LEARN Journal: Language Education and Acquisition Research Network
ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 19, No: 1, January – June 2026

Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

Mapping STOU-EPT Scores with the CEFR Framework: A Standard Alignment Approach

Thanyasinee Laosum[a], Anusorn Koedsri[b,*]

[a] thanyasinee.lao@stou.ac.th, Office of Registration, Record and Evaluation, Sukhothai Thammathirat Open University, Thailand
[b] anusorn.koe@stou.ac.th, Office of Registration, Record and Evaluation, Sukhothai Thammathirat Open University, Thailand
[*] Corresponding author, anusorn.koe@stou.ac.th

APA Citation:
Laosum, T., & Koedsri, A. (2026). Mapping STOU-EPT scores with the CEFR framework: A standard alignment approach. *LEARN Journal: Language Education and Acquisition Research Network, 19*(1), 90-118. https://doi.org/10.70730/ROLY2367

Received
27/02/2025

Received in revised form
08/08/2025

Accepted
04/09/2025

ABSTRACT

This study examined the alignment of the Sukhothai Thammathirat Open University English Proficiency Test (STOU-EPT) with the Common European Framework of Reference for Languages (CEFR) using Webb's alignment method, and mapped STOU-EPT scores to CEFR levels with a modified Angoff method. Ten panelists evaluated alignment, ten panelists participated in standard setting, and 1,272 STOU-EPT test-takers were used. The STOU-EPT aligned with the CEFR as follows. Categorical Occurrence: Items across listening, structure, and reading were aligned to the A2–B2. Listening was more evenly distributed across levels, more structure items were included in A2–B1, and more reading items were included in B1–B2. All three sections met Webb's minimum requirement, with high rater agreement on listening and moderate agreement on structure and reading. Depth of Knowledge (DOK): Cognitive demand on listening and structure aligned with CEFR expectations at B1 and B2 but did not at A2. Reading showed appropriate cognitive demand across all targeted levels. Balance of Representation: The distribution of items across CEFR levels was balanced on listening, structure, and reading. CEFR-mapped scores were identified for each skill and overall performance, categorizing total scores as A2 (1–55 items), B1 (56–76), and B2 (77–100).

Keywords: CEFR, mapping, standard setting, STOU-EPT, Webb's alignment

Introduction

Thailand has implemented English proficiency policies to enhance its global competitiveness. Since the 2016 academic year, the Office of Higher Education Commission mandated the elevation of English language standards across all academic programs (Office of the Higher Education Commission, 2016). In response, higher education institutions, including Sukhothai Thammathirat Open University (STOU), have adopted strategies to improve instruction and ensuring students develop the necessary skills to access global knowledge. Graduate students must pass English proficiency tests meeting required standards. Undergraduates enrolling from 2017 onward must complete an English assessment or submit CEFR-aligned test scores. Those without valid scores may take an internally developed test administered by STOU.

The STOU-EPT, developed by STOU, evaluates English proficiency in listening, structure, and reading across three levels (A2, B1, and B2). Preliminary score ranges were set as follows: Listening – A2 (1–8), B1 (9–20), B2 (21–25); Structure – A2 (1–10), B1 (11–23), B2 (24–35); Reading – A2 (1–7), B1 (8–21), B2 (22–40). This approach provided a practical starting point; however, the absence of a standardized cut score method raises questions regarding the validity and defensibility of such classification thresholds.

Since 2017, the STOU-EPT has been used to assess English proficiency. However, concerns remain regarding its validity and CEFR alignment. Jiraro and Angsuchoti (2022) evaluated four test forms using expert judgment and found a high proportion of items exceeded their intended CEFR levels—44%, 61%, 59%, and 25%, respectively—while only 1–3% fell below. With only four experts involved, generalizability is limited, and the presence of items above or below targeted levels raises concerns about content validity. Psychometric analyses based on Classical Test Theory (CTT) showed that most items had acceptable difficulty levels (p = 0.20–0.80), but reliability coefficients (KR-20) varied considerably (0.35–0.89), with notably low values for Part I: Listening sections of Forms 3 and 4 (0.35 and 0.53, respectively). Comparisons with IELTS and SWU-SET revealed weak concordance at the A2 level, suggesting the STOU-EPT may assess skills at a higher level than intended, potentially overestimating test-takers' proficiency. This concern is amplified by the broad score range for B1, which increases the risk of misclassification near cut-off points. Notably, the original study did not establish cut scores using recognized standard-setting methods, leaving gaps in interpretive validity. As Kane (2013) observed, unvalidated thresholds cannot compromise fairness and accuracy in high-stakes decisions. These limitations highlight the need for further research to verify CEFR alignment and apply defensible standard-setting procedures to support valid score interpretations.

To address concerns about the STOU-EPT's interpretive validity, two complementary processes are required: alignment and mapping. In this study, alignment refers to examining the correspondence between test items and the CEFR descriptors to determine whether the test content appropriately targets the intended proficiency levels (Webb, 1997). Webb's alignment method was selected for its systematic and widely accepted approach to evaluate the match between assessment content and external standards. Originally developed for use in large-scale educational policy studies in the United States—such as aligning alternate assessments for students with special needs (Roach & Elliott, 2004; Roach et al., 2005) and evaluating statewide testing programs (Forte, 2017; Traynor et al., 2020)—the method provides clear criteria for assessing content focus. In the context of CEFR-linked language assessments, it offers a structured, transparent, and replicable process for documenting alignment quality—features essential for validating the STOU-EPT's content claims. Despite its international recognition, Webb's alignment method has not yet been applied in CEFR-related language assessments in Thailand.

Mapping, on the other hand, involves establishing empirical cut scores to classify test-takers into CEFR levels, thereby enabling valid and defensible score interpretations (Cizek & Bunch, 2007). These processes—alignment and mapping—are conceptually distinct but interdependent; meaningful score use requires both content alignment and appropriate proficiency thresholds. To this end, the present study adopts Webb's alignment methodology—commonly used in educational policy—and the modified Angoff method, an approach where experts judge evaluated items to establish proficiency thresholds, a widely accepted standard-setting approach in language testing (Plake & Cizek, 2012). The modified Angoff method has been applied in both international and national CEFR-aligned standard-setting studies, such as TOEIC®, secondary-level assessments in the U.S., and CU-TEP in Thailand (Baron & Papageorgiou, 2016; Tannenbaum & Baron, 2011; Wudthayagorn, 2018). While Papageorgiou et al. (2019) used psychometric linking to map TOEFL iBT® scores to China's

CSE framework, their focus likewise reinforced the interpretive validity of proficiency classifications. Collectively, these applications demonstrate the method's flexibility and policy relevance, justifying its use in this study to ensure content alignment and to establish defensible cut scores for the STOU-EPT.

The research objectives are therefore:

1. To align the STOU-EPT with the CEFR framework using Webb's alignment method.
2. To map the STOU-EPT scores to the CEFR levels using the modified Angoff method.

