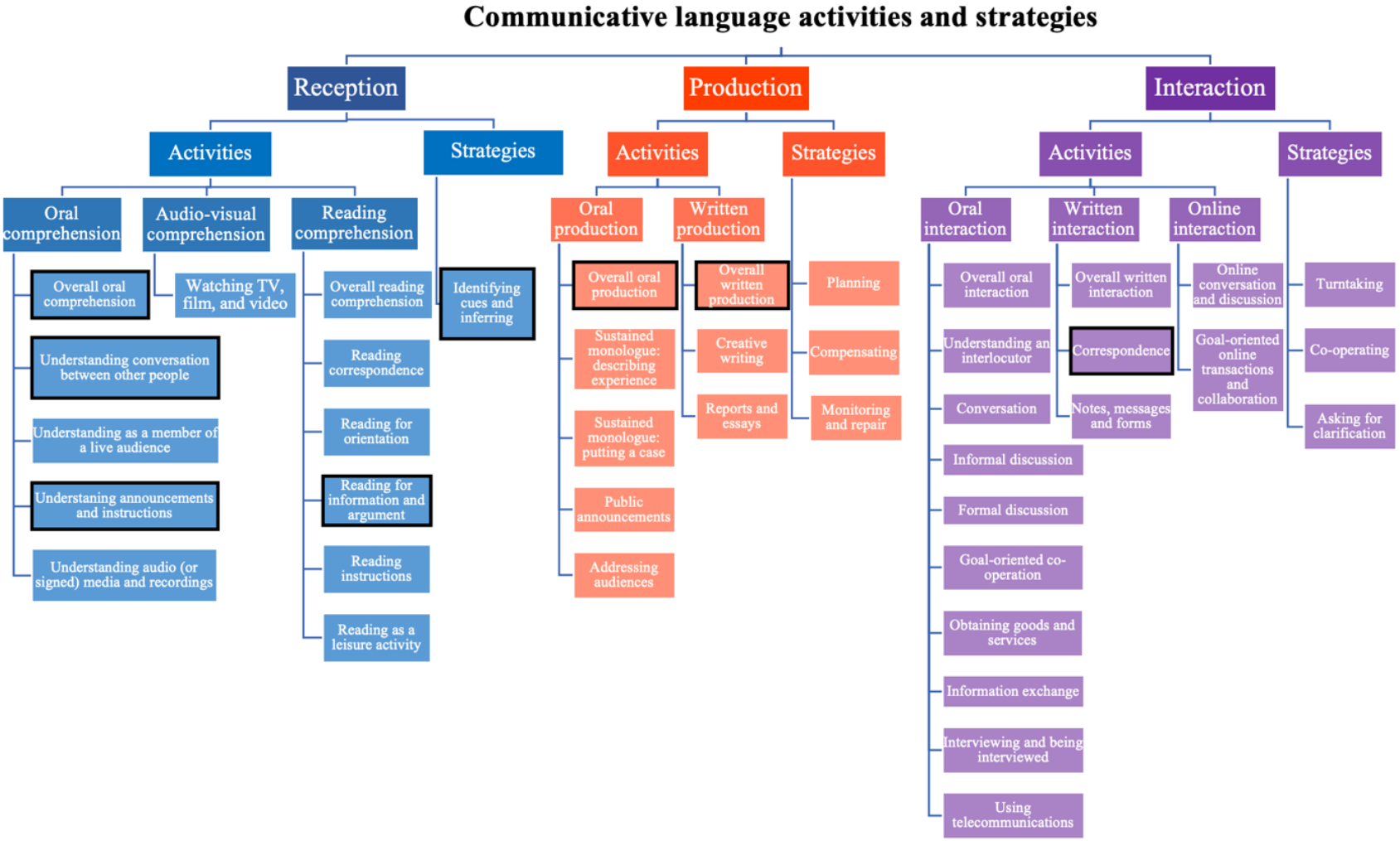
After analyzing the experts' agreement on the CEFR level of 100 test items, the results showed that the experts agreed on the level of most of the test items as B1 (49 items) or B2 level (42 items). With reference to the experts' notes, the researcher then matched each item to the description of communicative language competence and the communicative language activities and strategies. Figure 3 highlights (in black) the types of activities or strategies among 100 test items on

