PASAA Journal
Special Issue 2025/2026, 166–201

Development and Validation of an Algorithmic Randomization Model for a CEFR-Aligned English Proficiency Test for College Students

Jamjumrat Deeprom*

Srinakharinwirot University, Bangkok, Thailand

**Corresponding author: jamjumrat@g.swu.ac.th*

Article information	
Abstract	Amid growing demands for standardized, scalable, and equitable English language assessments in higher education, institutions face challenges related to item exposure, content imbalance, and the resource-intensive nature of manual test construction. This study developed and validated an algorithm-based item randomization and test assembly model for constructing CEFR-aligned assessments from a pre-validated item bank. Grounded in Assessment Engineering and expert input, the model comprises five components—Listening, Vocabulary, Usage and Functional Language, