English Language Teaching; Vol. 14, No. 7; 2021
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

The Development of STEP, the CEFR-Based English Proficiency Test

Kietnawin Sridhanyarat[1], Supakarn Pathong[2], Todsapon Suranakkharin[3] & Amornrat Ammaralikit[1]

[1] Department of Western Languages, Faculty of Archaeology, Silpakorn University, Thailand

[2] Department of English, Faculty of Arts, Silpakorn University, Thailand

[3] Department of English, Faculty of Humanities, Naresuan University, Thailand

Correspondence: Kietnawin Sridhanyarat, Department of Western Languages, Faculty of Archaeology, Silpakorn University, Thailand 10200.

Received: June 3, 2021 Accepted: June 29, 2021 Online Published: June 30, 2021

doi: 10.5539/elt.v14n7p95 URL: https://doi.org/10.5539/elt.v14n7p95

Abstract

This study aimed at developing the Silpakorn Test of English Proficiency (STEP), in alignment with the Common European Framework of Reference for Languages (CEFR), and in accordance with the theoretical framework established by Alderson et al. (2006). Four major steps were involved in the test construction. First, English language lecturers who served as content specialists were asked to design can-do statements presented in the CEFR. Then the specialists designed the test specification based on the can-do statements. Four skill areas: listening, semi-speaking, reading, and semi-writing were targeted as the test construct. At this juncture, the content specialists were required to write test items in accordance with the test specification. Next, the test items constructed were determined for their validity and reliability. Finally, a standard setting was carried out. The results demonstrated that the framework offered by Alderson et al. (2006) served as an effective reference document for developing the STEP. In terms of validity and reliability, the STEP was of statistical significance, that is, it could be aligned with the CEFR levels and measure test takers' English proficiency at a specific CEFR level. The current findings provide useful insights for test developers or researchers who wish to design proficiency tests in alignment with the CEFR.

Keywords: CEFR, CEFR-based proficiency test, test specification, the STEP

1. Introduction

1.1 Background of the Study

It is generally acknowledged that English plays a crucial role in international communication among people around the world. Within the ASEAN Economic Community (AEC), English is regarded as the working language among individuals (Hiranburana et al., 2017). English serves as the key to success in both education and professions (Pisuwan, 2014). However, communicating in English places a large burden on a number of Thai students (Pitsuwan, 2014; Prappal, 2003). Such a burden potentially derives from various factors; for example, learning strategies (Oxford, 2003), methods of teaching (Tayjasanant & Suraratdecha, 2016), or syllabus design (e.g., Abhakorn, 2017; Rattanaphumma, 2012).

Within the context of Silpakorn University, Thailand, the situation noted earlier is not surprising. The level of English proficiency among students in the current university context mainly ranges between A1 and B1, which suggests that many students need further improvements in terms of their English language skills. The CEFR consists of three main levels—A, B, and C—each further subdivided into two levels. A1 is the lowest proficiency level, while C2 is the highest (Council of Europe, 2001).

Relevant literature highlights that it is generally accepted that a good language test is considered to be a significant factor to help ameliorate such a challenging situation. For example, Heaton (1988) points out that a language test produces an effective impact on learning and teaching, which ultimately brings about students' improved learning habits. Spaan (2009) further supports this notion stating that a test serves as a reference document employed to make decisions relevant to learning and instruction. Therefore, tests are considered useful tools in providing insightful implications for the learning and teaching of a particular language.

In efforts to reinforce undergraduate students' English language skills, there is a lack of a good proficiency test suited to the context of Silpakorn University, and therefore, there is a need to develop a systematic test that can

help guide the teaching and learning of English for students there. The primary aim of this research project is to develop the Silpakorn Test of English Proficiency (henceforth STEP) and to see whether it is both valid and reliable. Adopted as a standard framework of reference for English language education in both Thailand and international contexts (Hiranburana et al., 2017), the current project intends to construct the STEP, associated directly with the CEFR. Although the CEFR has been widely adopted as a reference document by numerous test developers, teachers, and policy makers (Deygers et al., 2018), its concise guidelines enabling test specifications to be derived from each level of the CEFR have not been well established. Thus, this research also opts for a more specific theoretical and practical framework proposed by Alderson et al. (2006). Based entirely on the CEFR, the framework allows test developers to create test items that are associated closely with the constructs elucidated in the CEFR and systematically calibrated to the CEFR-based levels.