the CULI Test and their categorization within three domains of communicative language competence. The figure shows that eight of the 42 specific activities and strategies in the CEFR were included in the CULI Test. Reading for information and argument was the most common question type, with 24 items, followed by overall written comprehension with 20 items, while the lowest number was overall oral comprehension with five items. Regarding the three modes, most of the items were found to test reception skills, with only two production skills and one interaction skill being tested. Almost all the items were classified as activities, with only one strategy—identifying cues and inferring—being tested in the test items.

E-ISSN: 2287-0024

Figure 3

Activities and Strategies Used in CULI Test Items (Highlighted in Black)



After analyzing the question types, the next step was to determine what communicative language competence was being tested. As shown in Table 4, after conferring with the experts, three out of 13 aspects of communicative language competence were found to be tested: vocabulary range, grammatical accuracy, and vocabulary control. Vocabulary range was the most common, found in all parts of the test, followed by grammatical accuracy and vocabulary control, which were found in only two parts; however, one of these two parts—specifically, the sentence completion part—comprised 20% of the total items.

Most test items were found to test one aspect of communicative language competence under each activity. Nevertheless, it is worth noting that when some of the test items were found to assess more than one or even two aspects of communicative language competence, it could be difficult to determine what language aspect was required for the test takers to answer those items correctly. For example, four questions out of 20 in the sentence completion part were found to test both grammatical accuracy and vocabulary control. The process of reexamining these items along with the records of the test takers' answers should be done to validate these items.

In addition, when comparing the number of test items, the sentence completion and the text completion part tested the greatest number of aspects, which might lead to these two sections being the most difficult as the test takers would need to apply various types of communicative language competence to answer the questions.

E-ISSN: 2287-0024

Table 4

CULI Test Comparison to the CEFR Illustrative Descriptor Scales: Communicative Language Activities & Strategies and Competence from Illustrative Scales and CEFR Level

Part and no. of items	Communicative Language Activities and Strategies	Communicative Language Competence	CEFR range
Photographs (5)	- Overall oral comprehension	- Vocabulary range	A2-B2
Question-Response (15)	- Overall oral production	- Vocabulary range	A2-B2
Short Conversations (15)	- Understanding conversations - Identifying cues & inferring	- Vocabulary range	B1-C1
Short Talks (15)	- Understanding announcements & instructions	- Vocabulary range	B1-B2
Sentence Completion (20)	- Overall written production	- Grammatical accuracy - Vocabulary range - Vocabulary control	A2-C1
Text Completion (6)	- Correspondence	- Grammatical accuracy - Vocabulary range - Vocabulary control	B1-C1
Single Passages (14)	- Reading for information & argument - Identifying cues & inferring	- Vocabulary range	A2-C1
Double Passages (10)	- Reading for information & argument - Identifying cues & inferring	- Vocabulary range	A2-C1

The final column in Table 4 states the CEFR range of each section of the CULI Test. This CULI Test form contained items ranging from A2 to C1. Nevertheless, the numbers of test items assessing each level were not equal. The majority of the test items were found to assess B1 (49 items) or B2 (42 items) level, while few items were classified as either A2 (one item) or C1 (eight items) level. More details of how the questions correspond to the descriptors from the CEFR Illustrative Scales on communicative language activities, strategies, and competence are described below.

Table 5

Adapted Descriptions of A2 Levels from the CEFR Illustrative Scales Assessed in CULI Test Items (Adapted from CEFR Companion Volume, 2020)

Communicative Language Activities and Competence	**A2**	**Section**
Overall oral comprehension	Can understand phrases and expressions related to areas of most immediate priority (e.g., very basic personal and family information, shopping, local geography, employment), provided people articulate clearly in a generally familiar variety.	Photographs
Reading for information and argument	Can understand the main points of short texts dealing with everyday topics (e.g., lifestyle, hobbies, sports, weather). Can identify specific information in simpler materials they encounter such as letters, brochures, and short news articles describing events.	Single Passage Double Passages
Overall oral production	Can give a simple description or presentation of people, living, or working conditions, daily	Question-Response

