Language Assessment at a Thai University: A CEFR-Based Test of English Proficiency Development

BUDI WALUYO

School of Languages and General Education, Walailak University, Thailand

ALI ZAHABI*

School of Liberal Arts, King Mongkut's University of Technology Thonburi, Thailand

LUKSIKA RUANGSUNG

School of Languages and General Education, Walailak University, Thailand

Corresponding author email: ali.zaha@kmutt.ac.th

Article information

Article history:
Received: 5 Apr 2023
Accepted: 1 Sep 2023
Available online: 24 Jan 2024

Keywords:
CEFR
Classical Test Theory
English proficiency
Language testing
Test development

Abstract

The increasing popularity of the Common European Framework of Reference (CEFR) in non-native English-speaking countries has generated a demand for concrete examples in the creation of CEFR-based tests that assess the four main English skills. In response, this research endeavors to provide insight into the development and validation of a CEFR-based test aimed at evaluating undergraduate students' English proficiency for placement tests and exit exams. The CEFR served as the framework for item development while Classical Test Theory informed the test evaluation process. A sample of 2,248 first-year students participated in Testing 1 and 3,655 first- and second-year students took part in Testing 2. The results of the analysis of the multiple-choice listening and reading tests indicated favorable levels of item difficulty and discrimination indices, as well as high reliability coefficients obtained from Cronbach's alpha, Kuder-Richardson, and split-half reliability. The correlation and regression analyses revealed close relationships between the subtests and between each subtest and the total score, supporting the test's criterion validity. The study also demonstrated significant predictive validity on TOEIC scores. The findings of this study offer implications for the development of university-level English proficiency tests that integrate CEFR levels and CTT analysis.

INTRODUCTION

The Common European Framework of Reference (CEFR) for languages was introduced in 2001 and has since become widely adopted as a guide for language policies and teaching practices both within and outside Europe (Nagai, 2020). The CEFR provides a streamlined approach to language proficiency through its use of levels and descriptors, focusing on language use in real-life contexts. Moreover, the framework accommodates for multimodality and allows for flexibility in various situations (Figueras, 2012). In Asia, the CEFR has been adapted and implemented in language learning, teaching, and assessment in Japan (Negishi et al., 2013), Taiwan (Wu & Wu, 2007), and Vietnam (Nguyen, 2016). In some Asian countries, the CEFR serves as a framework for determining language proficiency, such as the General English Proficiency Test (GEPT) in Taiwan (Wu, 2019), the Vietnamese Standardized Test of English Proficiency (VSTEP) in Vietnam (Quynh, 2019), and the Test in Practical English Proficiency (EIKEN) in Japan (Dunlea et al., 2019), for use in universities, government, and industry. Thailand has applied the CEFR to its English teaching and learning practices at all levels since 2014 (Anantapol et al., 2018; Rofiah et al., 2022). However, despite its widespread usage, the CEFR has been criticized for lacking direction in language testing, providing only definitions and descriptors of language proficiency without specifying how to measure test item validity and reliability (Weir, 2005). The framework's multimodality and multi-interpretation capabilities are both strengths and limitations of the CEFR. As such, it is suggested that the CEFR be used as a source of inspiration and guidance, serving as a collection of descriptive samples for comprehending language proficiency (North, 2014). The most recent version of the CEFR, released in 2018, maintains these features (Council of Europe, 2018). Further empirical investigations are needed for the effective adoption and implementation of the CEFR in language teaching and assessment.

In Thailand, higher education institutions are mandated to implement a CEFR-aligned English proficiency examination (Cheewasukthaworn, 2022). This study endeavors to bridge the existing gaps between the adoption of CEFR standards and the actual development of such an assessment tool, encompassing listening, speaking, reading, and writing proficiencies. The impetus behind this undertaking stems from Walailak University's unwavering commitment to blend language assessment with academic excellence, concurrently addressing the linguistic requirements of its students while adhering to the standards set by the Thai Ministry of Education. The paramount significance of this project is underscored by Walailak University's pivotal role in elevating English language skills within Thailand's higher education landscape. Nevertheless, within the realm of English Language Teaching (ELT), despite a burgeoning interest among educators and researchers in crafting CEFR-based tests, a dearth of comprehensive studies is evident. Previous research has predominantly fixated on singular skills like speaking (Liu & Jia, 2017; Waluyo, 2020) or writing (Harsch & Seyferth, 2020), while listening and reading have garnered comparatively limited attention. In parallel, global English proficiency examinations have initiated efforts to align their scoring systems with CEFR levels, exemplified by TOEFL ITP (Tannenbaum & Wylie, 2008; Pratiwi & Waluyo, 2022), Cambridge ESOL (Khalifa & Ffrench, 2009), TOEFL IBT (Papageorgiou et al., 2015), CU-TEP (Wudthayagorn, 2018), GEPT (Brunfaut & Harding, 2014), and others. This phenomenon underscores a growing interest in leveraging CEFR levels to gauge English competence. Nevertheless, the dearth of empirical studies regarding the creation of a comprehensive English proficiency test, along with the absence of concrete examples demonstrating the alignment of a CEFR-based test with international standardized assessments by a higher educational institution, underscores the exigency for further research. This paper endeavors to bridge these gaps by documenting the development of a CEFR-based English proficiency test and assessing its predictive validity in relation to an international English examination.

In the study's context, the Ministry of Education in Thailand launched an English education reform that required university instructors to design their curriculum and teaching practices based on the CEFR framework. However, it was reported by Kanchai (2019) that these instructors lacked a thorough understanding of the CEFR. The poor English proficiency of students at Walailak University, even in their senior year, prompted the implementation of a policy mandating the completion of six English courses. After these courses, students were evaluated through an in-house examination, with scores mapped to the CEFR levels to determine their English proficiency. Despite the efforts to align with the government's policy, the University faced challenges in mapping their scores to CEFR levels as the only available option, the Chulalongkorn University Test of English Proficiency (CU-TEP), was insufficient. To address this limitation, Walailak University developed an English proficiency test aligned with the CEFR framework, which addresses the students' language needs while fulfilling the government and university's policy (Dimova et al., 2022).

This study presents the stages involved in the creation of the "WUTEP," which stands for Walailak University Test of English Proficiency, a CEFR-based assessment of English proficiency. It outlines the design, development, and validation of a CEFR-aligned English proficiency test that assesses four crucial skills: listening, speaking, reading, and writing. In December of 2017, Walailak University, located in Southern Thailand, established an academic team composed of foreign and Thai English lecturers with the goal of developing a standardized CEFR-based test of English proficiency. This test was intended to be utilized to gauge the English proficiency of Walailak University students on an annual basis as a measure of their English progress over the course of a single academic year. Additionally, the test was meant to serve as an option for fulfilling the university's English proficiency requirement for both incoming local and international faculty and graduate students. To meet these expectations, the test was first required to be based on a widely recognized framework. The CEFR was selected for this purpose as it was officially declared as the basis for English teaching and learning at all levels in Thailand by the Ministry of Education in 2014 (Anantapol et al., 2018; Waluyo & Apridayani, 2021). Secondly, Classical Test Theory (CTT) was employed as the evaluation basis of the test due to its comprehensive procedure for test item development, evaluation, and scaling, which is crucial in the validation process (DeVellis, 2006). Finally, the test development process entailed mapping the test onto other widely recognized international standardized tests, such as International English Language Testing System (IELTS) and Test of English for International Communication (TOEIC), to generate comparable scores that can be interpreted by other institutions and candidates.