The insights gained from this research project would give an understanding into problems regarding the teaching and learning of English for Thai students in general, and students of Silpakorn University in particular. Additionally, the STEP would be useful in describing learners' specific levels of English proficiency, in both Thailand and international contexts, and assist in connecting students' proficiency more closely with specific academic and professional qualifications, which serve as their academic and professional references.

1.2 Research Questions

The major goal of this research project is to construct a proficiency test in accordance with the CEFR. In this project, the STEP would be administered to assess students' ability in using English in various contexts. Specifically, the STEP would provide a standard reference document to offer useful information about learners' levels of proficiency in four skill areas: listening, semi-speaking, reading, and semi-writing. The current research project aims to answer the following questions:

1) Is the developed test (STEP) valid and reliable?

2) To what extent does the developed test (STEP) correspond to the CEFR?

2. Literature Review

The literature section discusses the notion of the CEFR, which served as a reference guideline in developing the STEP. The section also elucidates Alderson et al.'s (2006) framework that was crucial to the construction of the STEP. In this study, the STEP served as a useful tool for successful learning and teaching, relevant to English language learners' needs and society. To fulfill this function, the STEP was thus designed in accordance with the CEFR (Council of Europe, 2001).

Adopted locally and globally as a standard reference, the CEFR has been constructed as a common basis for the integration of language syllabuses, curriculum references, exams, books, etc. (Alderson et al., 2006; Hiranburana et al., 2017). The CEFR provides a comprehensive means for learners to use a certain language for communication. It also describes what language skills and knowledge learners need to improve so that they can communicate in a particular language effectively (Leńko-Szymańska, 2015). Within the framework, comprehensive descriptions are provided, which include the cultural context of a particular language. The CEFR describes scales of proficiency so learners' progress can be gauged at each stage of their language learning over time.

Presented in the CEFR, there are two main scales: global and illustrative scales. Both scales are designed to determine the various proficiency levels of language learners. The illustrative scales differ from the global scale in that they are designed specifically to describe learners' performance of listening, speaking, reading, and writing skills. In contrast, the global scale broadly describes language use essential for learners of a particular language, as presented in Table 1.

Table 1. Global scale

Scales		Descriptors
Proficient User	C2	-Can understand with ease virtually everything heard or read.
		-Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation.
		-Can express himself/herself spontaneously, very fluently and precisely, differentiate finer shades of meaning even in more complex situations.
	C1	-Can understand a wide range of demanding, longer texts, recognize implicit meaning.
		-Can express himself/herself fluently and spontaneously without much obvious searching for expressions.
		-Can use language flexibly and effectively for social, academic, and professional purposes.
		-Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors, and cohesive devices.
Independent User	B2	-Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization.
		-Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.
		-Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	-Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.
		-Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics, which are familiar, or of personal interest.
		-Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	-Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g., very basic personal and family information, shopping, local geography, employment).
		-Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters.
		-Can describe in simple terms, aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	-Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.
		-Can introduce himself/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has.
		-Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

In Table 1, there are three main proficiency levels: A, B, and C, whereby each level includes Level 1 and Level 2. In this respect, the former level is less advanced than the latter. Students whose proficiency ranges between A1 an A2 are regarded as Basic Users; those whose competence level is between B1 and B2 are perceived as Independent Users, and students whose proficiency is evaluated between C1 and C2 are viewed as Proficient Users (Council of Europe, 2001, p. 24).

In short, the CEFR provides a useful grid that test developers can use to describe their testing system. However, this conceptual framework for language learning and teaching should not be rigid. The CEFR-based framework should be flexible so that it can be adapted to a given context (Alderson et al., 2006; Council of Europe 2001;

Deygers et al., 2018; Leńko-Szymańska, 2015). The STEP project thus corresponded to this basic CEFR principle.

According to Alderson et al. (2006), test specifications which are derived from each level of the CEFR have not been well documented. Thus, the CEFR alone does not provide explicit guidance on the actual development of a test and test tasks (Alderson et al., 2006). One of the consequences is that it provides a broad range of communicative activities (Spaan, 2009; Weir, 2005), leading to ambiguity.

Alderson et al. (2006) also point out that there are four major problems concerning the various frames offered by the CEFR. The first type of problem concerns the inconsistent use of the *can-do* statements displayed in each scale of proficiency. Similar *can-do* descriptions sometimes occur in different scales, and something at B1 involving vocabulary in contexts, is not displayed at lower or higher scales. For example, the *can-do* statement *recognize* occurs only at A1, B1, and C1; however, it does not occur at A2, B2, and C2. Another instance is that *infer* is indicated at C1 only. As proposed by Alderson et al. (2006), *inferencing* should also occur as an operation at B1 and B2.