Communicative Language Activities and Competence	A2	Section
	routines, likes/dislikes, etc. as a short series of simple phrases and sentences linked into a list.	
Overall written production	Can produce a series of simple phrases and sentences linked with simple connectors like "and," "but," and "because."	Sentence Completion
Vocabulary range	Has sufficient vocabulary to conduct routine everyday transactions involving familiar situations and topics. Has sufficient vocabulary for the expression of basic communicative needs and coping with simple survival needs.	Photographs Question-Response Sentence Completion Single Passage Double Passages
Grammatical accuracy	Uses some simple structures correctly, but still systematically makes basic mistakes.	Sentence Completion

4.3 A2 Categorization

In this CULI Test form, only one item was assessed as A2, which was in the sentence completion part testing overall written production. Based on the descriptors, A2-level items were categorized into four activities and two communicative language competences located in five sections of the CULI Test: photographs, question-responses, sentence completion, single passages, and double passages. As summarized in Table 5, these sections involved questions and answers that were simple, short, and related to daily life. The first communicative language activity found was overall oral comprehension in the photographs

section, which involved listening to sentences and matching a sentence to the photo. Second was the reading for information and argument in the single passage and double passages, which included a question about the main point or main idea of a short text. Third was the overall oral production in the question-responses section, where both the questions and possible responses were related to everyday information and composed of one simple sentence. Next, the overall written production being assessed in the sentence completion section included simple connectors as choices, corresponding to the description of simple structures of grammatical accuracy at A2 level. The vocabulary in all questions at this level was determined to be related to basic and familiar situations.

4.4 B1 Categorization

The largest number of test items fell into the B1 level, with 48 items. As shown in Table 6, the additional activities being assessed at B1 level were understanding conversation between other people, understanding announcements and instructions, reading for information and argument, and correspondence. The only strategy tested in the CULI Test, identifying cues and inferring, started in B1 level with basic inferences or predictions. Three communicative language competences were in this level, as vocabulary control was present in the sentence completion, text completion, and single passage sections. Grammatical accuracy at this level was described as reasonably accurate, while vocabulary range was classified as good.

Table 6

Adapted Descriptions of B1 Levels from the CEFR Illustrative Scales Assessed in CULI Test Items (Adapted from CEFR Companion Volume, 2020)

Communicative Language Activities, Strategies, and Competence	B1	Section
Overall oral comprehension	Can understand straightforward information about common everyday or job-related topics, identifying both general messages and specific details, provided people articulate clearly in a generally familiar variety. Can understand the main points made in clear standard language or a familiar variety on familiar matters regularly encountered at work, school, leisure, etc., including short narratives.	Photographs
Understanding conversations	Can follow much of everyday conversation and discussion, provided it is clearly articulated in standard language or in a familiar variety.	Short Conversations
Understanding announcements & instructions	Can understand simple technical information, such as operating instructions for everyday equipment. Can follow detailed directions. Can understand public announcements at airports, stations, and on planes, buses, and trains, provided these are clearly articulated with minimum	Short Talks

Communicative Language Activities, Strategies, and Competence	B1	Section
	interference from [auditory/visual] background noise.	
Reading for information and argument	Can understand the main points of short texts dealing with everyday topics (e.g., lifestyle, hobbies, sports, weather). Can identify specific information in simpler material they encounter such as letters, brochures, and short news articles describing events.	Single Passage Double Passages
Overall oral production	Can reasonably fluently sustain a straightforward description of one of a variety of subjects within their field of interest, presenting it as a linear sequence of points.	Question-Response
Overall written production	Can produce straightforward connected texts on a range of familiar subjects within their field of interest, by linking a series of shorter discrete elements into a linear sequence.	Sentence Completion
Correspondence	Can compose basic formal e-mails/letters (e.g., to make a complaint and request action). Can compose personal letters describing experiences, feelings, and events in some detail.	Text Completion