THEORETICAL FRAMEWORKS

This section details the theoretical frameworks used as the basis of item constructions (i.e., the CEFR) and test validation (i.e., Classical Test Theory) of WUTEP.

The Common European Framework of Reference (CEFR) for Languages

The CEFR was initially used by ministers of the European Union as a standard to check language skills in 2001 (Council of Europe, 2001). In the last decade, the framework has gradually been accepted by countries within and outside of Europe and is considered one of the accepted standards of evaluation for foreign language proficiency, especially English (Little, 2006). As a language framework, the CEFR accentuates such aspects as an Action-Oriented Approach - emphasizing on what learners can do with the language (action oriented) and viewing learners as Social Agents – stressing on the importance of learners to take responsibility for their own learning, which may involve various personal measures, including goal setting as well as reflection of the language progress and process (Nagai et al, 2020). In the latest documents of the CEFR, the Council of Europe (2018) details that at the heart of the CEFR lies the Descriptive Scheme, which highlights the descriptions of overall language proficiency encompassing general competences, communicative language competences, communicative language activities, and communicative language strategies; and the Common Reference Levels, categorizing the learners into six levels of proficiency: A1–A2 (basic users), B1–B2 (independent users), and C1–C2 (proficient users) (Council of Europe, 2001). The latter one has been widely known and is used in recognizing learners' proficiency levels before and after learning the target language.

To date, the application of the CEFR as a framework in English proficiency tests in higher education, within and outside of Europe, has mainly been directed to measure students' proficiency levels as part of entry requirements (Deygers et al., 2018; Piccardo, 2020). Nevertheless, Harsch (2018, p. 1) argues, "... the CEFR alone cannot guarantee that different institutions and stakeholders will use it in a comparable way and come to comparable interpretations when employing and interpreting its proficiency scales." In this instance, it is important to underline that although the CEFR has a clear description of its global reference levels, different test formats and scoring systems should be expected across English proficiency tests. Furthermore, in English proficiency test development, to ensure the alignment of the test with the CEFR, different procedures are normally implemented for measuring different English skills. For instance, for the writing test, Harsch and Seyferth (2020) created a CEFR-based writing checklist that incorporated proficiency-oriented learning outcomes with classroom-based and achievement-oriented assessment goals. Meanwhile, Borger (2019) involved external raters who made assessments of recorded speaking tests against the CEFR scales in developing a speaking test. Then, listening and reading tests commonly include multiple-choice questions in which the constructs of each question are developed in accordance with the CEFR levels (Kim & Crossley, 2020). Figure 1 below illustrates the constructed CEFR-based development procedures from previous studies.

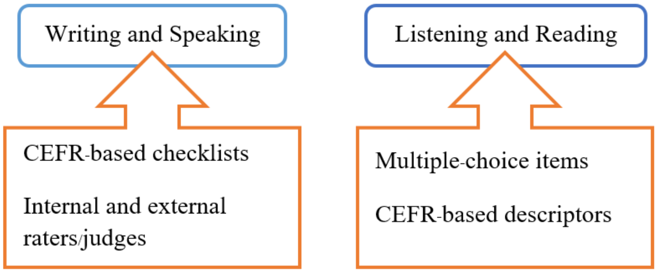

Figure 1 Construct CEFR-based development procedures

For a more detailed guideline, North et al. (2010) developed a core inventory elaborating features of materials in English language teaching (ELT) for each CEFR level, from A1 to C1. C2 is excluded since it is extremely rare to be found among English learners and in an ELT context. The inventory is comprised of functions, grammar, discourse markers, vocabulary, and topics. Some points elaborated in North et al.'s (2010) inventory can be seen in Table 1. This inventory has been used as a reference point in the development of standardized tests of English proficiency such as the Aptist test by the British Council (O'Sullivan & Dunlea, 2015). In Asia, Moser (2015) contends that the inventory may serve an important role in transforming a knowledge-based language curriculum to a competency-based one. Such a belief seems to be visible and followed-up in the paper written by Hiranburana et al. (2017), who discussed the CEFR from Thai perspectives and experiences. In the present study, the inventory serves as a foundational reference point in test item construction in listening and reading tests and assessment rubrics in speaking and writing tests.

Table 1

Some of the details from the core inventory by North et al. (2010)

	A1	A2	B1	B2	C1
Functions	- Directions - Describing habits and routines	- Describing habits and routines - Describing past experiences	- Checking understanding - Describing experiences and events	- Critiquing and reviewing - Describing experiences	- Conceding a point Critiquing and reviewing constructively
Grammar	- Adjectives: common and demonstrative - Adverbs of frequency	- Adjectives - Comparative, - Use of than and definite article - Adjectives - Superlative - Use of definite article	- Adverbs - Broader range of intensifiers such as too, enough	- Adjectives and adverbs - Future continuous	- Futures (revision) - Inversion with negative adverbials
Discourse Marker	- Connecting words, and, but, because	- Linkers: sequential – past time	- Connecting words expressing cause and effect, contrast etc.	- Connecting words expressing cause and effect, contrast etc.	- Linking devices - Logical markers
Vocabulary	- Food and drink - Nationalities and countries	- Adjectives: personality, description, feelings - Food and drink	- Collocation - Colloquial language	- Collocation - Colloquial language	- Approximating (vague language) - Collocation
Topic	- Family life - Hobbies - Pastimes	- Education - Hobbies - Pastimes	- Books - Literature - Education	- Arts - Literature	- Arts - Literature

Classical Test Theory (CTT)

In test development, Classical Test Theory (CTT) is one of the most prominent and frequently used measurement frameworks. Emerged in the 1940s, CTT focuses on test or form scores and has been dubbed "true score theory" for its assumption that each individual taking a test possesses a true score—scores obtained by examinees which depend on the difficulty level of the selected assessment tasks (Magno, 2009). This psychometric theory permits the prediction of testing outcomes, including the ability of the test takers and test item difficulty level; it manifests in the concept of a true score and an error through an observed score, indicating the reliability of a test (Alagumalai & Curtis, 2005). Nonetheless, CTT has been claimed to be sample dependent (Hambleton, 2000), implying that representative samples with an adequate number must be collected to perform the CTT analyses. In this instance, Hambleton and Jones (1993) suggest 200–500 as the sample size and argue that the addition of both item difficulty and discrimination into CTT, which are examined with test-score mean, standard deviation, and reliability, has provided sufficient desired statistical properties in the test development process.

The application of CTT in test development is generally to evaluate the test. It focuses on analyzing the total score, which involves frequency of correct responses to disclose question difficulty, frequency of responses that examine distracters, reliability of the test, and item-total correlation to identify discrimination at the item level (Impara & Plake, 1998). CTT provides the opportunities for standardization and calibration during test construction, which are two essential processes that indicate if the test material, the test administration circumstances, test sessions, and scoring methods are comparable to other standardized tests (standardization) and if the test instrument can place one person relative to others (calibration) (Alagumalai & Curtis, 2005; Waluyo & Bakoko, 2021). To achieve such objectives, CTT can be performed to explore the relationship between test length and test reliability, estimate difference scores and change scores, evaluate the properties of two or more measures, and assess the degree to which measurements can be affected by measurement errors (Stage, 2003).