Another problem is that the terms used to describe student competencies are relatively synonymous; for example, the verbs *understand* and *recognize* used to describe comprehension at B1. The third problem is that the definition of certain features is far from clear, or in other words, the terms given are not explicitly defined. For instance, the word *simple* is used very often in the scales; however, there is no meaningful distinction between *simple* and *less simple.* Clearly, this indicates that the CEFR does not contain any guidelines as to what should be simple, with respect to structures, lexis, or any other linguistic features. Thus, lists of grammatical features and lexical items for a certain language to be tested need to be supplemented in the CEFR. Alderson et al. recommend the use of electronic corpora as a reliable source "if terms such as *simple* and *frequent*" are used in the scales (p. 12). This problem also applies to a variety of other expressions presented in the CEFR. Many expressions such as *the most common*, *everyday*, *concrete*, or *highly colloquial* "need to be clarified, defined, and exemplified" if tasks and test items are to be linked to specific CEFR scales (p. 12).

The last problem is that there is a limitation which prompts the development of systematic test specifications. For example, the task, as what candidates have to do with texts is not systematically addressed. Also, there is no discussion of how tasks may be differentiated by scale. To elaborate, some of the illustrative scales address general items such as *reading for orientation* or *reading for information and argument*; however, other scales address specific texts such as *listening to announcements and instructions* or *reading instructions* (Alderson et al., 2006, p. 13). This disadvantage is considered to be a major problem for test writers and item bank builders. To overcome such a problem, Alderson et al. (2006) propose a more systematic framework which helps test developers to relate test specifications more closely to the CEFR and specific proficiency levels in the CEFR. The framework is demonstrated in 3.2 and 3.3 of the Method Section.

In summary, the CEFR has been designed as a common basis for the integration of language syllabuses, curriculum references, tests, books, etc; however, the CEFR alone does not provide systematic guidance which enables test developers to devise test specifications based exclusively on the *can-do* statements and specific levels in the CEFR. A more solid theoretical framework as established by Alderson et al. (2006) was a better option for this study.

3. Method

A criticism of the CEFR is that the CEFR-based scales for describing listening and reading abilities are structured by level rather than by activity. Specifically, all the *can-do* verbs presented in the CEFR focus mainly on comprehension, such as *infer*, *locate*, or *understand*. This creates a serious problem regarding how tasks can be distinguished by level. Accordingly, a more systematic test method provided by Alderson et al. (2006) could allow test developers and test writers to link their test directly to the CEFR and specific levels identified in the CEFR. In order to understand the entire methodology process, this section first addresses the project goal. It further explains the procedures for developing the current test. Finally, how the test is aligned with specific CEFR levels is presented.

3.1 Project Goal

The main purpose of the present project was to develop a test based on the CEFR. In this project, the STEP was used to measure students' ability in using English in various contexts. Essentially, the STEP was developed as a standard reference document to provide useful information about learners' levels of proficiency in the four skill areas: listening, semi-speaking, reading, and semi-writing. The current research project aimed to answer the following questions:

1) Is the developed test (STEP) valid and reliable?

2) To what extent does the developed test (STEP) correspond to the CEFR?

The insights gained from this project would give an understanding into the problems regarding the teaching and learning of English for Thai learners in general, and students of Silpakorn University in particular.

3.2 Developing Test Specification

The constructs aimed at for this project included listening, semi-speaking, reading, and semi-writing. As discussed, the CEFR does not provide adequate guidance for the development of tests (Alderson, 2005; Alderson et al., 2006). Therefore, this study opted for a conceptual framework proposed by Alderson et al. (2006), which served as a generic guideline for constructing the STEP. In their research project, Alderson et al. gathered expert judgements, which determined whether the CEFR was applicable for test construction. These researchers also identified what was missing in the CEFR, developed a conceptual frame for test analyses and specifications, and determined guidelines for item writers and sample tasks for different languages at the 6-CEFR levels. Accordingly, it is reasonable to claim that the framework proposed by Alderson et al. offered an insightful instrument for describing the relationship between test specifications and the CEFR.

The CEFR has been established to be *comprehensive*, *transparent*, and *coherent*. By *comprehensive*, the CEFR should be able to identify various ranges of language knowledge, skills, and usage. This concept requires language users to describe their goals as well. In terms of *transparent*, information must be precise and explicit, available, and readily understandable to language users. In relation to *coherent*, "the description is free from internal contradiction" (Council of Europe, 2001, p. 7).

Concisely, *comprehensive*, *transparent*, and *coherent* are essential concepts for test developers when developing any test which corresponds to the CEFR. The current project also applied this basic CEFR philosophy to test construction. As discussed previously, the CEFR alone does not provide a wealth of comprehensive description for test developers to characterize test content. For example, there exists no meaningful distinction between the verbs *understand* and *recognize* used to describe reading comprehension as elucidated in Level B1. These verbs are considered synonymous and thus unclear (Alderson et al., 2006). To avoid this constraint, Alderson et al. propose the following procedures for constructing test items:

1) Describe the text and the tasks employing the characteristics of a classification system.

2) Predict the level of a certain task which is directed by the classification system and the CEFR-based levels. This brings about an evaluated CEFR scale.

3) Pretest the task which explicitly describes the features of the sample piloted.

4) Calibrate the task.

5) Set the levels deriving from the calibration.

6) Select a psychometric level for the given task.

7) Select a certain level for the task.

Evidently, the framework proposed above provides more systematic and practical guidance on the actual construction of a test in response to the CEFR. This project thus intended to apply such a framework to constructing the STEP. In this study, there were four major steps taken to construct the STEP. In the first step, can-do statements for the STEP were developed. The second step dealt with the development of texts and tasks and the construction of test items. The third step involved the analysis of content validity and reliability of the tests. The final step was associated with standard setting.

3.3 Procedures for Developing the STEP

The following are step-by-step procedures for developing the STEP.

3.3.1 Describing Can-Do Statements

In this project, can-do statements were developed based entirely on the CEFR. The can-do statements of this study were developed on the basis that the CEFR has been adapted as a conceptual guideline for teaching general education English courses among students at Silpakorn University.

In this stage, 15 Thai lecturers serving as English language experts took part. These English language experts were educated to at least master's degree level, had been teaching English in Thailand for over five years. There were three steps involved in the development of can-do statements. First, the experts, divided into four groups (listening, semi-speaking, reading, and semi-writing), were required to describe can-do statements in accordance

with the CEFR. At a certain CEFR level, they had to describe what sub-skills students were expected to master. The can-do statements were based on the experts' judgements. Next, each group of experts was required to present their own can-do statements to the other groups so that useful feedback and suggestions could be obtained. This procedure assisted in determining whether the can-do statements developed were associated with the CEFR. Finally, a principled frame developed in accordance with the *can-do* statements of the CEFR was obtained.

3.3.2 Describing Test Specification and Item Construction

In this stage, test developers involved were encouraged to describe their test specification (the text and tasks) based on the can-do statements. Then they constructed test items in accordance with the specification designed. The description of texts and tasks was sometimes known as the Grid. In general, the Grid was organized into three parts: the texts to which the test was related, the test items, and the entire task which covered both the texts and the test items (Alderson et al., 2006). The test specification or the Grid in this project was developed based entirely on the CEFR.

The procedures for identifying the test specification are explained as follows. First, the judges were required to describe tasks and test items (test specification) using the can-do statements, which served as the operations. In this regard, they were also asked to assign levels to tasks, guided by the test specification and specific CEFR levels. The language experts who were responsible for developing each skill area were then required to present their own test specification (blueprint) to the other experts so that useful feedback and suggestions could be obtained. This procedure assisted in determining whether the blueprint developed was associated with the CEFR. Finally, a principled STEP blueprint developed in accordance with the can-do statements was obtained.

In this study, test items were constructed in response to the Grid or STEP blueprint, the theoretical reference document established based on the CEFR. There were two major steps taken in constructing test items. First, the experts had to write test objectives in accordance with the blueprint. The experts were then required to write test items based on the objectives they had developed. This helped in making the test items closely related to the target blueprint.

Under investigation, the STEP covers four skill areas: semi-speaking, semi-writing, listening, and reading. The semi-speaking test included only items that assessed oral skills indirectly (Qian, 2009). They aimed to assess students' oral skills, such as responding to questions, establishing social contacts, responding in conversations using basic expressions, handling very short social exchanges, and ending conversations appropriately. The semi-writing test involved items that measured written skills indirectly. Specifically, it aimed to measure test takers' abilities in using English for written communication; such as using grammar and vocabulary, as well as developing arguments. In this study, listening and reading tests were developed to measure students' ability to recognize multiple features of language knowledge in listening and reading (Laufer et al., 2004).

Each skill area had 30 items, and only A1 and C1 levels were targeted. In a certain skill area, there were six items constructed to measure students' ability in each CEFR level. In total, there were 120 items covering the four skill areas in the STEP.