Communicative Language Activities, Strategies, and Competence	**B1**	**Section**
	Can compose basic e-mails/letters of a factual nature (e.g., to request information or to ask for and give confirmation). Can compose a basic letter of application with limited supporting details.	
Identifying cues & inferring	Can make basic inferences or predictions about text content from headings, titles, or headlines. Can watch or listen to a short narrative and predict what will happen next. Can follow a line of argumentation or the sequence of events in a story, by focusing on common logical connectors (e.g., however, because) and temporal connectors (e.g., after that, beforehand). Can deduce the probable meaning of unknown words/signs in a text by identifying their constituent parts (e.g., identifying roots, lexical elements, suffixes, and prefixes).	Short Conversations Single Passage Double Passages
Vocabulary range	Has a good range of vocabulary related to familiar topics and everyday situations.	Photographs Question-Response

Communicative Language Activities, Strategies, and Competence	B1	Section
	Has sufficient vocabulary to express themselves with some circumlocutions on most topics pertinent to their everyday life such as family, hobbies, and interests, work, travel, and current events.	Short Conversation Short Talks Sentence Completion Text Completion Single Passage Double Passages
Grammatical accuracy	Uses reasonably accurately a repertoire of frequently used “routines” and patterns associated with more predictable situations.	Sentence Completion
Vocabulary control	Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations. Uses a wide range of simple vocabulary appropriately when discussing familiar topics.	Sentence Completion Text Completion Single Passage

4.5 B2 Categorization

As shown in Table 7, 42 items at B2 level were detected in all sections of the CULI Test, assessing four communicative language activities, one strategy, and three competences: descriptors of the communicative language activities,

strategies, and competences at this level mentioned unfamiliar topics, details, different structures, and a variety of strategies. The vocabulary range started to involve general topics while vocabulary control and grammatical accuracy were classified as relatively high.

Table 7

Adapted Descriptions of B2 Levels from the CEFR Illustrative Scales Assessed in CULI Test Items (Adapted from CEFR Companion Volume, 2020)

Communicative Language Activities, Strategies and Competence	**B2**	**Section**
Overall oral comprehension	Can understand standard language or a familiar variety, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic, or vocational life. Only extreme [auditory/visual] background noise, inadequate discourse structure and/or idiomatic usage influence the ability to understand.	Photographs
Understanding conversations	Can identify the main reasons for and against an argument or idea in a discussion conducted in clear standard language or a familiar variety. Can follow chronological sequence in extended informal discourse, e.g., in a story or anecdote.	Short Conversations

Communicative Language Activities, Strategies and Competence	**B2**	**Section**
Understanding announcements & instructions	Can understand announcements and messages on concrete and abstract topics delivered in standard language or a familiar variety at normal speed. Can understand detailed instructions well enough to be able to follow them successfully.	Short Talks
Reading for information and argument	Can understand articles and reports concerning contemporary problems in which particular stances or viewpoints are adopted. Can recognize when a text provides factual information and when it seeks to convince readers of something. Can recognize different structures in discursive text: contrasting arguments, problem–solution presentation, and cause–effect relationships.	Single Passage Double Passages
Identifying cues & inferring	Can use a variety of strategies to achieve comprehension, including watching out for main points and checking comprehension by using contextual clues.	Short Conversations Single Passage Double Passages

Communicative Language Activities, Strategies and Competence	**B2**	**Section**
Vocabulary range	Has a good range of vocabulary for matters connected to their field and most general topics.	Photographs Question-Response Short Conversation Short Talks Sentence Completion Text Completion Single Passage Double Passages
Grammatical accuracy	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding.	Sentence Completion
Vocabulary control	Lexical accuracy is generally high, though some confusion and incorrect word/sign choice does occur without hindering communication.	Sentence Completion Text Completion Single Passage

4.6 C1 Categorization

This CULI Test contained eight items at the C1 level. As shown in Table 8, most test items fell under the C1 level on four scales: understanding conversations, grammatical accuracy, vocabulary control, and identifying cues and inferring. Only one communicative language activity in the short conversation section could be scaled at C1 level, which specified that learners should be able to identify attitude. This means that a question would be categorized as C1 level if it asked about a speaker's attitude, mood, or intention. An example of such a question is "what does the man probably feel?" Vocabulary range at this level expanded to almost all situations with appropriate use of less common vocabulary for vocabulary control. Finally, grammatical accuracy was described as high, with rare errors.