In the development of the English proficiency test, CTT has been used to evaluate tests that measure students' overall and specific skill proficiency. As an example, Thirakunkovit (2016) evaluated the College English International Test (ACE-In) developed by Purdue University to measure international students' English proficiency in listening, reading, and writing. The items' reliability was found satisfactory, and they were identified as measuring the same underlying construct, i.e., English proficiency. In a specific skill, CTT has frequently been used to assess multiple-choice questions, such as on a reading test. Such an assessment was conducted by an early study (Perkins & Miller, 1984), followed by recent studies (e.g., Janssen et al., 2014; Zubairi & Kassim, 2016), which encouraged the use of CTT in the evaluation of high-stake tests. DeVellis (2006) argues that CTT is the standard comprehensive procedures in the process of developing, evaluating, and scaling test items, which are the prerequisite to a standardized test validation method. In this instance, to employ CTT in English proficiency test development, Suen (1990) suggests following the standard psychometric process, which begins with test item construction and continues with the examination of validity and reliability of the items through the exploration of true and observed scores. To scale the test items, paper exams and raters can be implemented, as illustrated in Figure 2.

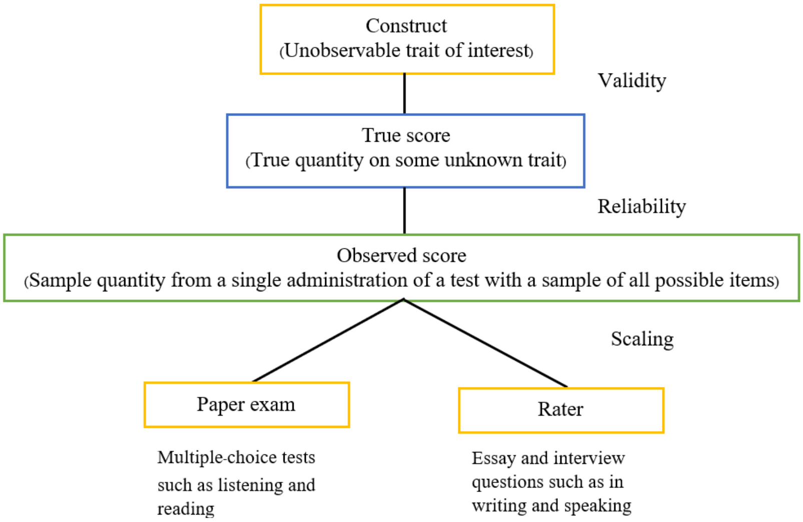
Construct
(Unobservable trait of interest)
Validity
True score
(True quantity on some unknown trait)
Reliability
Observed score
(Sample quantity from a single administration of a test with a sample of all possible items)
Scaling
Paper exam
Rater
Multiple-choice tests such as listening and reading
Essay and interview questions such as in writing and speaking

Figure 2 The psychometric process (Adapted from Suen, 1990)

Embracing CTT in test development is not only beneficial in the score examination but also valuable in the individual item analysis. DeMars (2018) contends that the item discrimination and difficulty indices derived from CTT can be used to determine whether items are useful or should be discarded and replaced. Item discrimination reveals the details of how well the items separate between examinees with high and low scores, while item difficulty refers to item easiness or item facility, reflected by the mean score on an item. In their latest study, Malec and Krzeminska-Adamek (2020) compared several methods of evaluating multiple-choice options in an English proficiency test, which included the use of several item analyses included in CTT. Their study confirmed that most of the evaluation methods gave similar results, signaling that employing one method may be adequate for multiple-choice item analysis. Due to its measures on both test score and item analysis, the present study applies CTT in the evaluation of multiple-choice questions in listening and reading tests. The exposition of the reliability concept embedded in CTT is meaningful as the basis for evaluating measuring instruments, such as for measuring students' listening and reading proficiency levels; CTT can provide sufficient information when the goal is to explore measurement error and test reliability (Wu et al., 2016).

METHOD

Research Objectives

The purpose of this study is to address the design, development, and validation of WUTEP as a CEFR-based test of English proficiency. The study pursues the following research objectives:

1. To evaluate the quality and functionality of items in the listening and reading tests of WUTEP through Classical Test Theory (CTT) analyses in testing 1 and 2.

2. To assess the effectiveness of the listening, reading, speaking, and writing tests of WUTEP in measuring students' English proficiency levels in testing 1 and 2.

3. To investigate the extent to which scores from WUTEP align with the predictive validity of TOEIC.

Research Design

Test framework

The purpose of WUTEP is to evaluate the English proficiency of those whose native language is not English. WUTEP scores are primarily used as a measure of the ability of both international and domestic students to use English in academic and work environments. WUTEP assesses students' levels of English proficiency in listening, reading, writing, and speaking, framed by the CEFR and CTT.

Test design

The design of WUTEP followed the standard procedures of the psychometric process in test development suggested by Suen (1990) as seen in Figure 2 and Irwing and Hughes (2018) as detailed in Figure 3. It began with construct definition, specification of test needs, and test structure. Afterward, test item construction was finished, and the test was piloted several times. The results were evaluated using CTT. After all the pilot studies had been conducted, the test was mapped onto an international standardized test, i.e., TOEIC (Test of English for International Communication) developed by an Educational Testing Service (ETS), who also created other international standardized tests such as TOEFL and GRE. Mapping onto TOEIC was the starting point before the WUTEP scores were mapped onto other international standardized tests. The selection of TOEIC was primarily caused by the fact that, compared to other international standardized tests, TOEIC has been widely accepted in Thailand, mainly for employment or job-related purposes.

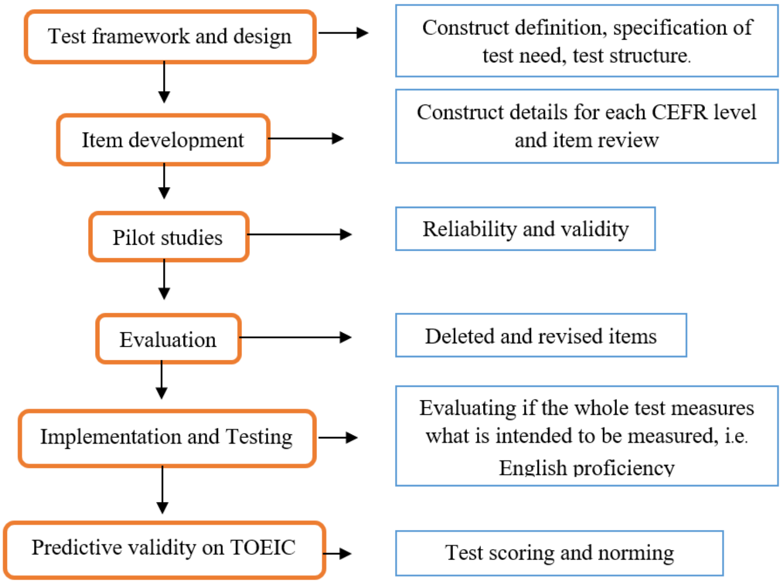
Test framework and design
Construct definition, specification of test need, test structure.
Item development
Construct details for each CEFR level and item review
Pilot studies
Reliability and validity
Evaluation
Deleted and revised items
Implementation and Testing
Evaluating if the whole test measures what is intended to be measured, i.e. English proficiency
Predictive validity on TOEIC
Test scoring and norming

Figure 3 The stages of WUTEP development following Irwing and Hughes (2018)

Item development: Test format and item constructions

Listening

The listening test measures test takers' ability to understand spoken English in academic and work environments. The listening section consisted of 50 multiple-choice questions. On the CEFR levels, the questions were divided into five levels with the following composition: A1 (20%), A2 (20%), B1 (20%), B2 (30%), and C1 (10%). The questions were distributed among four parts of the listening test, in which Part 1 comprised four statements and a picture for each question (5 questions), Part 2 contained a question or statement and three responses spoken in English (15 questions), Part 3 included conversations (15 questions), and Part 4 covered English talks (15 questions). The whole test would last about 40 minutes. Each question was constructed at each CEFR level, referring to guidelines from North et al. (2010). The question designers looked at the functions, topics, contexts, and CEFR level. During the question construction phase, the designers were given the details of the questions that they needed to create. For example, when constructing a listening question, the test designer would ensure that the listening question is within the assigned function, topic, and context (academic or work environment). Table 2 below depicts some of the details during the listening question construction. Key words in each constructed question were checked by using Vocab Kitchen for the CEFR level.