Literature Review

The CEFR as a Conceptual Framework for Language Test Validation

The CEFR provides a descriptive scale of language proficiency to support transparent communication of learner abilities across educational, professional, and national contexts (Council of Europe, 2020). Its six-levels (A1–C2), anchored in "can-do" descriptors, operationalize communicative competence across listening, speaking, reading, and writing. These descriptors define observable language behaviors, offering a framework for assessment (North, 2000). For example, a B1-level learner is expected to "understand the main points of clear standard input on familiar matters" and "deal with most situations likely to arise whilst travelling" (Council of Europe, 2020).

Although developed in Europe, the CEFR has been widely adopted worldwide, prompting efforts to align local assessments with its scale. Such alignment enhances the comparability, interpretability, and policy relevance of test scores (Green, 2020). In practice, CEFR alignment is typically conducted through expert judgment methods, which offer systematic evidence for construct relevance and level correspondence (Figueras, 2012). However, aligning test tasks with CEFR levels is not a straightforward matching exercise. It requires careful analysis of the linguistic, cognitive, and functional characteristics embedded in the tasks (Alderson, 2007).

Despite its widespread adoption, applying the CEFR in non-European contexts poses challenges such as curricular misalignment, cultural adaptation, and limited local expertise. Its conceptual abstraction and underspecification may also complicate alignment decisions (Alderson, 2005; Hulstijn, 2007). Nevertheless, when implemented with methodological rigor and contextual awareness, CEFR alignment can enhance transparency and inform high-stakes decisions such as university admission or graduation (Figueras, 2012).

Webb's Alignment Method in CEFR-Based Language Testing

Webb's alignment method was originally developed to evaluate the coherence between curriculum standards and assessments in K–12 education (Webb, 1997). It comprises five dimensions: content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability. Among these, content focus is the most widely applied in educational assessment research and consists of five operational criteria: (1) categorical concurrence—the match between test items and content standards; (2) Depth of Knowledge (DOK) consistency—the alignment of cognitive complexity, categorized into four levels: recall, skills and concepts, strategic thinking, and extended thinking; (3) range-of-knowledge correspondence—the breadth of content coverage; (4) balance of representation—equitable distribution of content; and (5) source of challenge—whether item difficulty arises from construct-relevant features.

The core principles of Webb's alignment method have been integrated—explicitly or implicitly—into recent CEFR alignment studies. For example, Tangsakul and Poonpon (2024) used the CEFR Content Analysis Grid to classify item coverage across CEFR levels, reflecting Webb's criteria of categorical concurrence and range-of-knowledge correspondence. Their findings show how structured content mapping supports test validation through systematic alignment with CEFR descriptors. This underscores the growing relevance of alignment frameworks in CEFR-based validation. However, Webb's framework remains underutilized in English language assessment in Thailand. This study addresses that gap by applying it to a high-stakes CEFR-referenced proficiency test, given its clarity in examining the breadth and cognitive demand of test content.

In this study, three of Webb's criteria—categorical concurrence, DOK, and balance of representation—were selected to evaluate the STOU-EPT's alignment with the CEFR, as they are most relevant to the study's focus on score interpretation and level-based inference. According to Kane's (2006) argument-based validity framework, evidential priorities must align with the intended interpretation. While the exclusion of the remaining criteria is theoretically justified, caution is warranted: CEFR-based test design alone does not ensure content breadth or fairness, and DOK levels may not fully capture the multidimensionality of CEFR descriptors (Alderson, 2007). Future research aiming to support broader validity arguments should consider incorporating all five criteria to enhance interpretive robustness.

Modified Angoff Method for Standard Setting

The Angoff method (Angoff, 1971) is widely used in judgment-based standard setting, especially in high-stakes testing. It requires panelists to estimate the probability that a minimally competent test-taker would answer each item correctly. The modified version (Livingston & Zieky, 1982) introduces multiple rounds, training, empirical feedback (e.g., item difficulty), and group discussions to improve consistency and reduce bias (Cizek & Bunch, 2007; Plake & Cizek, 2012). The use of the modified Angoff method for setting cut scores is particularly suitable for multiple choice tests and is widely adopted for aligning standardized test scores with reference frameworks such as the CEFR (Papageorgiou et al., 2019; Tannenbaum & Baron, 2011). In Thailand, Wudthayagorn (2018) used the method to map CU-TEP scores to CEFR levels, while Athiworakun and Wudthayagorn (2018) applied it to align the SWU-SET with levels A2, B1, and B2 across five skills.

As compared with the bookmark method, the contrasting groups method, the holistic method and the classical Angoff, the modified Angoff method's item-by-item judgment process offers a closer alignment with the performance-based philosophy of CEFR descriptors. However, this approach is not without its own challenges, as it places substantial cognitive demands on panelists and assumes a reliable, shared understanding of "minimal competence"—a construct that is inherently subjective and susceptible to cognitive biases (Hambleton & Pitoniak, 2006; Kane, 2013). The present study addresses these challenges proactively. To manage cognitive load and mitigate potential biases, the standard-setting process for the STOU-EPT will involve comprehensive panelist training on CEFR descriptors, the use of concrete performance exemplars, and multiple, discussion-based rating rounds with statistical feedback to foster a robust and shared understanding of proficiency levels.

Based on this analysis, the modified Angoff method was adopted for this study. Its emphasis content-based and item-level judgment provide a transparent and defensible process that directly supports the STOU-EPT's primary aim - to establish credible and meaningful links between test scores and the CEFR. This approach is consistent with best practices in other large-scale assessments, such as the TOEFL iBT, where methodological rigor and the interpretability of scores are paramount (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Papageorgiou et al., 2019).

Synthesizing Alignment and Standard Setting within a Validity Framework

To address both alignment and score interpretation within a unified validity framework, this study integrates three methodological perspectives—Webb's alignment method, the CEFR as a descriptive scale, and the modified Angoff method—under the conceptual guidance of the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014). These standards emphasize validity evidence to support test use and interpretation, particularly content relevance and decision consistency. Webb's method contributes content-related evidence through systematic analysis of alignment between test tasks and CEFR descriptors, while the modified Angoff method yields defensible cut scores aligned with defined proficiency levels. Together, these approaches offer complementary evidence for construct representation and classification accuracy, framing alignment and standard setting, as integral to a broader validation argument consistent with contemporary testing standards (Kane, 2006; American Educational Research Association et al., 2014).

The STOU-EPT: Objectives, Development, and CEFR-Based Design

The STOU-EPT is intended for (1) individuals seeking admission to undergraduate and graduate programs at STOU and (2) members of the general public who wish to assess their English proficiency and use the results as official evidence for academic admission, employment applications, or other relevant purposes.

The STOU-EPT was developed by a test committee at STOU through five stages. First, The STOU-EPT was developed by a test committee at STOU through five stages. First, committee members received training on CEFR-based test development from Dr. Cyril Weir and his colleagues at the Centre for Research in English Language Learning and Assessment, University of Bedfordshire, United Kingdom. The training focused on CEFR-aligned test design and the formulation of test specifications, which served as the foundation for item writing. Second, the test forms underwent expert review to ensure clarity, linguistic appropriateness, and level alignment. Language complexity and academic vocabulary were evaluated using Lextutor and the Academic Word List (AWL) (Jiraro & Angsuchoti, 2022). Third, the test was piloted under standardized conditions, and item statistics—such as difficulty and discrimination—were analyzed using CTT. Fourth, underperforming items were revised or discarded based on empirical results. Finally, validated items were stored in a secure item bank to maintain content balance and psychometric comparability in future administrations.

