PASAA
Volume 63
January – June 2022

Developing a Standardized English Proficiency Test in Alignment with the CEFR

Kanchana Cheewasukthaworn
Rangsit University
Email Address: kanchana.c@rsu.ac.th

Abstract

In 2016, the Office of the Higher Education Commission issued a directive requiring all higher education institutions in Thailand to have their students take a standardized English proficiency test. According to the directive, the test's results had to align with the Common European Framework of Reference for Languages (CEFR). In response to this directive, this study was conducted to demonstrate how to develop such a test. Grounded in the conceptual frameworks of test usefulness, standardized tests, test development processes, and the CEFR, the study explored six stages in developing a 100-item-multiple-choice standardized English proficiency test: test planning, test design and writing, test reviewing, test piloting, test revising and revised test try-out. The test was piloted during the test piloting and revised test try-out stages with 108 and 88 students, respectively, and underwent item analysis to examine its reliability, difficulty index, and discrimination index. The item analysis conducted after the revised test try-out stage showed that the test's reliability, assessed by Kuder-Richardson Formula 20 (KR-20), was 0.94, suggesting that the test was highly reliable with good

internal consistency. How the test was developed in close alignment with test usefulness and the CEFR is discussed in the study.

Keywords: Test development, test alignment, item analysis, the CEFR

Introduction

The Council of Europe's (2001) Common European Framework of Reference for Languages (CEFR) has been introduced to language education in Thailand for less than a decade. In 2014, the CEFR was adopted as a guiding framework for English language teaching of basic education in Thailand (Office of the Basic Education Commission, 2014, p. 1). Two years later, the Office of the Higher Education Commission (OHEC) issued a directive concerning its policy to enhance English standards in higher education institutions (Office of the Higher Education Commission, 2016). According to Clause 5 of the directive, each higher education institution is required to have its students take a standardized English proficiency test with results aligned with the CEFR.

In compliance with the OHEC directive, Rangsit University decided to develop its own standardized English proficiency test for two reasons. First, the university planned that upon completion of the test development and alignment with the CEFR, the test would be used to serve other purposes relating to the university's learning and teaching management, e.g., an admission test for new graduates and a placement test after enrollment of new undergraduates in each term. The second reason relates to the cost of taking other standardized English proficiency tests currently available. Precisely, the university was concerned that the cost of taking these tests may be too high for its students to afford. There was also a concern about a delayed and incomplete collection of the test results from students. For these reasons, the university deemed it appropriate to develop its own standardized English proficiency test with scores aligned with the CEFR.

Rangsit English Language Institute was assigned to develop this test which was later called "Rangsit University Test of English Proficiency" (RSU-TEP). However, this task could not be accomplished swiftly nor could it be achieved without careful planning and implementation. One major concern that indeed most test developers face when developing any new test was to ensure that the test met all the requirements of a good test such as test validity, reliability, and practicality (Bachman & Palmer, 2010; Weir, 1993). It was also found that as stated by scholars in language test development (e.g., Downing, 2006; Green, 2014; Orozco & Shin, 2019), the test development process involved numerous steps and was a time-consuming task.

This study aims to demonstrate how to develop a standardized English proficiency test (i.e., RSU-TEP) with scores that align with the CEFR. The study primarily focused on the test development process and the test's quality in terms of validity and reliability while how to align the test's scores with the CEFR will be conducted in another study. The study was guided by the research question: How can the RSU-TEP be developed as a standardized English proficiency test with scores that align with the CEFR?

Literature Review

Literature relevant to this study includes that on test usefulness, standardized testing, the test development process, and the CEFR. After a thorough review of literature in these areas, test usefulness was found to be of utmost importance and related to the other three areas of focus.

Test Usefulness

Test usefulness is a concept proposed by Bachman and Palmer in 1996. According to these scholars, when any new test needs to be developed, it is important to ensure that the test is useful; that is, it fully meets its intended objectives. To ensure that the test is useful, Bachman and Palmer (1996, p. 17) proposed a test usefulness model comprising six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. These qualities can be classified into three groups as shown below.

Qualities Relating to Test Content Design

Construct validity, authenticity, and interactiveness can be considered the qualities relating to test content design, i.e., what to be included in a test. Bachman and Palmer (1996) defined construct validity as "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (p. 21). Based on this definition, for any test to be regarded as having construct validity, it must have (i) test validity (i.e., measure what it claims to measure) and (ii) test items that can yield scores which provide sufficient justification for an interpretation of test-takers' abilities measured by the test.

Authenticity is another quality related to test content design. Giraldo (2019) summarized it as "the degree of correspondence between an assessment (its items, texts, and tasks) and the way language is used in real-life scenarios and purposes" (p. 216). Simply put, if a test is to be considered authentic, its tasks must require test-takers to use the language for communication in real-life situations. The more a test's tasks require test-takers to use such language, the more authentic the test is.

The other quality relating to the test content design is interactiveness which is "the extent and type of involvement of the test-taker's individual characteristics in accomplishing a test task" whereby the test-taker's individual characteristics include "the test-taker's language ability, topical knowledge, and affective schemata" (Bachman & Palmer, 1996, p. 25). In other words, if a test is regarded as interactive, it must activate and require test-takers to apply their language abilities, including other relevant knowledge and skills to accomplish the tasks in the test.