3.3.3 Content Validity and Reliability of the STEP

There were two steps taken to determine the content validity of the test items. First, constructed test items were determined for content validity by five editors. The criterion used to determine the content validity of the test items was based on the IOC, as developed by Rovinelli and Hambleton (1977). The IOC is a step taken to evaluate content validity of a particular test at the item development stage. The editors' responsibilities were as follows. First, the editors were required to determine whether the test items corresponded to the objectives specified in the blueprint. Then they had to examine whether the test items were grammatically correct and natural to native speakers of English. It should be noted here that all the steps were taken based on the IOC method. The IOC was indexed at 0.75, of which helped ensure that a certain test item was valid in terms of its content (Turner & Carlson, 2003). The results of the content validity are presented in the findings section.

The reliability of the STEP was then determined. In particular, the test was piloted with 33 students whose proficiency ranged between A1 and C1. In general, a pilot can be undertaken on a small scale. Piloting with appropriate samples of test takers is essential to know whether a certain item is truly at the level of difficulty anticipated. It also allows us to elicit as many comments as possible concerning the quality of the test instrument (Alderson et al., 2006). Therefore, knowing the level of the test takers allowed a judgement of adequacy for such items. Spaan (2009) defines pilot as a method to try out the test instrument to remove ambiguities, and to

determine the clarity and comprehensibility of test items. This process involved the interpretation of the difficulty level of the tasks and the evaluation of time overload.

To conclude, four major steps were taken to construct the STEP. The first step involved the description of can-do statements required for the STEP. The second step dealt with the construction of the STEP blueprint. During this step, test items were also constructed in response to the objectives identified in the blueprint. The next step ensured the content validity and reliability of the test. Finally, standard setting was carried out.

3.3.4 Assigning the STEP Levels to Specific Levels of the CEFR

The outcome of taking a certain test concerns a numerical score. With respect to tests employed for listening, semi-speaking, reading, and semi-writing, this score is usually derived from correct responses. Two decisions must be made in response to such a score. The first decision based on this score has to be made in order to evaluate the test taker's ability. This pass/fail decision helps ensure that the examinee performed satisfactorily on the exam.

If a test is to be related to the CEFR, a second decision must be made. To elaborate, the latter decision has to be made to ensure whether the test taker has reached a given CEFR level. Both decisions deal with the determination of *a cut score* which can be referred to as a *performance standard*. Based on a pass/fail decision, the cut score refers to the minimum score on the test which will result in the decision *pass*; however, scores that are less than the cut score result in the decision *fail*. For example, a cut score for Level B2 is the minimum score which will result in the decision that the candidate's ability is at B2 or higher, whereas scores lower than B2 are regarded as B1 or lower.

In relating the STEP to the CEFR, a cut score was set for A1, A2, B1, B2, C1. In this regard, every candidate's ability is interpreted either as A1, A2, B1, B2, or C1. In this study, the process of determining cut scores or setting performance standard was based on statistical tests and a group decision. In this study, STEP was equated with SPEEXX, a CEFR-based test so that specific CEFR levels for the STEP were obtained. The group required to make such a decision is generally known as a panel. For relating the STEP more closely to the CEFR, the panelists had to be familiar with the CEFR itself. Standard setting can have insightful implications for examinees and policy makers. Thus, the process of relating the STEP to the CEFR involved a set of procedures which needed to be rigorously conducted at different steps. The current project intended to relate the STEP to the rich description of the CEFR and the framework offered by our language experts. The approach adapted was considered a logical means to critically review and assess the content and level of the current test. To assign the STEP to the CEFR, the same group of experts were invited to have discussions together. The findings of the standard setting are revealed in the findings section.

4. Results

4.1 Findings Regarding Validity and Reliability of the STEP

This section presents results regarding validity and reliability. It also reports the difficulty and discrimination of the test. The results obtained from listening, semi-speaking, semi-writing, and reading tests are provided, respectively.

4.1.1 Listening Section

The listening section comprised 30 items. The validity and reliability of the listening test was determined based on the statistical tests shown in Table 2 below.