Table 8

Adapted Descriptions of C1 Levels from CEFR Illustrative Scales Assessed in CULI Test Items (Adapted from CEFR Companion Volume, 2020)

Communicative Language Activities, Strategies, and Competence	C1	Section
Understanding conversations	Can identify the attitude of each participant in an animated discussion characterized by overlapping turns, digressions, and colloquialisms that is delivered at a natural speed in varieties that are familiar.	Short Conversations
Identifying cues and inferring	Is skilled at using contextual, grammatical, and lexical cues to infer attitude, mood, and intentions and anticipate what will come next.	Short Conversations Single Passage Double Passages

Communicative Language Activities, Strategies, and Competence	C1	Section
Vocabulary range	Can select from several vocabulary options in almost all situations by exploiting synonyms of even words/ signs less commonly encountered. Can understand and use appropriately the range of technical vocabulary and idiomatic expressions common to their area of specialization.	Short Conversation Sentence Completion Text Completion Single Passage Double Passages
Grammatical accuracy	Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.	Sentence Completion Text Completion
Vocabulary control	Use less common vocabulary idiomatically and appropriately.	Sentence Completion Text Completion Single Passage

When examining the scale of communicative language activities and strategies and communicative language competence, it followed logically that most of the items were found to be testing B1 and B2 levels. The categorization of the test according to the CEFR illustrative descriptor scales for almost all sections started at B1, except for question-response, which started at A2, and sentence completion, which started at A1. It is worth noting that some of the activities used

as multiple-choice items, namely overall oral production and correspondence, were limited to the B1 level.

5. Discussion

The aim of this study was to align the CULI Test format to CEFR scales of communicative language competence. Both the quantitative and qualitative results suggested that the majority of the items on the CULI Test could be defined as B1 and B2 levels, with some items categorized as A2 and C1 levels. In fact, most of the items in the first four sections of listening tasks were at the B2 level. As most test takers are likely to be between B1 and B2 proficiency, this indicated that the test is mostly appropriate for differentiating test takers between these proficiency levels.

The CULI Test is in accordance with the Council of Europe's (2001) recommendation that a maximum of seven aspects of communicative competence be tested in each section of a test, as no section of the CULI test exceeds three aspects. Unlike Azman et al. (2021), where the aspects to be tested were arranged according to the cognitive processes involved, the CULI Test assesses certain aspects multiple times in several sections of the test. A large number of questions at two specific CEFR levels—B1 and B2—means that this test is mainly capable of differentiating test takers of intermediate and upper intermediate proficiency. Similar results have been found for other university-based tests in Thailand (Cheewasukthaworn, 2022), which can be explained through the fact that the target test takers for such tests are mostly undergraduates who are expected to require levels of English in the B1-B2 range to complete their studies.

If the aim of the CULI Test is to measure proficiency more broadly, more items assessing other levels should be added. Specifically, if the CULI Test aims to differentiate between four CEFR proficiency levels, more items at A2 and C1 level should be added. For example, Hidri (2021) describes a test that claims to assess four levels, B1-C2, and applies various tasks pertinent to each level.

The inclusion in the CULI Test of a large number of questions testing the same level may be considered excessive and redundant. If the goal of the test is to differentiate whether the test taker is of B1 or B2 level, the number of test items can be reduced, resulting in a shorter test. This approach may help alleviate the issue of stress among test takers, as addressed by Jabeen et. al. (2025). In the circumstance that the shorter version of the test remains reliable and valid in differentiating proficiency levels, as suggested by Min and Bishop (2024), fewer questions and less time in the test room would benefit all stakeholders, especially the test takers.