Qualities Relating to Test Scores

Reliability and impact can be considered the qualities relating to test scores. Reliability is the consistency of scores on a test when the test is administered to the same group of test-takers but at different occasions or settings (Bachman & Palmer, 1996). Several scholars (e.g., Bachman & Palmer, 1996; Chapelle, 2012; Giraldo, 2019) agreed that reliability should be analyzed at the outset of the test development

process. According to Chen (2009, p. 5), reliability could have profound impact on test validity; that is, if the test lacks consistency in test scores, the scores obtained from this test will become useless, and it will be impossible to claim that the test is valid or has construct validity.

With regard to impact, it is any possible consequence of a test, whether positive or negative, on all stakeholders involved in the use of a test, and it will come into being after the test is used (Bachman & Palmer, 1996; Chapelle, 2012; Chen, 2009; Giraldo, 2019). As a direct result available from a test, test scores could have positive or negative impacts on individuals, educational systems, and society (Bachman & Palmer, 1996) which illustrates how test scores and impact are interrelated.

Quality Relating to Test Implementation

Practicality directly relates to an implementation of a test. It is a relationship between the resources needed for the design, development, and use of test and the resources available to perform these activities (Bachman & Palmer, 1996). If the resources needed exceed the available resources, the test is not practical.

To conclude, Bachman and Palmer's (1996) test usefulness model can be considered a comprehensive guideline encompassing all key concepts to be considered during the language test development process to ensure that a test can be utilized fully as intended. All the six qualities in the model are interconnected, and should be considered in concert with each other rather than being evaluated independently.

Test Usefulness and Standardized Test

Many scholars (e.g., Kaukab & Mehrunnisa, 2016; Kennedy, 2003; Shepherd, 2017) have provided similar definitions of what defines *a standardized test*. Careful analysis of these definitions reveals two underlying traits of standardized tests. The first of this *uniformity* which refers to the fact that anybody taking a standardized test on any subject matter will be given the same type of questions, and they will be required to do that test under the same conditions. The second trait is *consistency* which refers to the fact that all answers to a standardized test are graded in the same way, and the score obtained by each test-

taker is interpreted in the same manner, using the same grading criteria.

Moreover, according to Kennedy's (2003), any standardized test must have four characteristics which are:

> (a) it has been carefully constructed to measure the construct of interest, (b) the conditions under which the examination should be administered are specified and carefully controlled, (c) the way in which responses are scored is specified, and (d) the way in which scores are interpreted, that is, their meaning, follows precise rules. (pp. 8-9)

Test Usefulness and Test Development Process

A review of the literature concerning the test development process revealed two commonalities among the test development processes proposed by several scholars (Council of Europe, 2011; Downing, 2006; Orozco & Shin, 2019). Firstly, the test development process consists of many steps. For instance, according to the Council of Europe's (2011) *Manual for Language Test Development and Examining*, a test development process consists of four steps, i.e., planning, designing, trying out, and informing stakeholders while Orozco and Shin's (2019) test development process comprises five steps namely "design statement, test specifications, test writing, piloting test and result analysis, and test revision" (p. 2). Despite having different numbers of steps involved, these processes are similar to each other. Moreover, many scholars (Alderson et al., 1995; Council of Europe, 2011; Downing, 2006; Green, 2014; Young et al., 2013) agree that the first step of the test development process (i.e., the planning stage) is highly significant, and the test's purpose including its specifications should be identified at this stage. Indeed, according to Alderson et al. (1995), test specifications are "the blueprint to be followed by test and item writers, and they are also essential in the establishment of the test's construct validity" (p. 9).

Along with construct validity, Bachman and Palmer (1996) contend that practicality can play a vital role in every stage of the test

design and development process, and it has a great influence over the form, format, and type of the test to be developed.

Test Usefulness and the CEFR

The CEFR is a framework of reference particularly designed for foreign language education "to provide a transparent, coherent and comprehensive basis for the elaboration of language syllabuses and curriculum guidelines, the design of teaching and learning materials, and the assessment of foreign language proficiency" (Council of Europe, 2021). As a framework of reference, the CEFR classifies language proficiency into three main levels, namely "A" for basic users, "B" for independent users, and "C" for proficient users (Council of Europe, 2001). Aside from these three main levels, each level is further divided into two sub-levels. Thus, in total, the CEFR comprises six levels: A1, A2, B1, B2, C1, and C2 (Council of Europe, 2001).

Each level of the CEFR is accompanied by a description of what foreign language learners at that level can do. The description of each level is collectively called *can-do statements* or *can-do descriptors* (Council of Europe, 2001). Presently, these can-do descriptors can be roughly divided into two types: global scale and illustrative scale descriptors. The global scale descriptors illustrate the overall abilities of what a person at each level can do in terms of using a foreign language for communication. For instance, language learners who are considered basic users (A2) are generally described as those who can use a foreign language to communicate very basic information about themselves and something happening in their everyday life by using very simple terms or short phrases (Council of Europe, 2001, p. 5). On the other hand, the illustrative scale descriptors provide details of what language learners can do in terms of the four basic communication skills: listening, speaking, reading, and writing. For instance, foreign language learners at the A2 level are expected to "catch the main point in short, clear, simple messages and announcements" (Council of Europe, 2001, p. 9).