Table 2. Validity, Reliability, Difficulty, and Discrimination of the Listening Test

Dimensions	Methods	Indicators	Meanings
Validity	Content Validity	IOCt > 0.75	The listening test is of statistical significance. That is, it is able to measure students' ability in English from A1 to C1 levels (Turner & Carlson, 2003).
	Concurrent-based Validity	r_{pb} = 0.47	The listening test could measure the target content of the test (HR-Guide, 2018).
	Concurrent Validity	^{r}xy = 0.84	The listening test corresponds to the CEFR and SPEEXX, which is designed based on the CEFR at the significance level (Hopkins, 2002).
Reliability	Kuder-Richardson 20	KR_{20} = 0.87	The listening test is reliable at the significance level KR20 = 0.869 (U of Washington, 2019).
	Split-Half	^{r}tt = 0.93	The listening test is reliable at the significance level KR20 = 0.925.
Difficulty	Classical Model	p = 0.66	The difficulty of the test is of moderate level (Kelley, 1939).
Discrimination	Classical Model	d = 0.48	The discrimination of the test is of high value. (Testing Services, 2019).

As can be seen in Table 2, the findings show that the listening test displayed content validity at the significance level (content validity: IOCt > 0.75, concurrent-based validity: ^{r}pb = 0.47, and concurrent validity: ^{r}pb = 0.84), was statistically reliable (Kuder-Richardson 20: KR_{20} = 0.87 and Split-Half: r_{tt} = 0.93), offered a moderate level of difficulty (classical model: p = 0.66), and exhibited a high value of discrimination (classical model: d = 0.48).

4.1.2 Semi-Speaking Section

The semi-speaking test consisted of 30 items. The results of the semi-speaking test in terms of its validity, reliability, difficulty, and discrimination are presented in Table 3.

Table 3. Validity, Reliability, Difficulty, and Discrimination of the Semi-Speaking Test

Dimensions	Methods	Indicators	Meanings
Validity	Content Validity	IOCt > 0.75	The semi-speaking test is able to gauge students' ability in English from A1 to C1 levels (Turner & Carlson, 2003).
	Content-based Validity	^{r}pb = 0.44	The semi-speaking test can significantly measure the target content (HR-Guide, 2018).
	Concurrent Validity	r_{xy} = 0.84	The semi-speaking test corresponds to the CEFR scales and SPEEXX at a significant level (Hopkins, 2002).
Reliability	Kuder-Richardson 20	KR_{20} = 0.87	The semi-speaking test shows highly significantly reliability (U of Washington, 2019).
	Split-Half	r_{tt} = 0.81	The semi-speaking test is reliable at ^{r}tt = 0.81.
Difficulty	Classical Model	p = 0.61	The difficulty of the semi-speaking test is of moderate level (Kelley, 1939)
Discrimination	Classical Model	d = 0.48	The discrimination of the semi-speaking test is of high value (Testing Services, 2019)

As demonstrated in Table 3, the semi-speaking test was of high validity (content validity: IOCi > 0.75, content-based validity: ^{r}pb = 0.44, and concurrent validity: ^{r}xy = 0.84), displayed statistically significant reliability (Kuder-Richardson 20: KR_{20} = 0.87 and Split-Half: r_{tt} = 0.81), offered statistically significant difficulty (classical model: p = 0.61), and exhibited a significant level of discrimination (classical model: d = 0.48).

4.1.3 Reading Section

In Table 4, the findings regarding the validity, reliability, difficulty, and discrimination of the reading test are provided.

Table 4. Validity, Reliability, Difficulty, and Discrimination of the Reading Test

Dimensions	Methods	Indicators	Meanings
Validity	Content Validity	IOCt > 0.75	The reading test is aligned with the CEFR levels at the significance level IOCt > 0.75 (Turner & Carlson, 2003).
	Content-based Validity	$^{r}pb = 0.40$	The reading test can measure the target content of the STEP at the level of significance $^{r}pb = 0.400$ (HR-Guide, 2018).
	Concurrent Validity	$r_{xy} = 0.76$	The reading test corresponds to the CEFR levels and SPEEXX at the level of significance $^{r}xy = 0.76$ (Hopkins, 2002).
Reliability	Kuder-Richardson 20	$KR_{20} = 0.821$	The reliability of the reading test is of statistical significance *KR20 = 0.821* (U of Washington, 2019).
	Split-Half	$r_{tt} = 0.75$	The reading test is statistically reliable at $^{r}tt = 0.75$.
Difficulty	Classical Model	p = 0.66	The difficulty of the reading test is of moderate level (classical model: p = 0.66) (Kelley, 1939).
Discrimination	Classical Model	d = 0.44	The discrimination of the reading test is of high statistical significance d = 0.44 (Testing Services, 2019).