Therefore, administrators and test developers may wish to consider redesigning the CULI Test to be a proficiency test specifically of B1 or B2 level proficiency. Doing so would mean that the CULI Test would resemble other language tests for non-native speakers of other languages, such as the Japanese-Language Proficiency Test for Japanese and the HSK (Hanyu Shuiping Kaoshi) test for Chinese. These two language tests are available in multiple versions according to their difficulty levels, which are determined by the amount of vocabulary knowledge required. Test takers need to determine their approximate level and choose the appropriate corresponding test level when applying to take the test. Redesigning the CULI Test in such a way may make it more suitable for higher education contexts, as Thailand's Ministry of Education introduced minimum graduate proficiency level requirements of B1 for diploma students or B2 for bachelor's degree students (Sae-Ong, 2025). The CULI Test developers should decide on one of these two options depending on the intended purposes and scope of the test.

In spite of some criticisms, CEFR has proved to be a fruitful resource in providing descriptors for assessing proficiency level (Hidri, 2021; Shak & Read, 2021). When developing a test, a list of CEFR illustrative scales should accompany the selection of questions so as to map the questions to their proficiency level.

This should include the can-do statements, which can be later provided to the test takers.

Vocabulary level remains a core element of assessment, as evidenced across all sections of the CULI test. Knowledge of vocabulary is undeniably essential to perform well on an English test (Chen et al., 2023). Another essential concern is the content and genre of the text. As Siripol et al. (2025) have noted, differences in background knowledge can affect comprehension, and familiarity with the content or genre may influence the score result. Therefore, determining the vocabulary level and text theme should be a fundamental step in developing a test.

6. Limitation and Future Research

This study did not include interviews with test takers to elicit their standpoints. After adaptations have been made to the CULI Test, further research should be conducted to explore test takers' perspectives and the washback effect of the updated test.

7. Conclusion

This study has determined that the CULI Test mainly consists of items assessing the B1 and B2 CEFR proficiency level. Test developers aiming to align tests with the CEFR should select items based on the aspects of communicative competence that the test aims to assess and should avoid overloading test-takers' cognitive system by testing multiple aspects of communicative competence all at once. The results indicate that expanding the test to include more test items on the A2 and C1 levels could assess the test takers' proficiency across a wider spectrum. Drawing from this research, test writers should refer to the relevant CEFR scales when determining the difficulty of test items. Although this study was conducted on multiple-choice items, similar guidelines apply to other test types, as shown in previous studies (Azman et al., 2021; Shak & Read, 2021). Finally, if the aim of a test is to determine whether test-takers meet a specific proficiency

level, a greater number of questions may not always be desirable. As demonstrated in this study, multiple questions testing the same skills on the same CEFR level may exhaust the test takers while yielding no further insight into their proficiency level. This study suggests that reducing the number of questions on the test may produce a better result for both the test takers and the test administrators in terms of time and resources. Ultimately, this study suggests that the CEFR descriptors include all the communicative competences, activities, and strategies needed to align a language test with the CEFR scales.

8. About the Author

Sasithorn Limgomolvilas is a full-time lecturer at Chulalongkorn University Language Institute (CULI). After receiving her Master's degree in TESOL from San Francisco State University, she returned to teach at CULI before graduating with a Ph.D. in English as an International Language, Chulalongkorn University. Her main research interests are in ESP and language assessment.

9. Acknowledgement

I am deeply grateful to Associate Professor Dr. Jirada Wudthayagorn for the initial encouragement and the continuous support. I also would like to express my gratitude to the reviewers for their insightful suggestions.

10. Declaration of AI Use

The author declares that Gemini and Consensus were used in searching for synonym suggestions and related articles, respectively.

11. References

Adams, R. J., Wu, M. L., Cloney, D., Berezner, A., & Wilson, M. R. (2020). *ACER ConQuest: Generalized item response modelling software* (Version 5.29) [Computer software]. Australian Council for Educational Research. https://www.acer.org/au/conquest

Anantapol, W., Keeratikorntanayod, W., & Chobphon, P. (2018). Developing English proficiency standards for English language teachers in Thailand. *Manutsayasat Wichakan, 25*(2), 1–35.