The illustrative descriptors are very useful for those desiring to prepare a test in alignment with the CEFR as they provide specific details in terms of what language learners can do in relation to each

communication skill. Nevertheless, it is worth noting that the global and illustrative descriptors of the CEFR are not intended to be an exhaustive list of what foreign language learners at each level can do; instead, they provide examples of what the learners at each level can do (Council of Europe, 2011, p. 13).

The CEFR descriptors can be viewed as relating to construct validity in the test usefulness model. Specifically, these descriptors provide details of what should be included in a test to conform with the CEFR. In short, if a test's results are to be aligned with the CEFR, the test's constructs or its contents must be carefully aligned with the descriptors of the CEFR levels if its scores are to align with these levels.

Methodology

The Framework of the Study

To develop the RSU-TEP, the test development process presented in the Council of Europe's (2011) *Manual for Language Test Development and Examining* (henceforth called "test development manual") was adopted as the framework of this study. This framework consists of four stages: planning, design, try-out, and informing stakeholders. In the planning stage, a test development team should identify all fundamental details concerning test construct, test impact, and test practicality (Council of Europe, 2011, p. 20). With regard to test construct and test impact, the team should specify the purpose of the test, the prospective test-takers, the language abilities to be assessed by the test, the scoring of the test, and the intended use of test results, including any possible impacts the test might bring about. While with regard to test practicality, the team should identify the number of prospective test-takers, the deadline for test development completion, the test administration procedures, and test delivery mode. All of these details constitute an initial set of test specifications that will lay a practical groundwork for the test development process.

In the second stage (i.e., design), the test development team is obliged to make decisions on the technical features of the test. These features include the test's duration, number of items, number of items per section, types of item (i.e., selected or constructed responses), the

test's format, the number of marks to give for each item, and characteristics of the rating scale (Council of Europe, 2011, p. 22). At the end of the second stage, a more comprehensive set of test specifications is ready for try-out in the third stage.

In the third stage, the test is tried out to gain useful information to improve the test's quality (Council of Europe, 2011, p. 24). Such information is usually available from the test's item analysis and can be obtained via piloting, and consulting colleagues and stakeholders. This the information is then used to revise the test. The piloting and consultation with colleagues or stakeholders can be repeated until the final version of test specifications is approved.

In the final stage, the final test specifications available at the end of the third stage are made known to stakeholders, especially those who will be directly affected by the test. For example, the prospective test-takers will be informed of the final test specifications so that they know what they will be assessed on in the test.

Context of the Study

This study was conducted in 2019 at Rangsit University (RSU). Rangsit English Language Institute (RELI) was assigned to develop a new standardized English proficiency test called the "RSU-TEP" because it is the academic division responsible for providing foundation English courses to students at all levels in RSU. In 2019, there were 23 full-time faculty members (21 Thais and two Filipinos) in RELI. To ensure that this task was achieved smoothly, a committee called the "RSU-TEP Development Committee" consisting of 13 members was appointed (henceforth called "committee"). The committee had 12 Thai and one Filipino staff members, all of whom held at least a master's degree in Teaching English as a Foreign Language, English Language Teaching (ELT), or ELT-related fields, and had been teaching foundation English courses at RELI for at least two years. Additionally, all of them attended a workshop on test development and test alignment with the CEFR given by an expert in language testing and the CEFR to ensure that they have a good grasp of test development and test alignment with the CEFR.

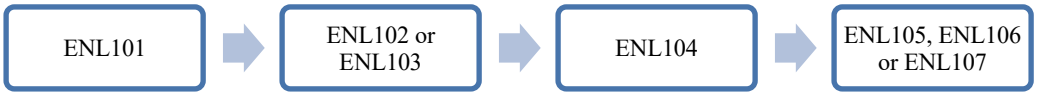
Participants

One hundred and ninety-six undergraduates from various faculties in RSU were purposively selected to participate in this study. These participants were in different years of their studies, ranging from freshmen to seniors. The main criterion used in selecting these participants was their different English abilities to comply with one fundamental requirement of a good multiple-choice test; that is, the test must distinguish test-takers at an advanced level from test-takers at a beginner level (Quaigrain & Arhin, 2017). This requirement suggests that the participants should have different levels of English abilities.

To ensure that participants with different English abilities were selected, two factors were considered: the foundation English courses they were taking and their programs of study. In 2019, RSU's undergraduates were required to take at least four foundation English courses as shown below for graduation.

Figure 1

Structure of Required Foundation English Courses for RSU Undergraduates in 2019



According to Figure 1, except for ENL101, there is a prerequisite for each foundation English course. That is, ENL101 is a prerequisite for ENL102 and ENL103, ENL102 or ENL103 is a prerequisite for ENL104, and ENL104 is a prerequisite for ENL105, ENL106, and ENL107. Apart from these courses, ENL100, a preparatory course, is an additional required English course for freshmen whose Ordinary National Education Test (O-Net) score in English is below 21. Nevertheless, ENL100 is not counted toward students' graduation.