The test comprises 100 dichotomous items, multiple-choice items distributed across three skills: listening (25 items), structure (35 items), and reading (40 items). The listening section—administered in 30 minutes—includes four item formats: short-dialogue comprehension, information matching, summary completion, and true-false identification. The structure and reading sections use four-option multiple-choice items and are completed within a combined two-hour time limit. The test allocates the greatest weight to reading, given the university's focus on self-directed learning and graduate-level study, followed by structure and listening.

Method

Research Informants and Participants

The key informants for studying the alignment between the STOU-EPT and the CEFR using Webb's method were 10 panelists selected purposively. Although Ngudgratoke (2018) recommends three to five panelists for alignment, this study employed 10 to address the complexity of CEFR alignment and ensure reliable and diverse judgments (Cizek & Bunch, 2007). Their qualifications include: (1) a bachelor's, master's, and/or doctoral degree in English language teaching, linguistics, or intercultural communication, (2) experience in CEFR-based testing, including test development, rater training, assessment review, and related research, and (3) at least five years of English teaching experience in higher education, combined with involvement in CEFR-based assessment or curriculum development.

The key informants for standard-setting, responsible for determining cut scores and evaluating implementation, were also 10 panelists selected purposively. They have the same qualifications as those in the alignment study but were not involved in developing the STOU-EPT. Given the high-stakes standard-setting typically involves 15–20 panelists to ensure the consistency of the standard-setting outcome (Cizek & Bunch, 2007), the smaller panel size in this study necessitated additional validation. To substantiate this, a classification accuracy (CA) analysis was used to compare classifications derived from the panel's cut scores against proficiency levels established by an Item Response Theory (IRT) model, thereby providing an empirical check on the consistency of the final judgments.

Participants comprised 1,272 volunteer test-takers—including STOU students and external candidates—who voluntarily registered for the STOU-EPT test administered between May and September 2024. The sample size aligns with psychometric guidelines for IRT calibration, which typically recommend 500–1,000 participants (DeMars, 2010). To mitigate threats to validity in this high-stakes context, person-fit analyses were conducted to identify atypical response patterns. Given the likelihood of local dependence across sections (Listening, Structure, Reading), item quality was evaluated using the Rasch testlet model to account for such dependencies and to avoid inflated precision estimates. In this process, each panelist reviewed the items and applied professional judgment, using the Rasch difficulty estimates as supporting evidence in her decision-making.

Instruments

Set 1: An in-house set of test forms. The STOU-EPT was developed in 2023 by faculty members from the School of Liberal Arts at STOU, based on CEFR-aligned test specifications. Four parallel test forms were developed to ensure comparability. For this study, one form was randomly selected. As all forms shared identical specifications, underwent the same review process, and targeted the same CEFR levels (A2, B1, B2), the selected form was considered representative of the full set and suitable for analysis. It comprised 25 listening, 35 structure, and 40 reading items.

Set 2: An evaluation form for assessing the alignment between CEFR indicators and DOK levels. The researchers administered the form to five experts who possessed the same qualifications as those in the original alignment study. Based on mode values, listening at A2 and B1 levels measured Skills and Concepts, while B2 measured Strategic Thinking. Structure at all levels (A2–B2) measured Skills and Concepts. For reading, A2 and B1 measured Skills and Concepts, and B2 measured Strategic Thinking. Inter-rater reliability results indicated high agreement for listening (ϱ = 0.911) and reading (ϱ = 0.911), and moderate agreement for structure (ϱ = 0.540), based on Evans's (1996) guidelines. The moderate reliability for structure suggests the need to strengthen raters' shared understanding of both CEFR indicators and DOK levels through structured discussions or clarification sessions prior to alignment tasks, to ensure consistent and conceptually grounded judgments.

Set 3: An evaluation form for assessing the alignment categorical concurrence and DOK, in conjunction with CEFR indicators. Indicators for assessing listening and reading were based on EAQUALS (2008) descriptors, while structure indicators were derived from the British Council and EAQUALS (2015). These indicators were categorized into A2, B1, and B2 levels. Content validity was reviewed by three English language experts using the Item Objective Congruence (IOC) technique. The IOC values for all indicators ranged from 0.67 to 1.00, exceeding the 0.50 threshold set by

Rovinelli and Hambleton (1977), demonstrating their suitability for evaluating alignment in listening, structure, and reading proficiency. The revised indicators were evaluated by three other experts for alignment and inter-rater reliability of categorical concurrence using mean of bivariate Spearman's rho rank correlations between raters (ϱ). Listening and structure showed high reliability (ϱ = 0.644 and 0.655), while reading showed moderate reliability (ϱ = 0.479). For the inter-rater reliability of DOK, listening showed high reliability (ϱ = 0.882), whereas structure showed moderate reliability (ϱ = 0.554) and reading showed low reliability (ϱ = 0.200), aligning with Evans' (1996) criteria. The moderate and low reliability for structure and reading may reflect variability in interpreting CEFR descriptors. Future applications may benefit from incorporating calibration rounds and benchmark items to improve consistency in rating complex receptive skills.

Set 4: A guide for setting cut scores using the modified Angoff method, which provided panelists with a framework for determining cut scores.

Set 5: Modified Angoff rating forms, used as structured instruments by panelists during the Angoff standard-setting process. On these forms, panelists record their percentage estimates for each test item, specifically indicating the probability that a minimally competent candidate (MCC) would answer that item correctly.

Set 6: A questionnaire on the standard-setting process comprised two parts. Part one collected panelists demographics via a three-item checklist: highest qualification, years of tertiary-level English teaching, and experience with CEFR-aligned assessment or instruction. Part two contained 12 items rated on a four-point Likert scale, spanning four domains: (1) clarity of orientation and materials (four items); (2) ease of completing probability estimates and judgments (two items); (3) opportunities for discussion and feedback (three items); and (4) perceived usefulness of supplementary data and confidence in final cut scores (three items). The questionnaire was administered immediately after Round 3, and aimed at evaluating procedural quality in order to inform future refinements of the modified Angoff method. Content validity was confirmed by three experts using IOC, with values ranging from 0.67 to 1.00, exceeding the 0.50 threshold (Rovinelli & Hambleton, 1977).

Sets 1–3 supported CEFR alignment evaluation, whereas Sets 1 and 4–6 were employed in the standard-setting process.