Athiworakun, 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. https://doi.org/10.55766/CTUU4836

Azman, H., Othman, Z., Shamsuddin, C. M., Wahi, W., Aziz, M. S. A., Mohamad, W. N. W., Othman, S., & Amin, M. H. M. (2021). Relating a sustained monologue speaking production rest to CEFR: Towards alignment. *Pertanika Journal of Social Sciences & Humanities*, *29*, 385–400. https://doi.org/10.47836/pjssh.29.S3.20

Baharum, N., Ismail, L., Nordin, N., & Razali, A. (2021). Aligning a university English language proficiency measurement tool with the CEFR: A case in Malaysia. *Pertanika Journal of Social Sciences and Humanities.* https://doi.org/10.47836/pjssh.29.s3.09

Bramley, T., & Wilson, F. (2016). Maintaining test standards by expert judgement of item difficulty. *Research Matters: A Cambridge Assessment Publication*, *21*, 48–54.

Chen, L.-C., Chang, K.-H., Yang, S.-C., & Chen, S.-C. (2023). A corpus-based word classification method for detecting difficulty level of English proficiency tests. *Applied Sciences, 13*(3), Article 1699. https://doi.org/10.3390/app13031699

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). Routledge. https://doi.org/10.4324/9780203848203

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

Council of Europe (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. https://rm.coe.int/1680459f97

Council of Europe (2020). *Common European Framework of Reference for languages: Learning, teaching, assessment – Companion volume*. https://www.coe.int/lang-cefr

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Education Testing Service (2020). *TOEFL® research insight series, Volume 6: TOEFL program history*. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v6.pdf

Flowerdew, J. (2002). Ethnographically inspired approaches to the study of academic discourse. In J. Flowerdew (Ed.), *Academic discourse* (pp. 235–252). Pearson Education.

Green, A. (2017). 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

Hidri, S. (2021). Linking the international English language competency assessment suite of examinations to the common European framework of reference. *Language Testing in Asia, 11*(1), 1–24. https://doi.org/10.1186/s40468-021-00123-8

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education*, *27*(1), 1–18.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353–366. http://www.jstor.org/stable/1435114

Jabeen, F., Ruqia, & Bahar, L. (2025). Examine the effects of high-stake resting on student's mental health at university level: A survey research. *Kashf Journal of Multidisciplinary Research, 2*(5), 1–19. https://doi.org/10.71146/kjmr430

Lee, T., Milanovic, M., & Pike, N. (2022). Equating Rasch values and expert judgement through externally-referenced anchoring. *International Journal of TESOL Studies*, *4*(1), 187–202. https://doi.org/10.46451/ijts.2022.01.12

Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (Vol. 33). Cambridge University Press.

Min, S., & Bishop, K. (2024). A shortened test is feasible: Evaluating a large-scale multistage adaptive English language assessment. *Language Testing, 41*, 627–648. https://doi.org/10.1177/02655322231225426

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels. *Research Memorandum No. RM-15-06*. Educational Testing Service.

Sae-Ong, U. (2025). University students' self-efficacy and assessment literacy: Insights from the English Exit Exam policy in Thailand. *Language Testing in Asia*, *15*, Article 55. https://doi.org/10.1186/s40468-025-00379-4

Shak, P., & Read, J. (2021). Aligning the language criteria of a group oral test to the CEFR: The case of a formal meeting assessment in an English for occupational purposes classroom. *Pertanika Journal of Social Sciences & Humanities*, *29*, 133–156. https://doi.org/10.47836/pjssh.29.s3.08

Siripol, P., Rhee, S., Thirakunkovit, S., & Liang-Itsara, A. (2025). Evaluating the consistency of automated CEFR analyzers: A study of English language text classification. *International Journal of Evaluation and Research in Education*, *14*(4), 3283–3294. https://doi.org/10.11591/ijere.v14i4.33528

Tangsakul, S., & Poonpon, K. (2024). Aligning academic reading tests to the Common European Framework of Reference for languages (CEFR). *rEFLections*, *31*(2), 614–638. https://doi.org/10.61508/refl.v31i2.275057

Wudthayagorn, J. (2018). Mapping the CU-TEP to the Common European Framework of Reference (CEFR). *LEARN Journal: Language Education and Acquisition Research Network, 11*(2*)*, 163–180. https://so04.tci-thaijo.org/index.php/LEARN/article/view/161641/116576