Therefore, students taking ENL100 were selected as the participants with a beginner level of English in this study because their O-Net scores in English were below 21. Additionally, from the researcher's experience in teaching ENL100, most ENL100 students

had limited vocabulary and difficulty in communicating in everyday English, whether in spoken or written form.

Participants with an intermediate level of English were selected from students taking ENL102, ENL103, and ENL104. ENL102 focuses on vocabulary, expressions, and language structures in science and technology, ENL103 focuses on vocabulary, expressions, and language structures in social science and humanities, and ENL104 emphasizes oral communications on general topics. Before taking these courses, students must pass the prerequisite for each course. The prerequisite requirement and the contents of these courses were believed to help improve the English capacities of students in these courses to reach an intermediate level. Thus, students taking these courses could be assumed to have an intermediate level of English in contrast to the beginners enrolled in ENL100.

Finally, the participants with advanced English were (i) first-year pharmacy students taking the custom-made ENL101 and (ii) aviation students taking ENL105. The pharmacy students were considered to have advanced English because these students had passed the university's admission test which included English; this fact suggests that their English capacities had already been assessed, and those who were admitted could be regarded as being relatively proficient in English. Also, the contents of the ENL101 course they were taking had been customized and geared toward pharmacy-related topics which are quite advanced. As for the aviation students, most of these students had studied in international schools or abroad before joining RSU. Additionally, the course they were taking (English for the International Workplace) was their last required course, meaning that they had passed all the three prerequisite courses. The prerequisite courses should, by and large, have helped enhance their English capacities. Besides, all RELI instructors unanimously agreed that both pharmacy and aviation students were generally proficient in English; thus, they were selected as participants with advanced English in this study.

After purposive selection, these participants were divided into two groups. The first group consisting of 108 students with different English capacities was the test-takers of the RSU-TEP's first draft

during the piloting stage. The second group consisting of 88 students with different English capacities was the test-takers of the revised RSU-TEP during the try-out stage. Both stages were implemented for test item analysis to examine the RSU-TEP's quality as suggested by the test development manual.

Instruments

Two instruments were employed in this study: CEFR resources and an item analysis web application.

CEFR Resources

During the course of preparing test items of the RSU-TEP, the committee had consulted various sources of the CEFR to observe the test's validity requirement. In this regard, three sources were used as the main reference. The first source was the Council of Europe's website (https://www.coe.int/en/web/common-european-framework-reference-languages) which provides comprehensive information about the CEFR. The second source was a publication entitled *Common European Framework of Reference for Languages: Learning, teaching, assessment – Structured overview of all CEFR scales.* This source, also produced by the Council of Europe, provides information about the CEFR global and illustrative scale descriptors, and was used as a practical guideline in preparing the test items. The third source was a publication called *A Core Inventory for General English* which was jointly produced by the British Council and EAQUALS. This source provided an inventory of topics, language functions, grammatical points, vocabulary and expressions, and scenarios relevant to the CEFR levels from A1 to C1. These three sources were used as the main references for CEFR because they were prepared by organizations with expertise in the CEFR.

RSU's Item Analysis Web Application

The item analysis web application developed by Asst. Prof. Dr. Wutthipong Chinnasri and Asst. Prof. Siriwan Wasukree of the College of Digital Innovation and Information Technology at RSU was used to examine the RSU-TEP's quality in terms of its reliability, difficulty index, and discrimination power. According to Chinnasri and

Wasukree (2015), this web application can be used with an objective test like a multiple-choice test having four or five options. The analysis results available from this web application include descriptive statistics (i.e., maximum score, minimum score, mean score of test-takers, and standard deviation); test's reliability measured by Kuder-Richardson 20 (KR-20) co-efficiency; difficulty index (p-value); item discrimination power (r); and a summary of item analysis results (Chinnasri & Wasukree, n.d., p. 6).

Concerning the test's reliability result available from the web application, it was calculated via the following formula:

KR-20 = $[n/n-1] * [1-(\Sigma p*q)/Var]$

where

KR-20 = internal consistency reliability co-efficiency of the test,

n = total number of items in the test,

p = proportion of test-takers answering the item correctly,

q = proportion of test-takers answering the item incorrectly,

Var = variance of the total scores of all test-takers.

Σ = sum up

Normally, the reliability's value can range from zero to one whereby zero signifies that a test has no reliability and one means that a test is highly reliable and homogeneous (Obon & Rey, 2019; University of Washington, 2021). Test reliability as measured by KR-20 shows how each item in the test relates to other items "in terms of who answers them correctly" (University of Washington, 2021). According to the web application, if the KR-20 value is greater than 0.60, the test is considered to have good reliability (Chinnasri & Wasukree, n.d.).

As for the difficulty index (p-value) available from the web application, it is calculated by dividing the number of test-takers who answer an item correctly by the total number of test-takers answering such an item. Generally, the p-value is in a range between zero and one. If the p-value of an item is high or close to one, the item is

interpreted as easy (Quaigrain & Arhin, 2017). To facilitate the application users, Chinnasri and Wasukree (2019) provided a detailed interpretation of the p-value available from the application as shown below.

Table 1

P-value Interpretation

P-value	Interpretation
0.00-0.19	Very difficult
0.20-0.39	Difficult
0.40-0.59	Moderate
0.60-0.79	Easy
0.80-1.00	Very easy

Item discrimination power (r) was calculated by the application using the formula:

$$r = P_H - P_L$$

where

r = item discrimination power,

P_H = number of correct answers made by test-takers in the upper group / total number of all test-takers in the upper group

P_L = number of correct answers made by test-takers in the lower group / total number of all test-takers in the lower group

The item discrimination power can range between -1.0 and 1.0 (Obon & Rey, 2019). If an item has a high discrimination power value, that item is considered a good item as it can differentiate the test-takers having mastery of the test construct from those who have not. To facilitate the application users, Chinnasri and Wasukree (2019) provided a detailed interpretation of the discrimination power value as shown below.

Table 2

Discrimination Power Value Interpretation

Discrimination power value (r)	Interpretation
-1.00 - 0.09	Very low discrimination power
0.10 - 0.19	Low discrimination power
0.20 – 0.39	Moderate discrimination power
0.40 – 0.59	Good discrimination power
0.60 – 1.00	Very good discrimination power

Data Collection and Data Analysis

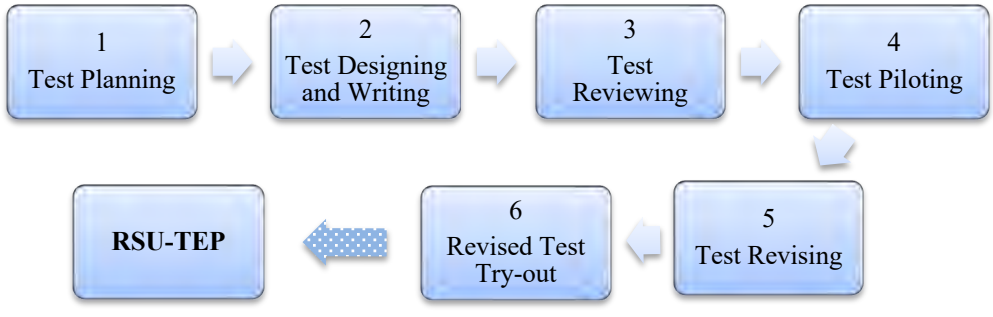
Data collection and data analysis were performed during the reviewing, piloting and revised test try-out stages of the RSU-TEP development process. Data analysis was performed through an item-objective congruency (IOC) test and the item analysis web application to examine the RSU-TEP's validity and reliability, respectively.

Results and Discussion

Guided by the four-stage test development process presented in the Council of Europe's (2011) test development manual, the RSU-TEP underwent a six-stage test development process as shown in Figure 2.

Figure 2

RSU-TEP Development Process



Stage 1: Test Planning

The first meeting of the RSU-TEP Development Committee was held in February 2019 to reach a consensus on the RSU-TEP's test specifications. Table 3 presents the committee's decisions on key specifications of the RSU-TEP.

Table 3

RSU-TEP's Specifications

RSU-TEP's specifications	Committee's decisions
Test's name	RSU-TEP
Prospective test-takers	RSU's undergraduates and graduates
Test's objective	To assess test-takers' proficiency in English with scores aligned with the CEFR
Cut-off scores needed from the test for CEFR alignment	Four cut-off scores corresponding to A2, B1, B2, and C1 of the CEFR
Test's constructs	Listening comprehension, language structure and expressions, and reading comprehension
Test's format and item number	A test of 100 multiple-choice items in which each item has four options
Test's duration	One hour and thirty minutes
Test's delivery mode	Computer-based
Test's scoring	One point for a correct answer to each item and zero point for an incorrect answer
Test's completion date	In six months (i.e., September 2019)

According to the reviewed literature, the test's objective should be identified at the beginning of the test development process (Council of Europe, 2011; Downing, 2006; Green, 2014; Young et al., 2013). In this regard, the committee agreed that the RSU-TEP's objective was to assess test-takers' proficiency in English with test scores aligned with the CEFR. For the present study, it was also crucial to identify the number of cut-off scores needed for the CEFR alignment in this stage because this information can have profound impacts on the test content and the number of items to be included in the test. On this issue, the committee decided that four cut-off scores, namely, A2, B1, B2, and C1, were needed from this test for the CEFR alignment. This

decision was made in recognition of a big gap in the English abilities of the RSU's students which could result from the fact that students with any level of English ability are eligible to study at RSU. The four cut-off scores were also believed to provide a clearer picture of the English abilities of the RSU's students.

After identifying the test's objective and the number of cut-off scores needed, the committee discussed the number of test items to be included in the RSU-TEP. In general, the more cut-off scores are required, the more test items are needed to ensure the test's reliability. This is in line with McCowan and McCowan's (1999, p.10) postulation that longer tests usually yield higher reliability values. Because four cut-off scores were needed from the test, the committee decided that the RSU-TEP should consist of 100 items which would be classified into four groups according to the cut-off scores needed: 10 items for A1, 40 items for B1, 40 items for B2 and 10 items for C1. Items for B1 and B2 levels represented the majority of the test items because B1 and B2 were the targets set for undergraduates and graduates, respectively. These 100 items were also believed to fit the test duration of one hour and thirty minutes which was viewed by the committee as proper because this duration could help minimize the test-takers' anxiety and fatigue from doing the test. Other specifications (e.g., delivery mode, scoring method, and completion date) were decided from the committee's thorough consideration of the available resources of RSU and RELI in accordance with the practicality quality in the test usefulness model.