Table 2
An example of the item constructions for the listening test (following North et al. (2010))

No.	Functions	Topics	Context	CEFR Levels
1	Directions	Hobbies	Academic Environment	A1
2	Directions	Holidays	Work Environment	A1
3	Describing people	Education	Work Environment	A2
4	Describing people	Leisure Activities	Academic Environment	A2
5	Describing places	Film	Work Environment	B1
6	Describing things	Leisure Activities	Academic Environment	A2
7	Suggestion	Holidays	Work Environment	A2
8	Describing places	Shopping	Work Environment	A2
9	Checking understanding	News	Work Environment	B1
10	Expressing opinions, Using intensifier 'far too much'	Lifestyles	Work Environment	B1
11	Comparative, using 'will' for prediction	Media	Work Environment	B1
12	Describing Experience/ Expressing Opinions/ Lifestyle	Books	Work Environment	B2
13	Expressing Opinion/ Expressing reaction	Literature	Work Environment	B2
14	Expressing agreement and disagreement	Arts	Work Environment	B2
15	Emphasizing a point or feeling	Scientific Development	Work Environment	C1
16	Expressing opinions tentatively/ hedging	Scientific Development	Work Environment	C1
17	Responding to counterarguments	Film	Work Environment	C1

Reading

The reading test focuses on test takers' ability to understand university-level academic texts and practical work-related texts. On the CEFR levels, the composition of the questions was the same as in the listening test, ranging from A1 to C1. There were three parts, where Part 1 contained incomplete sentences (20 questions), Part 2 comprised an e-mail that missed a word or phrase (5 questions), and Part 3 included reading comprehension questions with both single and double passages (25 questions). The whole test lasted 50 minutes. The construction of the reading questions also followed the guidelines from North et al. (2010). The question designers considered the grammar, topics, vocabulary, function, context, and CEFR level of each question. For example, when constructing a reading text with five questions, the test designer would ensure that the reading topic is within the assigned topic and context (academic or work environment) and covers the assigned vocabulary and function. Table 3 shows some of the details from the question construction process.

Table 3
An example of the item constructions for the reading test (following North et al. (2010))

No.	Grammar	Topics	Vocabulary	Function	Context	CEFR Level
1	Adjectives	Family life	Personal information	Giving personal information	Academic Environment	A1
2	Preposition	Holidays	Personal information	Telling the time	Academic Environment	A1
3	WH-question in past	Work and job	Travel and services	Describing past experiences	Work Environment	A2
4	Modals - should	Education	Food and drink	Suggestions	Academic Environment	A2
5	1st conditional	Education	Feelings	Obligation and necessity	Academic Environment	A2
6	Reported speech	Media	Collocation	Describing experiences and events	Work Environment	B1
7	Modals – might have + verb 3	Books	Collocation	Expressing opinions	Academic Environment	B1
8	Passives	Arts	Collocation	Critiquing/ reviewing	Academic Environment	B2
9	Inversion with negative adverbials	Technical and legal language	Idiomatic expression	Depending a point of view persuasively	Academic Environment	C1
10	Mixed conditionals	Scientific development	Approximating (vague language)	Emphasizing a point	Academic Environment	C1

Writing

The writing test evaluates test takers' ability to write a short essay in academic and work environments. The essay requires test takers to draw on their own knowledge and experience to support their opinion on a specific issue. The test takers will choose one topic to respond to and write a short passage of at least 150 words to elaborate on their responses. For writing, it was the assessment rubric that was constructed based on the CEFR level, where the range covered A1 to C1 level. The assessment criteria encompassed task achievement, grammar, vocabulary, logic, and mechanics (spelling, punctuation, and capitalization). Below is the assessment rubric.

Table 4
The writing assessment rubric

No.	Criteria	Points			
		0.5	1	1.5	2
1.	**Task Achievement**	has serious disorganization or underdevelopment ,irrelevant; does not meet the appropriate length (< 50 words)	has limited developmentin response to the topic and task; Has limited length (≥ 50 - < 100 words)	addresses the topic andtask well, though some points may not be fullyelaborated; Meets the minimum length (≥100 - <150 words)	effectively addresses the topic and task; Meets the appropriate length (≥150 words)
2.	**Grammar**	contains serious and frequent grammatical errors	may demonstrate inconsistent facility in sentence formation that may result in lack of clarity	displays facility in the use of language, demonstrating syntactic variety, though it will probably have occasional noticeable minor errors in structure	Displays consistent facility in the use of language, demonstrating syntactic variety, though it may have minor grammatical errors
3.	**Vocabulary**	shows very poor knowledge of words, word forms, and is not understandable	shows a limited range of vocabulary and contains confusing words and word forms	shows few misuse of vocabularies and forms, but not change the meaning	shows effective choice of words and forms
4.	**Logics**	displays inadequate organization or connection of ideas	displays unity, progression, and coherence, though connection of ideas may be occasionally obscured	shows unity, progression, and coherence, though it may contain occasional redundancy, or unclear connections	displays unity, progression and coherence.
5.	**Mechanics (Spelling, Punctuation, and Capitalization)**	is dominated by errors of spelling, punctuation, and capitalization	has frequent errors of spelling, punctuation, and capitalization	has occasional errors of spelling, punctuation, and capitalization	uses correct spelling, punctuation, and capitalization

Speaking

The speaking section assesses test takers' ability to use English effectively in academic and work environments. The speaking test consisted of three parts: self-introduction (1 minute), speaking about two topics (6 minutes), and question-answer (3 minutes). Students will have a discussion with a lecturer. It will be interactive and as close to a real-life situation as a test can get. The speaking assessment rubric is adapted from the IELTS speaking rubric that has been mapped onto the CEFR levels, ranging from A1 to C2. The criteria included fluency and coherence, lexical resources, and pronunciation, with a detailed description for each criterion adapted from the IELTS speaking rubric, which has been used by previous studies in assessing English speaking proficiency (Dashti & Razmjoo, 2020).

Training for test assessors. The writing and speaking tests are distinct from the listening and reading assessments, as they necessitate the utilization of assessment rubrics and do not include multiple-choice questions. To ensure the reliability of assessment, training sessions were provided to the assessors, who were comprised of foreign English lecturers from a diverse range of countries, including the USA, UK, Australia, India, Philippines, Indonesia, Bhutan, Ghana, among others. To preserve consistency among the assessors, the training was repeated prior to each administration of the test. Additionally, samples of varying levels of quality in essay and speaking responses were made available to the assessors as reference materials.

Pilot Studies and Evaluations

Pilot 1

In the initial exploratory study, a sample of 80 undergraduate participants was selected. The evaluation methodology for the various assessments varied due to the differing nature of the questions posed. As described in the literature review, CTT was employed to assess the multiple-choice listening and reading tests, which comprised a total of 100 questions. The results, as presented in Table 5, demonstrated a high degree of internal consistency among the test items, as indicated by the Cronbach's alpha coefficient (α = .911). A more in-depth analysis revealed that 61 items in the listening test and 48 items in the reading test exhibited favorable item discrimination indices ranging from .20 to .80. In light of these findings, necessary modifications were made.

This study also undertook a distinct validation process for the writing and speaking tests, as they were not in a multiple-choice format. To validate the speaking test, a survey was conducted among 474 undergraduate students to gauge their level of familiarity with the 20 speaking topics provided. These topics were separated into two groups: the most familiar and the least familiar. The topics were sourced from the students' English course materials at the university. The speaking test was structured such that 10 topics were selected, ranked by familiarity, as depicted in Chart 1, and evaluated using the CEFR level-based speaking assessment rubric. In contrast, the writing test prompts were validated through Inter-rater reliability analysis conducted among a panel of foreign English lecturers from Iran, Indonesia, India, and Vietnam.

Table 5

Results from pilot 1

	Listening	Reading
Cronbach's Alpha	.911	.831
Item Discrimination Indices	P level = .20 - .80 61 Items have high discrimination indexes.	P level = .20 - .80 48 Items have high discrimination indexes.
Difficulty Index	P level = .40	P level = .36

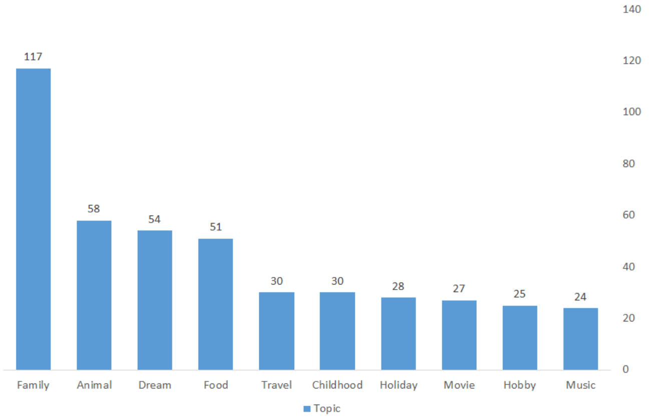

Chart 1 Results from the speaking topic survey

Pilot 2

The subsequent pilot study engaged both students and foreign English lecturers as participants. The evaluation protocols for the listening and reading tests remained consistent with those utilized in the initial pilot. As reflected in Table 6, notable improvements were observed from Pilot 1 to Pilot 2 in terms of item reliability (Listening: α = .911 to α = .96; Reading: α = .831 to α = .96), item discrimination (Listening: 61 to 67 items with high discrimination; Reading: 48 to 78 items with high discrimination), and item difficulty levels (Listening: .40 to .69; Reading: .36 to .61). In response to these results, necessary adjustments were made, and the number of questions on each test was reduced to 50.

Table 6

Results from pilot 2

	Listening	**Reading**
Cronbach's Alpha	0.963	0.959
Item Discrimination Indices	P level = .20 - .80 67 Items have high discrimination indexes.	P level = .20 - .80 78 Items have high discrimination indexes.
Difficulty Index	P level = .69	P level = .61

Data: Implementation and testing

This paper analyzed the data of [**Authors' University**] TEP scores from Testing 1 and 2. The details are elaborated below.

Testing 1

The test was administered in August 2018. It involved 2,248 first-year undergraduate students who were comprised of 1,664 females and 584 males. The students came from 13 different schools at Walailak University, including the School of Liberal Arts (331), School of Science (338), School of Pharmacy (139), School of Engineering and Technology (171), School of Informatics (72), School of Allied Health Science (368), School of Political Science and Law

(259), International Veterinary College (12), School of Architecture and Design (53), School of Nursing (169), School of Management (236), School of Public Health (257), School of Medicine (75), and School of Agricultural Technology and Food Industry (34). With regard to proficiency level, 1,053 students were at A1, 887 students were at A2, 265 students were at B1, and 43 students were at B2 levels. Their ages ranged from 18 to 20 years old.

Testing 2

The second test was conducted in May 2019. A different set of tests was used. The participants were a mix of first- and second-year students from Walailak University, with a total number of 3,655. For the first-year students, there were 2,054 of them, of which 1,203 were female and 383 were male. Meanwhile, the second-year students consisted of 1,527 females and 570 males, making up a total of 1,601 students. The students were from 14 different schools, such as the School of Liberal Arts (229), School of Science (24), School of Pharmacy (101), School of Engineering and Technology (139), School of Informatics (167), School of Allied Health Science (342), School of Political Science and Law (215), International Veterinary College (7), School of Architecture and Design (48), School of Nursing (165), School of Management (299), School of Public Health (241), School of Agricultural Technology and Food Industry (15), School of Medicine (47). For the first-year students, their proficiency levels consisted of A1 (375), A2 (1370), B1 (294), and B2 (15). Among the second-year students, there were 367 students at A1, 1,031 students at A2, 1,600 students at B1, and 2 students at B2.

RESULT AND DISCUSSION

Results

Item analysis

Classical Test Theory (CTT) was employed to examine the multiple-choice questions in the listening and reading tests of WUTEP through Testing 1 and 2. From Testing 1, it was obtained that the difficulty and discrimination indexes were .39 and .16, respectively, which indicated that items were difficult, but very discriminating (Haladyna & Rodriguez, 2013). The Item Facility (IF) calculation displayed that the quality of the sample was at 50% high and 50% low. Then, the reliability statistics of each question were explored. The results revealed a high level of internal consistency among the items (α = .83) with the point-biserial correlation at .23. The Kuder-Richardson reliability statistics revealed that the items were reliable for the test (*KR20* = .83) and most of the items shared the same level of difficulty (*KR21* = .81). Also, the coefficient from the split-half reliability was high at .83. Afterward, the scores for each question were analyzed. The test score mean was 38.64, in which there were test takers who obtained 86 (the highest) and 13 (the lowest). Out of 100 questions, 32 questions were considered ideal for the test, with 12 questions in the listening and 20 questions in the reading tests. Then, the rest of the items were subjected to revision and prepared for Testing 2.

In Testing 2, some improvements in the results were observed. The difficulty index was at the ideal level (*p* = .46), with a good discrimination index (*d* = .28). The internal consistency among the items remained at a high level (*α* = .85; consistently, the Kuder-Richardson reliability statistics kept the reliable level of the test (*KR20* = .85) and most of the items shared the consistent level of difficulty (*KR21* = .82). The test score mean was 44.03 with 15 and 89 as the minimum and maximum scores, correspondingly. The strength of the point-biserial correlation had increased to .31. Of 100 questions, 59 questions appeared to be representative for the test, with 27 questions in the listening and 32 questions in the reading test. The results from Testing 2 suggest that more than 50% of the multiple-choice questions in the listening and reading tests were difficult, yet very discriminating, which was considered desirable for a proficiency test designed to distinguish students' proficiency based on the CEFR levels. The detailed results are presented in Table 7.