Data Collection

The alignment study between the STOU-EPT and CEFR followed the four steps of Webb's method: (1) Orientation – panelists were briefed on

CEFR indicators, DOK levels, and the evaluation method, and were provided with relevant materials and guidelines; (2) Training – panelists practiced evaluating sample items to confirm their understanding of the process and to ensure consistent application of the criteria; (3) Discussion – panelists exchanged views, asked questions, and clarified interpretations to refine their shared understanding of the evaluation criteria and DOK levels, using descriptors, definitions, and sample items as reference points; and (4) Evaluation – each panelist independently reviewed all 100 items—25 listening, 35 structure, and 40 reading—to assess her correspondence with CEFR indicators and four DOK levels (recall, skills and concepts, strategic thinking, and extended thinking). The 14-day process included: Days 1–2 for familiarize herself with the evaluation materials, CEFR indicators, and DOK definitions; Days 3–10 to independently judge item-by-item across all three skills; Days 11–12 to review initial ratings, re-examine challenging items, and seek clarification if needed; and Days 13–14 to finalize and submit ratings. Panelists could request clarification or submit follow-up questions to the researchers at any time during the 14-day period via email or pre-arranged online consultation. All clarifications were recorded in written form and disseminated to all panelists to ensure consistent interpretation while preserving the independence of individual judgments.

The standard-setting process to determine cut scores was conducted as follows:

(1) Panelist Orientation – the researchers disseminated guidelines for determining cut scores, introduced the manual and forms, explained B1 and B2 test-taker characteristics on listening, structure, and reading. They defined borderline test-takers as those whose abilities lie on the threshold between two adjacent proficiency levels—A2 and B1, and B1 and B2—and described the expected knowledge, skills, and abilities of minimally competent candidates at each level. This ensured that all panelists shared a consistent understanding, which formed the basis for making consistent and defensible cut score judgments in subsequent rounds. Panelists engaged in structured calibration discussions and exemplar-based rating exercises to establish a shared interpretive framework prior to scoring.

(2) Cut Score Training:

Step 1: Panelists reviewed test content and item difficulty based on their teaching experience and concurrently considered the characteristics of borderline test-takers at the B1 level to inform their subsequent probability judgments.

Step 2: Panelists estimated the percentage (0–100%) of minimally competent B1 test-takers likely to answer each item correctly, recording their estimates on the modified Angoff rating forms provided.

For B2 cut score determination, the same procedure was followed as for B1, with panelists considering borderline B2 test-takers in their judgments.

Responses were analyzed for mean (M), standard deviation (SD), minimum (Min), maximum (Max), and standard error of judgment (SEJ) to assess consistency. Discussions helped refine understanding and achieved consensus. Processes (1) and (2) were completed within a one-day session.

(3) STOU-EPT Cut Score Determination – Two cut score types (specific skill and total scores) were established to ensure reliability and validity through a three-round process.

Round 1: Panelists determined cut scores for the three skills—listening, structure, and reading—at B1 and B2 levels. They first completed the B1 cut score determination before proceeding to B2 to maintain focus and consistency. They reviewed the content and difficulty of 100 test items based on their teaching experience and considered the characteristics of borderline B1 and B2 test takers, defined as minimally competent candidates for each level whose abilities lie at the threshold between two adjacent CEFR levels. Panelists then estimated the percentage (0–100%) of such candidates likely to answer each item correctly and recorded their estimates on the corresponding B1 and B2 rating forms.

The researchers compiled all responses and calculated the M, SD, Min, Max, and SEJ. Individual and group results were presented to panelists within the same round, allowing those with divergent ratings to provide their explanations and exchange perception with others. This feedback session aimed to clarify misunderstandings, refine shared interpretations, and ensure consensus before proceeding to Round 2 cut score determination. This round was completed in a one-day session.

Round 2: The Round 1 procedure was adopted, with additional information provided to support decision-making, including individual and group response data from Round 1 presented in the same round.

Panelists were allowed to adjust their responses for each item by increasing or decreasing the estimated percentage of minimally competent test-takers answering each item correctly, taking into account the feedback and clarifications discussed during the session. The revised responses were then analyzed using the same statistical measures as in Round 1. This round was completed in a one-day session.

Round 3: Panelists evaluated the appropriateness of the proposed B1 and B2 cut scores and discussed their plausibility and relevance for assessing English proficiency. They reviewed the aggregated results from Round 2 and engaged in collective deliberation, drawing on both statistical evidence and professional judgment, to ensure that the scores accurately reflected the required proficiency levels. Panelists were given the opportunity to revise their ratings during the same session, and all changes were immediately incorporated into the final dataset. The feedback session in this round served to finalize shared agreement on the cut scores before adoption. The final responses were analyzed using the same statistical measures as in Round 1, and the adjusted cut scores were adopted. This round was completed in a one-day session.

Data Analysis

The analysis comprised six sections. First, the categorical occurrence was analyzed through mean and standard deviation. The mean and proportion analyses were used with DOK levels. The mean of bivariate Spearman's rho rank correlations between raters (ϱ) was used for inter-rater reliability.

Second, the balance index (Q index) was used to analyze the representational balance. The Q index is an alignment criterion used to determine whether STOU-EPT items are evenly distributed across proficiency levels for each skill. A Q value near 1 indicates balanced representation, reflecting equal emphasis across skills, while a value near 0 suggests imbalance (Webb, 2007). The index is computed using this formula:

$$Balance\ Index = 1 - \left(\sum_{k=1}^{O} \left| \frac{I(k)}{O} - \frac{I(k)}{H} \right| \right) \div 2$$

Where O = Total number of objectives hit the set of CEFR descriptors for each skill

I (k) = Number of items hit corresponding to objective (k)

H = Total number of items hit the set of descriptors for each skill

Third, the quality of the STOU-EPT items was evaluated using the Rasch testlet model through four sequential steps: (1) unidimensionality was assessed via exploratory factor analysis; (2) local item independence was examined through the absolute values of Yen's Q3 statistics; (3) monotonicity was evaluated using ItemH coefficients to confirm that the probability of a correct response increases with ability; and (4) item difficulty parameters (b-parameters) and testlet variances were estimated.

Item difficulty parameters estimated using the Rasch testlet model ranged from −1.50 to 0.89 logits (M = -0.35, SD = 0.46), reflecting a modest spread centered near the average ability level. Easier items (e.g., Item 48:

−1.11; Item 50: −1.50) were answered correctly by most test-takers, whereas more difficult items (e.g., Item 3: 0.78; Item 5: 0.89) required higher proficiency. Standard errors of item difficulties were low (SE = 0.04–0.05), indicating stable parameter estimates. The potential scale reduction factor ($\hat{R}$) was approximately 1.00, confirming convergence of the estimation process. These difficulty estimates were not used to set cut scores directly but were provided to the Angoff panel in the second round as supplementary input. Panelists used this information to support or adjust their probability judgments for borderline B1 and B2 test-takers.

Forth, the cut scores for each skill were analyzed using descriptive statistics: M, SD, Min, Max, and SEJ.