As shown in Table 4, the results show that the reading test embedded in the STEP were statistically valid (content validity: IOCt > 0.75, content-based validity: $^{r}pb = 0.40$, and concurrent validity: $^{r}xy = 0.76$), revealed statistically significant reliability (Kuder-Richardson 20: $KR_{20} = 0.82$ and Split-Half: $^{r}tt = 0.75$), presented a moderate level of difficulty (classical model: p = 0.66), and exhibited a highly significant level of discrimination (classical model: d = 0.44).

4.1.4 Semi-Writing Section

The semi-writing test consisted of 30 items. The results of the semi-writing test in terms of its validity, reliability, difficulty, and discrimination are presented in Table 5.

Table 5. Validity, Reliability, Difficulty, and Discrimination of the Semi-Writing Test

Dimensions	Methods	Indicators	Meanings
Validity	Content Validity	IOCt > 0.75	The semi-writing test can truly measure a certain CEFR level (Turner & Carlson, 2003).
	Content-based Validity	$^{r}pb = 0.40$	The semi-writing test can measure the target content of STEP (HR-Guide, 2018).
	Concurrent Validity	$r_{xy} = 0.85$	The semi-writing test corresponds to the CFER levels and SPEEXX (Hopkins, 2002).
Reliability	Kuder-Richardson 20	$KR_{20} = 0.84$	The reliability of the semi-writing test is of statistical significance (U of Washington, 2019).
	Split-Half	$r_{tt} = 0.85$	The semi-writing test is statistically reliable at $^{r}tt = 0.85$.
Difficulty	Classical Model	p = 0.54	The difficulty of the test is of moderate level (Kelley, 1939).
Discrimination	Classical Model	d = 0.46	The test is of significant level of discrimination (Testing Services, 2019)

As shown in Table 5, the semi-writing test was considered significantly valid (content validity: IOCt > 0.75, content-based validity: r_{pb} = 0.40, and concurrent validity: r_{xy} = 0.85), reliability was of statistical significance (Kuder-Richardson 20: KR_{20} = 0.84 and Split-Half: r_{tt} = 0.85), the difficulty was of moderate level (classical model: p = 0.54), and its discrimination was of an acceptable level (classical model: d = 0.46).

In summary, validity, reliability, difficulty, and discrimination were considered in constructing the STEP. From the analysis, the STEP was statistically valid and reliable. The difficulty and discrimination of the test were also of statistical significance.

4.2 Standard Setting

Table 6 below illustrates the alignment of the STEP-based scales with the CEFR.

Table 6. The Assignment of the STEP Proficiency Levels to the CEFR

CEFR levels	Standard levels	STEP
C1	Proficient user	101-120
B2	Independent user	76-100
B1	Independent user	51-75
A2	Basic user	26-50
A1	Basic user	1-25

In Table 6, students whose STEP score ranged from 1 to 25 were regarded as Basic Users (A1), A2 learners if their total score is between 26 and 50, B1 if between 51 and 75, B2 if between 76 and100, and C1 if between 101 and 120. In this study, standard setting was conducted according to two steps. First, the number of items in the STEP were divided into the five CEFR levels (A1, A2, B1, B2, and C1). Second, these score ranges were then decided by the experts in this study, offering five score ranges as follows: A1 = 1-24, A2 = 25-48, B1 = 49-72, B2 = 73-96, and C1 = 97-120. In this study, statistical tests were also conducted to decide these cut-off scores.

5. Discussion

The purpose of this research project was to develop the STEP, a proficiency test in alignment with the CEFR. More importantly, this study intended to investigate whether the developed STEP was valid and reliable, and to what extent it corresponded to the CEFR.

A theoretical framework offered by Alderson et al. (2006) was adapted. Four major steps were taken to construct the STEP: identifying can-do statements, designing test specification and test items, ensuring content validity and reliability, and carrying out standard setting.

The results illustrate that the STEP was statistically valid and reliable, and statistically significant in terms of difficulty and discrimination. Specifically, as the STEP was constructed based on the CEFR levels, the STEP served as a valid and reliable test that corresponded to the CEFR.

Researchers interested in developing a proficiency test in connection with the CEFR can thus apply the framework of Alderson et al. (2006) as a reference. This will help in describing learners' specific levels of English language proficiency. In addition, the framework will assist in connecting students' English language competency more closely with specific academic and professional qualifications, which serve as their academic and professional references.

There are several suggestions that should be considered for further studies. The STEP should be equated with other standardized tests designed and based on the CEFR to ensure its concurrent validity. Further studies should analyze administered test items by means of the Classical Model, as the main purpose of this analysis was to ensure the quality of the test items. Researchers interested in language test development can employ the IRT Model to see what test items are suitable for specific CEFR levels.