Stage 2: Test Design and Writing

In compliance with the test's specifications agreed upon in Stage 1, the committee decided that the RSU-TEP be divided into three parts: listening comprehension, structure and expressions, and reading comprehension; all of which were the test's constructs. Because the RSU-TEP was intended to be aligned with the CEFR and four cut-off scores were needed, its content was designed to correspond with the CEFR A2, B1, B2, and C1 descriptors. After much deliberation, the committee resolved that RSU-TEP be designed as per the details below.

Table 4

RSU-TEP Components

Description	No. of Questions	Marks	Time
Part I Listening Comprehension			
Section 1: Short Conversations	15	40	25 minutes
Section 2: Short Talks	15		
Section 3: Lectures	10		
Part II Structure and Expressions			
Section 1: Incomplete Sentences	15	30	20 minutes
Section 2: Text Completion	15		
Part III Reading Comprehension			
Section 1: Short Texts	10	30	45 minutes
Section 2: Short Passages	10		
Section 3: Academic Passage	10		
Total	100	100	90 minutes

Based on Table 4, the listening comprehension part is the part containing the most marks in the RSU-TEP (i.e., 40 marks). The committee's decision to allocate most marks to this part was derived from the fact that listening skills are emphasized in all English foundation courses provided by RELI, and this skill corresponds with the communicative teaching approach adopted at RELI. The committee also felt that listening skills are of paramount importance to students, especially in this era of globalization when knowledge and information come in a multimedia form which can include voice and sound.

To ensure that the listening comprehension part's content was as valid, interactive, and authentic as possible in the multiple-choice format, this part was designed to include (i) short conversations on everyday life topics which correspond to the CEFR A2 and B1 descriptors; (ii) short talks on concrete and abstract topics which correspond to the CEFR B2 descriptor; and (iii) lectures on academic topics which correspond to the CEFR C1 descriptors. The committee also agreed that the test-takers would be allowed to listen to each audio once only. Additionally, either native or near-native English speakers would be sought to produce audio tracks of the content in this part as far as practical within the available budget.

The structure and expressions part, containing thirty items and representing thirty percent of the RSU-TEP's total points, was designed to assess the test-takers' ability in using the language correctly and appropriately. The committee believed that this part could implicitly reflect the test-takers' writing ability given that good writing requires a solid knowledge of language structure and expressions. As for the number of test items in this part, the committee felt that this part should contain fewer items than the listening comprehension part because RELI's teaching communicative approach emphasizes fluency more than accuracy.

To ensure that the structure and expressions section of the test had test items that correspond to the CEFR A2 to C1 descriptors, these test items cover a variety of expressions and language structures used to perform various functions as outlined by North et al. (2010), e.g., describing places (A2), expressing opinions (B1), speculating (B2), and expressing probability (C1). Apart from adhering to the CEFR descriptors, the expressions and language structures in this part were carefully selected to make sure that they were within the test-takers' real-life experience so as to make the test as authentic as possible.

The reading comprehension part was designed to assess test-takers' reading ability from beginning to advanced levels, e.g., reading for gist and details, understanding vocabulary in context, and making inferences. This part was included in the RSU-TEP because reading skills are very important at present when knowledge and information are often disseminated in a written form.

Three types of text were included in the test's reading comprehension section. These texts were carefully chosen to ensure that they corresponded to the CEFR descriptors from A2 to C1 and that they were within the realm of the test-takers' real-life experience. Examples of these texts were an email to a friend which corresponds to A2, a job advertisement which corresponds to B1, a machine operation manual which corresponds to B2, and an argumentative essay which corresponds to C1.

Stage 3: Test Reviewing

This stage was added to the original four-stage process in the Council of Europe's (2011) test development manual. Principally, it was added to (i) enhance the validity of the RSU-TEP's content and (ii) ensure that the test was free from common flaws (e.g., spelling mistakes and format). To achieve these objectives, after the committee finished writing all test items, they combined the items to create the first draft of the RSU-TEP. This first draft was then circulated among the committee's members to double-check each item's consistency with the required CEFR descriptors and make revisions as needed. After that, it was given to five ELT experts to perform an item-objective congruency (IOC) test of the RSU-TEP to examine its content validity. These experts held at least a master's degree in ELT or ELT-related field, and their English teaching experience was over 10 years. They also had expertise in test writing and the CEFR. The IOC test revealed that all items on the RSU-TEP were rated in the range between 0.6 and 1 which was above the acceptance value of 0.5 suggested by Rovinelli and Hambleton (1977, as cited in Turner & Carlson, 2003); thus, the RSU-TEP could be considered to have good validity.