Table 7
Results from Testing 1 and 2 (listening and reading)

Testing 1	Mean	Min.	Median	Max.	SD	Variance
Test scores	38.64	13	36	86	10.91	119.03
Difficulty index (p)	.39	.11	.37	.88	.15	.02
Delta	13.49	.00	14.15	18.00	3.54	12.56
Discrimination index (d)	.16	.01	.16	.42	.10	.01
Biserial (r)	.29	.09	.28	.63	.17	.03
Point-biserial RPB	.23	.08	.24	.46	.12	.02
Testing 2	**Mean**	**Min.**	**Median**	**Max.**	**SD**	**Variance**
Test scores	44.02	15	42	89	11.52	132.71
Difficulty index (p)	.46	.12	.47	.94	.18	.03
Delta	12.70	.00	13.15	17.90	3.55	12.63
Discrimination index (d)	.28	.14	.25	.63	.17	.03
Biserial (r)	.31	.17	.30	.82	.18	.02
Point-biserial RPB	.25	.15	.27	.54	.13	.02

Correlations among the subtests

In Testing 1, strong and positive correlations were observed between listening and reading scores (*r* = .64, *p* < .001), while the strengths of other correlations among the subtests of WUTEP were at a positive, moderate level. Each of the subtests' scores was strongly correlated with the WUTEP total scores. The strongest correlation was noted between writing and WUTEP total scores (*r* = .84, *p* < .001). Multiple-linier regression was performed to see how much variance of the WUTEP total scores could be explained by the subtests of WUTEP. The results revealed that with the listening, reading, writing, and speaking scores as predictors of WUTEP total scores, the model could significantly explain 99.8% (R^2 = .998) of the variance in the outcome variable (*F* (2243) = 251189.501, *p* < .001).

In testing 2, listening and reading scores were strongly correlated (*r* = .69, *p* < .001), while other subtests were moderately correlated. Among others, reading and WUTEP total scores had the strongest correlation (*r* = .83, *p* < .001), yet the other subtests' scores also reflected strong correlations with the total scores. In this second testing, the multiple-linier regression results showed that the model could explain 100% ($R^2 =$ 1.00) of the variance in the outcome variable (*F* (3650) = 5.678E+16, *p* < .001). The results from Testing 1 and 2, hence, confirmed that each of the subtests of WUTEP was associated closely and reliably measured what it was designed to measure, i.e., English proficiency level. This association also indicated that low performance in a particular subtest would affect the proficiency results, thereby distinguishing test takers' proficiency levels in both specific English skills and overall English proficiency. The following tables provide the detailed results.

Table 8

Correlations among the subtests in Testing 1

Testing 1	**Reading**	**Writing**	**Speaking**	**WUTEP Total Scores**
Listening	.642**	.462**	.467**	.757**
Reading		.505**	.517**	.810**
Writing			.483**	.835**
Speaking				.758**
Testing 2	**Reading**	**Writing**	**Speaking**	**WUTEP Total Scores**
Listening	.689**	.439**	.436**	.788**
Reading		.495**	.463**	.833**
Writing			.388**	.723**
Speaking				.783**

**. Correlation is significant at the 0.01 level (2-tailed).

Table 9

Regression results

Testing 1	**R²**	**F**	**B**	**Std. Error**	**Beta**	**t**
	.998	251189.501**	-.616	.045		-13.572**
Listening			.997	.006	.234	174.208**
Reading			1.015	.005	.287	204.057**
Writing			.976	.003	.444	362.190**
Speaking			.996	.004	.286	231.701**
Testing 2	**R²**	**F**	**B**	**Std. Error**	**Beta**	**t**
	1.00	5.678E+16	2.647E-13	.000		.000**
Listening			1.000	.000	.264	89063450.767**
Reading			1.000	.000	.313	101289547.599**
Writing			1.000	.000	.312	127239187.585**
Speaking			1.000	.000	.391	157302603.398**

Predictive validity of WUTEP on TOEIC

To examine the predictive validity of the test on TOEIC, this study selected 32 students to take TOEIC tests subsequently. The selection process commenced by assessing students' interest in participating in the TOEIC test and their availability to take it. Additionally, the students were chosen based on their intermediate-level English proficiency. We were unable to use a larger sample size due to the limited financial budgets to cover the participants' expenses. The students took the TOEIC tests on November 17, 2020. The analysis results exhibited that, performed using the IBM SPSS software, first, in predicting TOEIC listening scores by WUTEP listening scores, the model was significant ($F = 62.613$, $p < .000$); 67% of the variance in TOEIC listening results could be explained by WUTEP listening results ($R^2 = .67$) and both scores had a strong, positive relationship ($r = .822$, $p < .000$). For every unit increase in students' WUTEP listening results, a .59 unit increase in students' TOEIC listening results could be predicted ($B = .59$). Second, WUTEP reading scores could predict TOEIC reading scores by 53% ($R^2 = .53$) and the model was significant ($F = 35.762$, $p < .000$). For every unit increase in students' WUTEP reading results, a .72 unit increase in students' TOEIC reading results could be estimated ($B = .72$). Both variables were strongly related ($r = .737$, $p < .000$). Lastly, 69% of the variance in TOEIC total scores could be explained by WUTEP total scores ($R^2 = .69$). For very unit increases in WUTEP total scores, a .71 unit increase in TOEIC total scores could be projected ($B = .71$). A strong, positive relationship between WUTEP and TOEIC scores was noted ($r = .836$, $p < .000$). These results signify the predictive validity of WUTEP on TOEIC.

DISCUSSION

The main objective of this study was to present the stages involved in the design, development, and validation of a CEFR-based test of English proficiency, i.e., WUTEP.

First, the study evaluated the multiple-choice questions in the listening and reading tests of WUTEP in tests 1 and 2 by using CTT analysis. The results disclosed that the item difficulty and discrimination indexes were satisfactory with higher reliability coefficients after testing 2. More than 50% of the total questions were considered ideal for the test, while the rest were either deemed too easy or too difficult for the test takers. Both Testing 1 and 2 involved homogenous participants who were the main target test takers of WUTEP, so the results of the CTT analyses reflected the ability of the undergraduate students and the test item difficulty level for that particular group of participants (Magno, 2009). The results of various reliability statistics, e.g., the Kuder-Richardson reliability, Cronbach's alpha, and Split-half reliability, confirmed that the whole questions were reliable. There has not been a study specifically evaluating a CEFR-based test of English proficiency; yet evaluating multiple-choice tests by using CTT analyses has been done by previous studies (Janssen et al., 2014; Perkins & Miller, 1984; Thirakunkovit, 2016; Zubairi & Kassim, 2016), which validates the validation process conducted in this study. Nevertheless, the improvements that occurred from Testing 1 to Testing 2 might have been caused by the revisions and the increased number of participants. Thus, the results of the first research question also support the argument that in evaluating a high-stake test containing multiple-choice items, CTT can help determine whether items are useful or should be discarded and replaced (DeMars, 2018), but the results may depend on the sample size (Hambleton, 2000).

The following research question looked at the relationships between the WUTEP subtests and the overall score. This inquiry emphasized whether all the subtests were closely associated and measured the concept that they were designed to measure. Strong correlations among the listening, reading, writing, and speaking scores of WUTEP were obtained; all the subtests were also strongly related to WUTEP total scores. The regression results were reinforced by revealing that all the subtests explained 100% of the variance in test takers' WUTEP results. These results provide the criterion validity of the WUTEP from the correlation coefficient method. Furthermore, the quantitative analyses strengthen the content validity of WUTEP. During the item development stage, the listening and reading questions were developed based on the core inventory by North et al. (2010); the speaking topics and writing prompts involved internal raters, in which the assessment rubrics were adjusted following the CEFR level. Several studies (Rofiah & Waluyo, 2020; Waluyo, 2019) have used WUTEP scores to measure the level of English proficiency of Thai EFL learners. This shows that the assessment for English proficiency is accepted.