To help give insight into the specification of the STEP, item biases (i.e., effects of item biases on students' genders, their faculties, or their background) should be taken into account as well. Future research should investigate issues of test item specifications in detail to ensure that the STEP is a true parallel form. Researchers are advised to consider issues relevant to Bloom's Taxonomy (Anderson et al., 2001). Studies into test equating should also be taken into consideration to equate the STEP with other standardized tests.

6. Conclusion

This study aimed to develop the STEP, a proficiency test designed based on the CEFR. Specifically, this study attempted to determine whether the test developed was of statistical validity and reliability, and to what extent the current English proficiency test corresponded to the CEFR.

Under investigation, a theoretical framework established by Alderson et al. (2006) was adapted. There were four major steps involved in the construction of the STEP (i.e., identifying can-do statements, designing test specification and test items, ensuring content validity and reliability, and carrying out standard setting).

The findings demonstrate that the STEP is of statistical validity and reliability. Regarding difficulty and discrimination, the results indicate that the STEP is statistically significant. Therefore, it can be concluded that the STEP serves as a valid and reliable test that corresponds to the CEFR. In particular, the STEP is able to measure students' ability in using English at specific CEFR levels.

References

Abhakorn, J. (2017). Language syllabus from student teachers' perspective. *Electronic Journal of Foreign Language Teaching, 14*(2), 175-186.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum Books.

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

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives* (Complete edition). New York: Longman.

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

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

Heaton, J. B. (1988). *Writing English language tests (2nd ed.).* London, England: Longman.

Hiranburana, K., Subphadoongchone, K., & Tangkiengsirisin, S., Phoocharoensil, S., Gainey, J., Thogsngsri, J., Sumonsriworakun, P., Somphong, M., Sappapan, P., & Taylor, P. (2017). A framework of Reference for English language education in Thailand (FRELE-TH)—based on the CEFR, the Thai experience. *LEARN Journal, 10*(2), 90-119.

Hopkins, W. G. (2002). *A New View of Statistics.* Retrieved from https://www.sportsci.org/resource/stats/effectmag.html

HR-Guide. (2018). *Understanding test quality-concepts of reliability and validity*. Retrieved from https://www.hr-guide.com/Testing_and_Assessment/Reliability_and_Validity.htm

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17-24. https://doi.org/10.1037/h0057123

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*(2), 202-226. https://doi.org/10.1191/0265532204lt277oa

Leńko-Szymańska, A. (2015). The English Vocabulary Profile as a benchmark for assigning levels to learner corpus data. In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 115-140). Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/scl.70.05len

Oxford, R. L. (2003). *Language learning styles and strategies*. Mouton de Gruyter. https://doi.org/10.1515/iral.2003.012

Pitsuwan, S. (2014). Dr. Surin Pitsuwan, Opening Keynote Speaker; Teachers of English to Speakers of Other Languages (TESOL). *International Convention 2014*, Portland, Oregon, USA.

Prappal, K. (2003). English proficiency of Thai learners and directions of English teaching and learning in Thailand. *Journal of English Studies, 1*(1), 1-6.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6*(2), 113-125. https://doi.org/10.1080/15434300902800059

Rattanaphumma, R. (2012). Designing a language course in English in a Lingua Franca (ELF) setting: Perception and practice. *ABAC Journal, 32*(2), 1-10.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*(2), 49-60.

Spaan, M. (2009). Test and item specifications development. *Language Assessment Quarterly, 3*(1), 71-79. https://doi.org/10.1207/s15434311laq0301_5

Tayjasanant, C., & Suraratdecha, S. (2016). Thai EFL teachers and learners' beliefs and readiness for autonomous learning. *3L: The Southeast Asian Journal of English Language Studies, 22*(3), 153-169. https://doi.org/10.17576/3L-2016-2203-11

Testing Services. (2019). *Optimum Item Difficulty*. University of Wisconsin. Retrieved from https://www.uwosh.edu/testing/faculty-information/test-scoring/score-report-interpretation/item-analysis-1/item-difficulty

Turner, R. C., & Carlson, L. (2003). Index of Item-Objective Congruence for multidimensional items. *International Journal of Testing, 3*(2), 163-171. https://doi.org/10.1207/S15327574IJT0302_5

U of Washington. (2019). *Understand Item Analysis.* Retrieved from https://www.washington.edu/assessment/scanning-scoring/scoring/reports/itemanalysis/

Weir, C. J. (2005). Limitations on the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281-300. https://doi.org/10.1191/0265532205lt309oa