After the first draft of the RSU-TEP was validated through the IOC test, it was given to a native English speaker for proofreading. The RSU-TEP was then revised as per the comments of the native English speaker. At the end of stage 3, the first draft of the RSU-TEP was ready for piloting in the following stage.

Stage 4: Test Piloting

In August 2019, the RSU-TEP's first draft which had been prepared in RELI's Moodle Learning Management System (LMS) was piloted with 108 undergraduates with mixed abilities of beginner, intermediate and advanced levels. The test piloting was conducted in RELI's self-study computer labs in three different sessions within the same week. Before taking the test, the participants were informed of the research's objective and were requested for consent to participate in the study. After giving their consent, the participants registered on the LMS by giving their personal data (e.g., name, student identification number, and faculty). After registration, they could do the test part by

part. They had to finish Part I before moving on to Part I and Part III, respectively. As for the scoring method, one mark was awarded for a correct answer while zero marks were awarded for an incorrect answer. The time for doing each part was limited as shown in Table 4. While doing each part, the participants could see a countdown clock to know how much time they had left. After finishing the three parts, the participants saw their scores in each part, and they had to log out of the LMS.

The test piloting revealed that the duration of one hour and thirty minutes set by the committee is appropriate because most participants finished the test five to ten minutes before the set time expired. After all participants took the test, the item analysis web application was used to examine the RSU-TEP's reliability and other related aspects. The item analysis results showed that the reliability and difficulty index of the RSU-TEP's first draft were quite satisfactory; however, the results concerning item discrimination power warranted improvement.

Stage 5: Test Revising

After obtaining the item analysis results from the web application, the committee held its second meeting to discuss how to improve the RSU-TEP's quality, especially for the items which were indicated by the web application as very easy, very difficult, or having low or very low discrimination power. After these items were revised according to the meeting's suggestions, the revised RSU-TEP was ready for try-out in the next stage.

Stage 6: Test Try-out

In September 2019, the revised RSU-TEP was piloted with a second group of participants (i.e., 88 undergraduates) in the same manner as its first draft. The item analysis web application was used again to examine the revised RSU-TEP's reliability. A comparative summary of the item analysis results of the RSU-TEP's first draft and the revised version is presented in Table 5.

Table 5

Item Analysis Results of RSU-TEP's First Draft and Revised Version

Description	Item Analysis Results	
	1st Draft (n = 108)	Revised Version (n = 88)
Descriptive statistics		
Maximum score	87 (out of 100)	91 (out of 100)
Minimum score	17 (out of 100)	17 (out of 100)
Mean score	38.92	48.36
Standard deviation (SD)	13.13	18.23
Test's reliability		
KR-20 co-efficiency	0.88	0.94
Difficulty index (p-value)		
Very difficult	5 items (5%)	2 items (2%)
Difficult	50 items (50%)	29 items (29%)
Moderate	35 items (35%)	43 items (43%)
Easy	9 items (9%)	25 items (25%)
Very easy	1 item (1%)	1 item (1%)
Item discrimination power (r)		
Very good discrimination power	2 items (2%)	3 items (3%)
Good discrimination power	14 items (14%)	31 items (31%)
Moderate discrimination power	49 items (49%)	42 items (42%)
Low discrimination power	15 items (15%)	15 items (15%)
Very low discrimination power	20 items (20%)	9 items (9%)
Suggestions from item analysis web application		
Items that are suitable for use	65 items (65%)	76 items (76%)
Items that should be revised before using	29 items (29%)	22 items (22%)
Items that should not be used	6 items (6%)	2 items (2%)

RSU-TEP's First Draft

The first item analysis conducted after the test piloting stage revealed that the reliability and difficulty index of the RSU-TEP's first draft were acceptable. That is, the KR-20 co-efficiency was 0.88, and only 9% and 1% of the total items were analyzed as easy and very easy items, respectively. However, the result concerning item discrimination power warrants a revision of the test given that 35% of the total items were analyzed as having low and very low discrimination power. This

result is fully in line with the suggestions made by the web application that 65 items had good quality and were suitable to be included in the test whereby 29 items should be revised before use, and 6 items should not be used.

RSU-TEP's Revised Version

The second item analysis conducted after the test try-out stage showed that the revised RSU-TEP had remarkable improvement in terms of its reliability and item discrimination power.

Test Reliability

The revised RSU-TEP's reliability, assessed by KR-20, was 0.94 which was improved from 0.88 of the first draft. This high value of KR-20 showed that all items in the test "are measuring the same thing" (Mukherjee & Lahiri, 2015, as cited in Obon & Rey, 2019, p. 503). The KR-20 of 0.94 is also considered appropriate for the RSU-TEP which is a high-stake test given that most high-stake tests are generally required to have an internal reliability of 0.90 or higher (Quaigrain & Arhin, 2017, p. 4).

Item Difficulty Index

One obvious result of the revision of the RSU-TEP's first draft was that the number of "difficult" and "very difficult" items decreased from 5 and 50 items to 2 and 29 items, respectively. This reduction in the number of "difficult" and "very difficult" items altogether accounted for 31% of the total test items. On the other hand, the number of "moderately difficult" and "easy" items increased from 35 and 9 items to 43 and 25 items, respectively. These results could be regarded as conducive to aligning RSU-TEP's results with the CEFR in future because if the test had too many "difficult" items, it could be viewed as particularly designed for a certain group of test-takers. Also, a large number of "difficult" items could affect the number of items to be set for the A2 cut-off score which may consist of a few items only.