The last research question explored the predictive validity of WUTEP on TOEIC. The results confirmed that WUTEP total scores could explain about 70% of the variance in TOEIC listening and reading total scores. WUTEP listening and reading scores were observed to be closely correlated with TOEIC listening and reading scores. These results suggest that WUTEP scores can be comparable to TOEIC scores, and that both tests share identical features for measuring English proficiency. Most of the previous studies have attempted to map existing proficiency tests onto the CEFR level (Brunfaut & Harding, 2014; Khalifa & Ffrench, 2009; Papageorgiou et al., 2015; Tannenbaum & Wylie, 2008; Wudthayagorn, 2018), while the present study adds to the practice of predicting an existing international test by a CEFR-based test of English proficiency. Later on, it can lead to the practice of mapping a CEFR-based test score onto existing internationally recognized, standardized tests. Using predictive validity for validating an English proficiency test has also been conducted by empirical studies (Schoepp, 2018), making it one of the appropriate options when developing a CEFR-based test.

IMPLICATION

At the macro level, as interest in the adoption of the CEFR is growing across non-native English countries around the globe (Nagai, 2020), the stages of test development presented in this study can be a practical example for developing a CEFR-based test of English proficiency. Instead of mapping existing tests onto the CEFR, this study has proven that developing a comprehensive CEFR-based test that measures the four main English skills is feasible by combining the CEFR with CTT analysis. The CEFR can be applied as the foundational framework in item development for multiple-choice tests and in creating the criteria for assessment rubrics for speaking and writing. The CEFR core inventory from North et al. (2010) has been recognized as a reference point for language function, grammar, vocabulary, etc. for each CEFR level. Meanwhile, CTT analysis can be performed for test evaluation and validation. The stages of English proficiency test development highlighted in this study should provide an alternative solution to the insufficient guidelines of the CEFR informing language testing (Panmei & Waluyo, 2022; Weir, 2005).

Furthermore, in the past decade, there has been a gradually increasing trend of developing and implementing in-house/national/local standardized English proficiency tests. At a national level, some of the examples are the Canadian Academic English Language (CAEL) Assessment in Canada, the College English Test (CET) in the People's Republic of China, and the General English Proficiency Test (GEPT) in Taiwan (Cheng et al., 2014). There are also proficiency tests developed by universities; for example, i-TEPS by Seoul National University, Korea (Kim, 2018) and The University of Tehran English Proficiency Test (UTEPT), Iran (Rezaei & Shabani, 2010). At the micro level, in Thailand, high-ranking universities have established their own English proficiency tests, such as Chulalongkorn University Test of English Proficiency (CU-TEP), Prince of Songkhla University Test of English Proficiency (PSU-TEP), Thammasat University General English Test (TU-GET), and the latest one is the Srinakharinwirot University Standardized English Test (SWU-SET). It is assumed that other universities are still trying to find a way of developing their own English proficiency tests while still struggling to figure out the proper way of developing a standardized English proficiency test.

Templer (2004) argues that high-stake testing, such as English proficiency tests, often requires high fees and has created a worldwide industry involving educational commodification and marketisation on a global scale. It has been commonly known that English proficiency tests created by certain educational institutions, e.g., TOEFL by ETS and IELTS by the British Council, have been spread in non-native English countries, seemingly becoming a high-stake industrialized testing business. The high fees and centralization of the test administration are probably among the reasons for the growing trend in the development of in-house proficiency tests. Therefore, at this point, the findings of the present study should be incorporated into such an area of interest and encourage the development of CEFR-based tests both among countries and universities that are interested in having their own proficiency test.

CONCLUSION AND LIMITATION

This study has presented the development and validation of the WUTEP as a CEFR-based English proficiency test. The analysis showed that over 50% of the multiple-choice questions in the listening and reading tests met the "ideal" criteria, with good discrimination among test takers. The difficulty and discrimination indices from the last test yielded satisfactory results. To ensure that all subtests effectively measure English proficiency, correlations and regressions were conducted, confirming their alignment with the intended concept. Additionally, the study's predictive validity on the TOEIC indicates the WUTEP's ability to explain variance in TOEIC scores. It also explored the link between test results and CEFR proficiency levels, revealing specific item requirements for reaching each CEFR level, including A1, A2, B1, B2, and beyond. This comprehensive analysis highlights both the test's alignment with CEFR standards and certain contradictions within these relationships.

For limitations, this research solely interpreted quantitative results since test takers' qualitative data was not included in test development. The predictive validity has only been studied on TOEIC, suggesting that further tests on other international standardized English proficiency tests, such as TOEFL and IELTS, may be required to increase the criterion-related validity. Despite its large size, the sample was solely undergraduate EFL students at Walailak University, Thailand, suggesting that the test's validity may not be generalizable.

THE AUTHORS

Budi Waluyo completed his MA at the University of Manchester in the UK and his PhD at Lehigh University in the USA through grants from the IFP Ford Foundation, USA, and the Fulbright PhD Presidential Scholarship, USA. English Language Teaching, Educational Technology, and International Education are his areas of expertise as a lecturer and researcher. Dedicated to providing the best possible teaching and learning experiences, he is recognized as a Fellow of the UK's Higher Education Academy (FHEA).
budi.business.waluyo@gmail.com

Ali Zahabi, PhD is a lecturer at School of Liberal Arts, King Mongkut's University of Technology Thonburi (KMUTT), Thailand. He is a PhD graduate in Applied Linguistics from Universiti Sains Malaysia, and has more than 15 years of teaching experience at the academic level mostly in Southeast Asia (Malaysia & Thailand). His research interests include task-based language teaching, semiotics, and language assessment.
ali.zaha@kmutt.ac.th

Luksika Ruangsung obtained her Master's degree in teaching English as a Foreign Language from Thailand's Walailak University, with her research passion centered on the field of English Language Teaching.
luksika.ru@wu.ac.th

REFERENCES

Alagumalai, S., & Curtis, D. D. (2005). Classical test theory. In R. Maclean, R. Watanabe, R. Baker, Boediono, Y. C. Cheng, W. Duncan, J. Keeves, Z. Mansheng, C Power, J. S. Rajput, K. H. Thaman, S. Alagumalai, D. D. Curtis & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 1-14). Springer.

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

Borger, L. (2019). Assessing interactional skills in a paired speaking test: Raters' interpretation of the construct. *Apples: Journal of Applied Language Studies, 13*(1), 151-174.

Brunfaut, T., & Harding, L. (2014). Linking the GEPT listening test to the Common European Framework of Reference. *LTTC-GEPT Research Reports RG-05*, 1-75.

Cheewasukthaworn, K. (2022). Developing a standardized English Proficiency Test in alignment with the CEFR. *PASAA: Journal of Language Teaching and Learning in Thailand, 63*, 66-92.

Cheng, L., Klinger, D., Fox, J., Doe, C., Jin, Y., & Wu, J. (2014). Motivation and test anxiety in test performance across three testing contexts: The CAEL, CET, and GEPT. *TESOL Quarterly, 48*(2), 300-330.

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

Council of Europe. (2018). *The common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe.

Dashti, L., & Razmjoo, S. A. (2020). An examination of IELTS candidates' performances at different band scores of the speaking test: A quantitative and qualitative analysis. *Cogent Education, 7*(1), 1770936.

DeMars, C. E. (2018). Classical test theory and item response theory. *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. https://www.researchgate.net/publication/323222170_Classical_Test_Theory_and_Item_Response_Theory

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care, 44*(11), S50-S59.

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly, 15*(1), 3-15.

Dimova, S., Yan, X., & Ginther, A. (2022). Local tests, local contexts. *Language Testing, 39*(3), 341-354.

Dunlea, J., Fouts, T., Joyce, D., & Nakamura, K. (2019). EIKEN and TEAP: How two test systems in Japan have responded to different local needs in the same context. In L. I. W. Su, C. J. Weir & J. R. W. Wu (Eds.), *English language proficiency testing in Asia* (pp. 131-161). Routledge.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477-485.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38*, 60-65.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Instructional Topics in Educational Measurement, 12*(3), 38-47.