Fifth, weighted overall cut scores were calculated by using the method proposed by Cizek and Bunch (2007). The weighted score is computed using this formula:

$$Weighted\ Score = \frac{(Weight_{Skill} \times Cut\ Score_{Skill})}{Max\ Score_{Skill}}$$

Where $Weight_{Skill}$ = the proportion of total items allocated to such a skill
$Cut\ Score_{Skill}$ = the modified Angoff cut score for such a skill
$Max\ Score_{Skill}$ = the total number of items for such a skill

$$Total\ Weighted\ Cut\ Score = \sum_{i=1}^{n} \quad Weighted\ Score_i$$

Where n = the number of skills

Finally, the cut score-setting process was evaluated based on the panelists' feedback, with descriptive statistics.

Results

CEFR Alignment Results Based on Webb's Method

Categorical Occurrence Alignment Results

Listening section aligned with the CEFR, with eight items of A2 (M = 8.00), eight of B1 (M = 8.10), and nine of B2 (M = 8.90).

Structure section aligned with the CEFR, with 12 items of A2 (M = 12.20), 12 of B1 (M = 12.40), and 10 of B2 (M = 10.40).

Reading section aligned with the CEFR, with six items of A2 (M = 5.70), 19 of B1 (M = 19.20), and 15 of B2 (M = 15.10).

All sections in categorical occurrence aligned with at least six items per standard of Webb's criteria. An inter-rater reliability analysis of the ten raters indicated a high level of agreement on listening (ϱ = 0.902) and a moderate level on structure (ϱ = 0.588) and reading (ϱ = 0.432). The results also aligned with those specified by Evans (1996), as illustrated in Tables 1 and 2.

DOK Alignment Results

Listening section aligned with the DOK, with test items of B1 (% hit = 100) and B2 (% hit = 100) meeting the criterion, while A2 did not (% hit = 0) (below 50%).

Structure section aligned with the DOK, with test items of B1 (% hit = 100) and B2 (% hit = 92.31) meeting the criterion, while A2 did not (% hit = 0) (below 50%).

Reading section aligned with the DOK, with test items of A2 (% hit = 100), B1 (% hit = 100), and B2 (% above = 100) meeting the criterion.

All sections in DOK met Webb's criteria (1997), which posited that an assessment instrument possessed sufficient DOK when at least 50% of its test items matched or exceeded the depth specified in the corresponding indicators. An inter-rater reliability analysis of the ten raters indicated a moderate level of agreement on listening (ϱ = 0.525) and structure (ϱ = 0.442) and a low level on reading (ϱ = 0.249). These results aligned with those specified by Evans (1996), as illustrated in Tables 1 and 2.

Table 1

Categorical Occurrence and DOK Alignment Assessment

Skills	**CEFR Level**	**Categorical Occurrence Alignment**			**DOK Alignment**			
		Mean CEFR-Aligned Items (N = 10)	**SD**	**Results**	**% Below**	**% Hit**	**% Above**	**Results**
Listening	A2	8.00	3.56	Pass	100	0	0	Fail
	B1	8.10	2.85	Pass	0	100	0	Pass
	B2	8.90	2.51	Pass	0	100	0	Pass
Structure	A2	12.20	3.46	Pass	100	0	0	Fail
	B1	12.40	5.48	Pass	0	100	0	Pass

	B2	10.40	3.78	Pass	7.69	92.31	0	Pass
Reading	A2	5.70	6.70	Pass	0	100	0	Pass
	B1	19.20	7.00	Pass	0	100	0	Pass
	B2	15.10	8.84	Pass	0	0	100	Pass

Table 2

Inter-Rater Reliability for Categorical Occurrence and DOK Alignment

Skills	Alignment	Number of Test Items	Number of Raters	ϱ
Listening	Categorical Occurrence	25	10	0.902
	DOK	25	10	0.525
Structure	Categorical Occurrence	35	10	0.588
	DOK	35	10	0.442
Reading	Categorical Occurrence	40	10	0.432
	DOK	40	10	0.249

The balance of representation showed Q index values of 0.853 on listening, 0.952 on structure, and 0.817 on reading, indicating well-distributed STOU-EPT test items across proficiency levels. This aligns with Webb (2007), who considers Q ≥ 0.70 acceptable, as illustrated in Table 3.

Table 3

Balance of Representation Assessment

Skills	Number of Test Items	Proficiency Levels	Q
Listening	25	3	0.853
Structure	35	3	0.952
Reading	40	3	0.817

Mapping STOU-EPT Scores onto CEFR Levels

Mapping of STOU-EPT scores with CEFR levels using the modified Angoff method was presented in two formats:

Specific Skill Score

Listening: The recommended cut score for B1 in Round 3 was 13.49 (≈13), showing a decline from Round 1 (15.75) and Round 2 (15.91). The recommended cut score for B2 in Round 3 was 19.24 (≈19), remaining consistent with Round 1 (19.15) and Round 2 (19.30). The cut score difference between B2 and B1 was six points. The SD and SEJ for both B1 and B2 in Round 3 declined from Rounds 1 and 2.

The recommended cut score for B1 in Round 3 was 20.48 (≈20), showing no significant change from Round 1 (19.86) but a slight decline from Round 2 (22.12). For B2, the recommended cut score in Round 3 was 27.23 (≈27), consistent with Round 2 (27.30) but slightly higher than Round 1 (25.57). The cut score difference between B2 and B1 was seven points. The SD and SEJ for both levels in Round 3 declined from Rounds 1 and 2.

Reading: The recommended cut score for B1 in Round 3 was 23.26 (≈23), declining from Round 1 (24.53) and Round 2 (26.27). The B2 cut score in Round 3 was 31.43 (≈31), slightly higher than Round 1 (30.28) but lower than Round 2 (32.02). The cut score difference between B2 and B1 was eight points. The SD and SEJ for both levels in Round 3 were lower than those in Rounds 1 and 2, as illustrated in Table 4.

Table 4

Standard-Setting

Levels	**B1**			**B2**		
	Round 1	**Round 2**	**Round 3**	**Round 1**	**Round 2**	**Round 3**
Listening						
M	63.00	63.62	53.96	76.61	77.22	76.95
SD	14.88	5.36	3.36	16.15	6.86	6.42
Min	39.20	50.40	48.80	44.40	60.40	60.40
Max	84.40	69.20	60.00	95.40	88.16	85.68
SEJ	4.70	1.69	1.06	5.11	2.17	2.03
Cut scores	15.75	15.91	13.49	19.15	19.30	19.24
Structure						
M	56.74	63.19	58.52	73.05	78.00	77.79
SD	19.84	4.63	3.22	21.46	5.70	4.40
Min	8.00	58.29	54.14	15.60	68.71	69.86
Max	81.20	73.57	64.57	95.00	89.29	85.63
SEJ	6.28	1.46	1.02	6.79	1.80	1.39
Cut scores	19.86	22.12	20.48	25.57	27.30	27.23
Reading						
M	61.32	65.68	58.16	75.70	80.05	78.57
SD	18.86	7.72	5.58	19.25	7.32	4.90
Min	21.60	54.38	49.88	31.20	64.38	66.63
Max	85.80	79.50	65.38	95.20	91.38	84.65
SEJ	5.96	2.44	1.77	6.09	2.32	1.55
Cut scores	24.53	26.27	23.26	30.28	32.02	31.43

Remarks:

M = Mean of panelists' recorded percentage estimates

SD = Standard deviation of panelists' recorded percentage estimates

Cut score = Sum of panelists' mean probability estimates across all items

Total Score

The cut scores were set at 56 for B1 and 77 for B2.