Item Discrimination Power

The other noticeable result from the revision of the RSU-TEP's first draft was the increase in the number of items with "good" discrimination power from 14 to 31 items. Also, the items with "very

low" discrimination power were decreased from 20 to 9 items. Item discrimination power is considered one fundamental trait of a good test; that is, the test must be able to distinguish test-takers who have mastery of the test from those who have not. It can be said these improvements in discrimination power also helped enhance the RSU-TEP's reliability.

Conclusion

This study aimed to illustrate how to develop a standardized English proficiency test that can be aligned with the CEFR. Such a test was developed through six stages: test planning, test designing and writing, test reviewing, test piloting, test revising, and revised test try-out. The resulting test was found to be highly reliable given that its reliability, assessed by KR-20, was 0.94. It could also be considered a test with good validity because all test items had been carefully designed in compliance with the CEFR A2 to C1 descriptors. Moreover, the IOC test applied to it affirmed its content validity. All these results certainly pave the way for the future alignment of its results with the CEFR. Despite the limited time available for test development, this study shows how to develop a local standardized English proficiency test in multiple choice format in a Thai university setting.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher, & F. Davison (Eds.), *The Routledge handbook of language testing* (pp. 21–33). Routledge.

Chen, L. (2009). *Assessing the test usefulness: A comparison between the old and the new College English Test Band 4 (CET-4) in*

China. Retrieved from diva-portal.org website: https://www.diva-portal.org/smash/get/diva2:228454/FULLTEXT01.pdf

Chinnasri, W., & Wasukree, S. (2015). A development of a web application for item analysis on objective test. *VRU Research and Development Journal Science and Technology, 10*(1), 1–17.

Chinnasri, W., & Wasukree, S. (2019, September). *Objective test item analysis through the use of web applications*. Handout provided during the workshop of Rangsit English Language Institute, Phathumthani.

Chinnasri, W., & Wasukree, S. (n.d). *Manual for a web application for item analysis on objective test*. Retrieved from https://isdc.rsu.ac.th/itemanalysis/objective/

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

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

Council of Europe. (2021). *Common European Framework of Reference for Languages (CEFR)-Introduction*. https://www.coe.int/en/web/common-european-framework-reference-languages/introduction-and-context

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Lawrence Erlbaum Associates.

Giraldo, F. (2019). Designing language assessments in context: Theoretical, technical, and institutional considerations. *HOW, 26*(2), 123–143.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge.

Kaukab, S. R., & Mehrunnisa, S. (2016). History and evolution of standardized testing–A literature review. *International Journal of Research-Granthaalayah, 4*(5), 126–132.

Kennedy, E. (2003). *Raising test scores for all students: An administrator's guide to improving standardized test performance*. Corwin Press.

McCowan, R. J., & McCowan, S. C. (1999). *Item analysis for criterion-referenced tests*. https://files.eric.ed.gov/fulltext/ED501716.pdf

North, B., Ortega, A., & Sheehan, S. (2010). *A core inventory for general English*. https://www.eaquals.org/wp-content/uploads/EAQUALS_British_Council_Core_Curriculum_April2011.pdf

Obon, A. M., & Rey, K. A. M. (2019). Analysis of multiple-choice questions (MCQs): Item and test statistics from the 2nd year nursing qualifying exam in a University in Cavite, Philippines. *Abstract Proceedings International Scholars Conference 7*(1), 499–511. https://doi.org/10.35974/isc.v7i1.1128

Office of the Basic Education Commission. (2014). *Guiding practice issued under the Ministry of Education's Notification re: Policy for English education reform*. http://www.phatthalung2.go.th/myoffice/2557/data/tkk2/25570408_114829_9281.pdf

Office of the Higher Education Commission. (2016). *Notification of the Office of the Higher Education Commission re: Policy to enhance English standards in higher education institutions*. http://academic.swu.ac.th/Portals/43/35.pdf

Orozco, R. A. Z., & Shin, S. Y. (2019). Developing and validating an English proficiency test. *MEXTESOL Journal*, *43*(3), 1–11.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education 4*(1), 1–11.

Shepherd, K. J. (2017). *Measuring up: Standardized testing and the making of postwar American identities, 1940–2001* [Unpublished doctoral dissertation]. University of South Florida.

Turner, R. C., & Carlson, L. (2003). Indexes of item-objective congruence for multidimensional items. *International Journal of Testing, 3*(2), 163–171.

University of Washington. (2021). *Understanding item analyses*. https://www.washington.edu/assessment/scanning-scoring__trashed/scoring/reports/item-analysis/

Weir, C. J. (1993). *Understanding and developing language tests*. Prentice Hall.

Young, J. W., So, Y., & Ockey, G. J. (2013). *Guidelines for best test development practices to ensure validity and fairness for international English language proficiency assessments*. Educational Testing Service. https://www.ets.org/s/about/pdf/best_practices_ensure_validity_fairness_english_language_assessments.pdf