Harsch, C. (2018). How suitable is the CEFR for setting university entrance standards?. *Language Assessment Quarterly, 15*(1), 102-108.

Harsch, C., & Seyferth, S. (2020). Marrying achievement with proficiency–Developing and validating a local CEFR-based writing checklist. *Assessing Writing, 43*, 1-15.

Hiranburana, K., Subphadoongchone, P., Tangkiengsirisin, S., Phoochaeoensil, S., Gainey, J., Thogsngsri, J., ... & Taylor, P. (2017). A Framework of Reference for English Language Education in Thailand (FRELE-TH)-Based on the CEFR, the Thai experience. *LEARN Journal: Language Education and Acquisition Research Network, 10*(2), 90-119.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69-81.

Irwing, P., & Hughes, D. J. (2018). *Test development. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. https://onlinelibrary.wiley.com/doi/10.1002/9781118489772.ch1

Janssen, G., Meier, V., & Trace, J. (2014). Classical test theory and item response theory: Two understandings of one high-stakes performance exam. *Colombian Applied Linguistics Journal, 16*(2), 167-184.

Kanchai, T. (2019). Thai EFL university lecturers' viewpoints towards impacts of the CEFR on their English language curricula and teaching practice. *NIDA Journal of Language and Communication, 24*(35), 23-47.

Khalifa, H., & Ffrench, A. (2009). Aligning Cambridge ESOL examinations to the CEFR: Issues & practice. *Cambridge ESOL Research Notes, 37*, 10-14.

Kim, E. Y. J. (2018). Utility and bias in a Korean standardized test of English: The case of i-TEPS (Test of English Proficiency developed by Seoul National University). *Asian Englishes*, 1-14.

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

Little, D. (2006). The Common European Framework of Reference for languages: Content, purpose, origin, reception and impact. *Language Teaching, 39*, 167–190.

Liu, L., & Jia, G. (2017). Looking beyond scores: Validating a CEFR-based university speaking assessment in Mainland China. *Language Testing in Asia, 7*(1), 1-16.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Malec, W., & Krzeminska-Adamek, M. (2020). A practical comparison of selected methods of evaluating multiple-choice options through classical item analysis. *Practical Assessment, Research, and Evaluation, 25*(1), 7.

Moser, J. (2015). From a knowledge-based language curriculum to a competency-based one: The CEFR in action in Asia. *Asian EFL Journal, 88*, 1-29.

Nagai, N. (2020). *CEFR-informed learning, teaching and assessment: A practical guide*. Springer Nature.

Nagai, N., Birch, G. C., Bower, J. V., & Schmidt, M. G. (2020). *The CEFR and practical resources*. Springer.

Negishi, M., Takada, T., & Tono, Y. (2013, January). A progress report on the development of the CEFR-J. In E. D. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE Kraków Conference* (pp. 135-163). Cambridge University Press.

Nguyen, A. T. (2016). Towards an examiners training model for standardized oral assessment qualities in Vietnam. *Malaysian Journal of ELT Research, 11*(1), 41-51.

North, B. (2014). *The CEFR in practice* (Vol. 4). Cambridge University Press.

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

O'Sullivan, B., & Dunlea, J. (2015). *Aptis general technical manual version 1.0*. British Council.

Panmei, B., & Waluyo, B. (2022). The pedagogical use of gamification in English vocabulary training and learning in higher education. *Education Sciences, 13*(1), 1-22.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels*. https://www.ets.org/Media/Research/pdf/RM-15-06.pdf

Perkins, K., & Miller, L. D. (1984). Comparative analyses of English as a second language reading comprehension data: Classical test theory and latent trait measurement. *Language Testing, 1*(1), 21-32.

Piccardo, E. (2020). The Common European Framework of Reference (CEFR) in language education: Past, present, and future. *TIRF: Language Education in Review Series, 15*, 1-13.

Pratiwi, D. I., & Waluyo, B. (2022). Integrating task and game-based learning into an online TOEFL preparatory course during the COVID-19 outbreak at two Indonesian higher education institutions. *Malaysian Journal of Learning and Instruction (MJLI), 19*(2), 37-67.

Quynh, N. T. N. (2019). Vietnamese standardized test of English proficiency: A panorama. In L. I. W. Su, C. J. Weir & J. R. W. Wu (Eds.), *English language proficiency testing in Asia* (pp. 71-100). Routledge.

Rezaei, A., & Shabani, E. A. (2010). Gender differential item functioning analysis of the University of Tehran English Proficiency Test. *Pazhuhesh-e Zabanha-ye Khareji, 56*, 89-108.

Rofiah, N. L., & Waluyo, B. (2020). Using Socrative for vocabulary tests: Thai EFL learner acceptance and perceived risk of cheating. *The Journal of AsiaTEFL, 17*(3), 966-982.

Rofiah, N. L., Sha'ar, M. Y. M. A., & Waluyo, B. (2022). Digital divide and factors affecting English synchronous learning during Covid-19 in Thailand. *International Journal of Instruction, 15*(1), 633-652.

Schoepp, K. (2018). Predictive validity of the IELTS in an English as a medium of instruction environment. *Higher Education Quarterly, 72*(4), 271-285.

Stage, C. (2003). Classical Test Theory or Item Response Theory: The Swedish experience. www.cepchile.cl

Suen, H. K. (1990). *Principles of test theories*. Routledge.

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology. *ETS Research Report Series, 2008*(1), 1-75.

Templer, B. (2004). High-stakes testing at high fees: Notes and queries on the international English proficiency assessment market. *Journal for Critical Education Policy Studies, 2*(1), 1-8.

Thirakunkovit, S. (2016). *An evaluation of a post-entry test: An item analysis using Classical Test Theory (CTT)* [Doctoral dissertation, Purdue University]. https://docs.lib.purdue.edu/open_access_dissertations?utm_source=docs.lib.purdue.edu%2Fopen_access_dissertations%2F862&utm_medium=PDF&utm_campaign=PDFCoverPages

Waluyo, B. (2020). Thai EFL learners' WTC in English: Effects of ICT support, learning orientation, and cultural perception. *Humanities, Arts and Social Sciences Studies, 20*(2), 477-514.

Waluyo, B. (2019). Examining Thai first-year university students' English proficiency on CEFR levels*. The New English Teacher, 13*(2), 51-71.

Waluyo, B., & Apridayani, A. (2021). Teachers' beliefs and classroom practices on the use of video in English language teaching. *Studies in English Language and Education, 8*(2), 726-744.

Waluyo, B., & Bakoko, R. (2021). Vocabulary list learning supported by gamification: Classroom action research using Quizlet. *Journal of Asia TEFL, 18*(1), 289-299.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281-300.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Classical Test Theory. *Educational Measurement for Applied Researchers*, 73–90.

Wu, J. R., & Wu, R. Y. (2007). Using the CEFR in Taiwan: The perspective of a local examination board. *The Language Training and Testing Center Annual Report, 56*, 1-20.

Wu, R. Y. F. (2019). The general English Proficiency Test in Taiwan: Past, present, and future. In L. I. W. Su, C. J. Weir & J. R. W. Wu (Eds.), *English language proficiency testing in Asia* (pp. 9-41). Routledge.

Wudthayagorn, J. (2018). Mapping the CU-TEP to the Common European Framework of Reference (CEFR). *LEARN Journal: Language Education and Acquisition Research Network, 11*(2), 163-180.

Zubairi, A. M., & Kassim, N. L. A. (2016). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research, 2*(1), 1-20.