Using both the specific skill and total score approaches, STOU-EPT scores were mapped onto CEFR levels and categorized into score ranges for A2, B1, and B2, as illustrated in Table 5.

Table 5

Mapping of STOU-EPT Scores with the CEFR Framework

Levels	Specific Skill Score			Total Score
	Listening	Structure	Reading	
A2	1-12	1-19	1-22	1-55
B1	13-18	20-26	23-30	56-76
B2	19-25	27-35	31-40	77-100
CA	0.972 (97.20%)	0.956 (95.60%)	0.937 (93.70%)	-

This section details the analysis of classification accuracy (CA) for the STOU-EPT, evaluating how well the established cut scores differentiate test-taker proficiency levels.

Given the relatively small panel size in the Modified Angoff standard-setting process, a classification analysis was conducted to examine the validity of the result cut scores for the STOU-EPT. The analysis compared classifications based on test-takers' observed raw scores with those based on their IRT-derived expected scores, applying the same panelist-established cut-offs for both. Agreement rates were high for Listening (97.20%), Structure (95.60%), and Reading (93.70%). These findings offer relevant validity evidence supporting the consistency of the panelists' judgments and the defensibility of decisions based on these scores.

Evaluation of the Cut score-setting Process

The evaluation results of the cut score determination process in which the panelists expressed their opinions at the highest level (≥ 60%) were these five aspects (1) the probability estimation forms for correctly answering test items in Rounds 1, 2, and 3 were easy to complete (60%, M = 3.60, SD = 0.52); (2) the opportunity to provide feedback on the results of cut score determination for each round (80%, M = 3.80, SD = 0.42); (3) sufficient time was allocated to provide feedback on the results of cut score determination in each round (70%, M = 3.70, SD = 0.48); (4) the post-Round 1 meeting for cut-score determination was beneficial to the process (80**%**, M = 3.80, SD = 0.42); and (5) the post-Round 2 meeting for cut-score determination was beneficial to the process (70**%**, M = 3.70, SD = 0.48). There were three other aspects in which panelists expressed their opinions at a high level (≥ 60%) including the (1) facilitator provided a clear explanation of the methods used for determining cut scores (60%, M = 3.20, SD = 0.63); (2) the orientation and practice sessions improved understanding of the cut score determination process (60%, M = 3.50, SD = 0.71); and (3) the manual for cut-score determination using the modified Angoff method was clear and easy to understand (60%, M = 3.40, SD = 0.84), as illustrated in Table 6.

Table 6

Evaluation of the Cut Score-Setting Process

	Questions	**Evaluation Results**				**M**	**SD**	**Opinion Levels**
		1 (%)	2 (%)	3 (%)	4 (%)			
1	The orientation session clarified the objectives of the standard-setting process for determining cut scores.		1 (10)	5 (50)	4 (40)	3.30	0.67	high
2	The facilitator provided a clear explanation of the methods used for determining cut scores.		1 (10)	6 (60)	3 (30)	3.20	0.63	high
3	The orientation and practice sessions enhanced understanding of the cut score determination process.		1 (10)	3 (30)	6 (60)	3.50	0.71	high
4	The manual for the cut-score determination using the modified Angoff method was clear and easy to understand.		2 (20)	2 (20)	6 (60)	3.40	0.84	high
5	The probability estimation forms of correct answer test items in Rounds 1, 2, and 3 were easy to complete.			4 (40)	6 (60)	3.60	0.52	highest
6	The process of determining cut scores was simple and straightforward.	2 (20)	2 (20)	3 (30)	3 (30)	2.70	1.16	high
7	I had the opportunity to provide feedback on cut score determination results in each round.			2 (20)	8 (80)	3.80	0.42	highest
8	Sufficient time was allocated for me to provide feedback			3 (30)	7 (70)	3.70	0.48	highest

	on the results of cut score determination in each round.							
9	The post-Round 1 meeting for cut-score determination was beneficial to the process.			2 (20)	8 (80)	3.80	0.42	highest
10	The post-Round 2 meeting for cut-score determination was beneficial to the process.			3 (30)	7 (70)	3.70	0.48	highest
11	The information provided in Round 2, including Round 1 cut scores and item difficulty analysis using the Rasch testlet model was helpful for cut score determination.			5 (50)	5 (50)	3.50	0.53	high
12	I am confident in the cut scores determination using the modified Angoff method and can justify their appropriateness.		1 (10)	4 (40)	5 (50)	3.40	0.70	high
Total		2	8	42	68	3.47	0.40	high

Discussion

The Alignment Study Between STOU-EPT and CEFR

Interpreting CEFR Alignment Across Content, Cognitive, and Representational Dimensions

This study employed Webb's (1997, 2007) alignment framework to evaluate the extent to which the STOU-EPT reflects the CEFR's intended construct coverage. The three selected criteria—categorical concurrence, DOK consistency, and balance of representation—jointly represent content fidelity, cognitive demand, and structural fairness. These dimensions align with the Standards for Educational and Psychological Testing (American Educational Research Association et al, 2014) and Kane's (2006) argument-based validity framework.

The results indicated that the STOU-EPT achieved satisfactory categorical concurrence across all three skill domains. In accordance with Webb's (1997) guidelines, each CEFR level was represented by at least six items per skill. The listening section demonstrated high inter-rater reliability (ϱ = 0.902), suggesting consistent application of CEFR descriptors in auditory contexts. Structure also met the coverage criterion, but yielded moderate inter-rater reliability (ϱ = 0.588), implying variability in raters' interpretations of grammar complexity and CEFR level differentiation. Conversely, reading—though meeting coverage requirements—yielded lower reliability coefficients (ϱ = 0.432), possibly due to the broader and less operationalized nature of CEFR reading descriptors in academic settings (Al Lawati, 2023; North, 2000). This pattern aligns with Alderson's (2007) critique of CEFR descriptors, particularly their underspecification for receptive skills.

Interpreting DOK Alignment and A2-Level Challenges

The DOK alignment analysis revealed strong cognitive consistency at the B1 and B2 levels, with over 90% of items meeting the expected complexity criteria. This finding reinforces the STOU-EPT's validity in measuring higher levels of proficiency. However, A2-level items in both listening and structure failed to meet DOK expectations, raising concerns about their ability to engage the fundamental cognitive processes described in CEFR descriptors. Similar challenges have been reported in other CEFR alignment studies involving lower levels, where A2 tasks frequently fail to represent minimally proficient cognitive engagement (Tangsakul & Poonpon, 2024).

Moreover, the inter-rater reliability for DOK judgments remained moderate to low (e.g., ϱ = 0.442 for structure; ϱ = 0.249 for reading), indicating difficulties among panelists in interpreting cognitive complexity. This finding echoes Alderson's (2007) argument that CEFR descriptors lack precision in defining cognitive processes such as "strategic thinking" or "skills and concepts." Hulstijn (2007) similarly cautions that alignment decisions based solely on expert intuition are susceptible to inconsistency without empirical or training-based scaffolding.

Synthesizing Alignment Dimensions and Implications for Interpretive Validity

Synthesizing across all three alignment dimensions provides a nuanced picture of the STOU-EPT's interpretive validity. The B1 and B2 levels demonstrated coherent alignment across content, cognitive, and representational criteria, supporting their appropriateness for internal high-stakes decisions within the university context—particularly for graduation and English proficiency certification purposes, in alignment with national policy directions (Office of the Higher Education Commission, 2016). The convergence of categorical coverage, DOK consistency, and strong Q indices at these levels reflects the test's strengths in construct representation.

Nevertheless, the persistent misalignment at the A2 level—particularly in listening and structure—undermines the interpretive defensibility of scores for minimally proficient test-takers. Despite achieving numerical balance (Q = 0.853 for listening; 0.952 for structure), these sections failed to meet DOK expectations and showed poor inter-rater agreement. This convergence of weaknesses suggests that the A2-level items may inadequately represent the target construct, thereby threatening the fairness and accuracy of CEFR-based classifications (Kane, 2013; American Educational Research Association et al., 2014). Misclassification risks are particularly concerning near cut-off points, where ambiguous item characteristics could lead to overestimation or underestimation of learner ability (Green, 2020).

The consistently high Q index values across all sections—including 0.817 for reading—confirm the structural balance of CEFR-level representation, exceeding Webb's (2007) threshold of 0.70. However, as Alderson (2007) reminds us, structural representation alone does not guarantee meaningful score interpretation. Construct validity requires that test items not only cover the appropriate levels but also reflect the intended cognitive and functional characteristics embedded in CEFR descriptors.

These findings underscore that alignment is inherently multidimensional and interdependent. A deficiency in any one criterion—such as DOK consistency at A2—can compromise the interpretive integrity of scores. In this regard, the study affirms Kane's (2006, 2013) argument that test validation must be grounded in evidence synthesized across multiple dimensions of alignment.

To enhance the STOU-EPT's alignment quality and interpretive validity, these recommendations are proposed. First, A2-level items should be revised to ensure they engage appropriate cognitive processes and reflect the CEFR's conceptualization of minimal proficiency. Second, rater calibration should be strengthened through exemplar-based training, particularly in the interpretation of DOK levels. Third, item development guidelines should explicitly operationalize CEFR descriptors to support task design and rating consistency, particularly at the lower levels of proficiency. These steps will enhance the STOU-EPT's capacity to function as a defensible and equitable CEFR-referenced proficiency assessment in Thai higher education.

Mapping of STOU-EPT Scores with CEFR

Skill-Specific Cut Scores and Interpretive Trends

The iterative standard-setting procedure resulted in stable and interpretable cut scores across the three skill areas—listening, structure, and reading. Notably, the final cut scores for B1 and B2 levels (e.g., Listening: 13 and 19; Structure: 20 and 27; Reading: 23 and 31) demonstrated clear intervals between levels, which aligns with the CEFR's construct of progressive language proficiency. Decreases in standard deviations and standard errors of judgment across rounds indicated increasing panelist consensus, a trend that echoes findings from prior studies on the reliability benefits of iterative judgment rounds (Cizek & Bunch, 2007; Papageorgiou et al., 2019).

However, interpretive discrepancies were observed, particularly in the reading section. The consistently higher SDs for reading may be attributed to the broader and more abstract nature of CEFR descriptors in academic reading contexts (North, 2000), thus posing challenges to panelists when conceptualizing the minimally competent person (MCP). This echoes Alderson's (2007) concern that CEFR descriptors, especially at lower and higher levels, often lack sufficient specificity for receptive skills. The implications of these findings suggest that while cut scores were statistically robust, the construct interpretations represent require ongoing theoretical scrutiny.

Total Score Mapping and its Consequences

The total score thresholds for B1 (56) and B2 (77) were used to classify candidates into CEFR bands—A2 (1–55), B1 (56–76), and B2 (77–100). This compensatory model, which allows strengths in one skill to offset weaknesses in another, is both pedagogically practical and administratively efficient for institutional decisions, such as graduation eligibility or language support placement. It aligns with STOU's flexible learning model and accommodates diverse learner profiles.

Still, this approach raises key validity and policy concerns. While total scores offer interpretive simplicity, they may obscure domain-specific weaknesses, contradicting the CEFR's emphasis on distinct communicative competencies (Council of Europe, 2020). Misinterpretation of the total score as uniformly representative of all skills can lead to inappropriate educational or employment decisions and weakens diagnostic precision.

Furthermore, although supported by prior findings, the STOU-EPT's total score use is currently confined to internal decisions. Broader applications—such as employment screening or integration into national frameworks—require external validation and stakeholder consultation. Without such safeguards, there is a risk of misalignment with evolving CEFR-aligned language policy in Thailand (Office of the Higher Education Commission, 2016).

Additionally, the compensatory design may unfairly advantage candidates with strong performance in one domain while masking deficiencies in others. This could disadvantage test-takers with more balanced but moderate skills, especially in contexts where specific language abilities are crucial (e.g., interpreting, academic reading).

To strengthen fairness and interpretive utility, institutions should consider reporting both total and domain-level scores, accompanied by clear interpretive guidelines. Such a dual-reporting model better supports CEFR-aligned instruction and placement, while also upholding Kane's (2006) principle of contextualized validity arguments that consider test use, interpretation, and consequences.

Reflections on the Standard-Setting Process

The panelist evaluation data indicated generally positive perceptions of the modified Angoff procedure. Panelists valued the opportunity for reflection across multiple rounds, the clarity of facilitation, and the structured use of supplementary information, including Rasch-based item difficulty estimates. These findings reinforce previous findings suggesting that standard-setting exercises benefit from transparent procedures, data-informed feedback, and opportunities for expert calibration (Papageorgiou et al., 2019).

Critical reflection also reveals areas for improvement. First, the reliance on social consensus across rounds may lead to regression to the mean, where shifts in judgments stem more from conformity pressures than deeper conceptual clarity (Cizek & Bunch, 2007). Second, while item difficulty indices were provided, the extent to which panelists analytically integrated these data remains unclear. These observations suggest the need for future research to explore the metacognitive processes underlying expert judgment and to develop scaffolds that support evidence-based interpretation.

Taken together, the findings suggest that the modified Angoff method can be feasibly implemented in institutional contexts such as STOU for CEFR-referenced cut score setting. However, given the limitations in scope, generalizability, and certain construct representation issues, further research and refinement are essential to strengthen both the interpretive and consequential validity of the STOU-EPT. Specifically, the following areas warrant attention: (1) item refinement at lower levels: Several A2-level items—particularly in listening and structure—may not fully reflect the cognitive and functional descriptors outlined by the CEFR. Development of these items should align more closely with CEFR's "can-do" statements and involve validation through think-aloud or response process data, (2) panelist calibration: Variation in borderline judgments across rounds, while decreasing, indicates the need for enhanced training using exemplars or video-based cases to ground panelists' interpretations and (3) external validation: The current study relied exclusively on internal judgments. Future studies should examine the predictive validity of STOU-EPT scores against independent outcomes (e.g., academic achievement, employment success) to establish consequential validity.

Finally, the present study provides foundational evidence for the application of standard-setting procedures in aligning an institutional English proficiency test with the CEFR. The procedures employed—particularly the use of the modified Angoff method—yielded coherent cut scores and received positive expert feedback. However, the findings also reveal limitations with potential policy and educational implications. The use of total scores should be carefully qualified, and future developments should prioritize external validation, the refinement of low-level items, and the adoption of more diagnostic score reporting models. These steps are essential not only to strengthening the interpretive and consequential validity of the STOU-EPT but also ensuring the test supports equitable and transparent decision-making in alignment with national language education policies.

Conclusion and Implications

This study examined the alignment of the STOU-EPT with the CEFR using Webb's alignment model. Results demonstrated satisfactory alignment across three criteria: categorical occurrence, DOK, and balance of representation.

Item distributions across CEFR levels were as follows: Listening – A2 (8), B1 (8), B2 (9); Structure – A2 (12), B1 (12), B2 (10); Reading – A2 (6), B1 (19), B2 (15). DOK alignment was achieved at B1 and B2 levels in Listening and Structure (% hit = 100, 92.31), while A2 items in these sections showed no alignment (% hit = 0). Reading met DOK criteria at all levels. Q-index values indicated adequate representation: Listening (0.853), Structure (0.952), and Reading (0.817). Cut scores were set on the total score scale (0–100): A2 = 1–55, B1 = 56–76, B2 = 77–100.

Taken together, these findings yield important implications for language assessment practice - theoretical advancement, and educational policy. Practically, the use of CEFR-aligned cut scores supports fairer decisions in course placement and graduation. Theoretically, the study demonstrates the utility of Webb's criteria in language testing—a field where such applications remain limited. Policy-wise, the results respond to the Office of the Higher Education Commission's (2016) call for standardized English benchmarks in Thai universities.

While the expert panel comprised individuals from multiple institutions, its small size might have limited the diversity of perspectives. Moreover, the lack of cognitively appropriate A2 items highlights the need for improved item design at foundational levels. Future research should validate the cut scores with a broader CEFR-calibrated sample and examine the cognitive demands of item formats to enhance the validity of CEFR-referenced assessments.

Ethical Considerations

The study was ethically reviewed and approved by the Ethical Review Committee for Research Involving Human Subjects, STOU, Thailand (Protocol No. STOUIRB 2566/011.2501).

Acknowledgements

The research was funded by the Distance Education Division of STOU. Many thanks are extended to the Fund administration.

About the Authors

Thanyasinee Laosum: An assistant Professor at the Office of Registration, Record and Evaluation, STOU, Nonthaburi, Thailand, holds a Ph.D. in Educational Measurement and Evaluation from Chulalongkorn University. Her research focuses on educational measurement, learning development, student well-being, online test design, and support for students with special needs.

Anusorn Koedsri: An assistant Professor at the Office of Registration, Record and Evaluation, STOU, Nonthaburi, Thailand, holds a Ph.D. in Educational Measurement and Evaluation from Chulalongkorn University. His research focuses on educational measurement, computerized adaptive testing, R programming in educational assessment, and analyzing learning behaviors in MOOCs using machine learning and deep learning.

References

Al Lawati, Z. A. (2023). Investigating the characteristics of language test specifications and item writer guidelines, and their effect on item development: A mixed-method case study. *Language Testing in Asia*, *13*(1), 21. https://doi.org/10.1186/s40468-023-00233-5

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

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663. http://www.jstor.org/stable/4626093

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.

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

Baron, P. A., & Papageorgiou, S. (2016). *Setting language proficiency score requirements for English-as-a-second-language placement decisions in secondary education* (Research Report ETS RR-16-17). Educational Testing Service. https://doi.org/10.1002/ets2.12102

British Council & EAQUALS. (2015). *British Council – EAQUALS core inventory for general English* (2nd ed.). https://www.teachingenglish.org.uk/sites/teacheng/files/pub-british-council-eaquals-core-inventoryv2.pdf

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

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

DeMars, C. (2010). *Item response theory*. Oxford University Press.

EAQUALS. (2008). *EAQUALS bank of descriptors – As checklists.* https://www.eaquals.org/wp-ontent/uploads/EAQUALS_Bank_as_checklists.pdf

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, *66*(4), 477–485. https://doi.org/10.1093/elt/ccs037

Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems.* Council of Chief State School Officers.

Green, A. (2020). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). American Council on Education/Praeger.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern*

Language Journal, *91*(4), 663–667. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x

Jiraro, S., & Angsuchoti, S. (2022). The concurrent validity of Sukhothai Thammathirat Open University's English proficiency test with the computer system under the Common European Framework of Reference for Languages. *ASEAN Journal of Open and Distance Learning (AJODL)*, *14*(1), 98–107.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education /Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

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

Ngudgratoke, S. (2018). *Assessment and evaluation for standard-based education management*. Chatuporn Design.

North, B. (2000). *The development of a common framework scale of language proficiency*. https://www.researchgate.net/publication/312489070

Office of the Higher Education Commission. (2016). *Policy on enhancing English standards in higher education institutions 2016*. http://www.dqe.mhesi.go.th/front_home/Data%20Bhes_2559/04052559.pdf

Papageorgiou, S., Wu, S., Hsieh, C.-N., Tannenbaum, R. J., & Cheng, M. (2019). *Mapping the TOEFL iBT®test scores to China's standards of English language ability: Implications for score interpretation and use* (Research Report No. RR–19-44). Educational Testing Service. https://doi.org/10.1002/ets2.12281

Plake, B. S., & Cizek, G. J. (2012). The modified Angoff, extended Angoff, and yes/no standard setting methods. In G.J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 181–253). Routledge.

Roach, A. T., & Elliott, S. N. (2004). *Alignment analysis and standard-setting procedures for alternate assessments* (WCER Working Paper No. 2004–1). Wisconsin Center for Education Research, University of Wisconsin–Madison.

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate sssessment. *The Journal of Special Education*, *38*(4), 218–231.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Tijdschrift voor Onderwijsresearch*, 2, 49–60.

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

Tannenbaum, R. J. & Baron, P. A. (2011). *Mapping TOEIC® ITP scores onto the Common European Framework of Reference* (Research Memorandum ETS RM—11–33). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-11-33.pdf

Traynor, A., Webb, N. L., Christopherson, S., & Sato, E. (2020). *Using the results of content alignment analyses to inform ongoing item-level improvements to an assessment program: A guide for state departments of education and for assessment vendors.* Wisconsin Center for Education Products and Services.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Council of Chief State School Officers. https://files.eric.ed.gov/fulltext/ED414305.pdf

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7–25. http://dx.doi.org/10.1207/s15324818ame2001_2

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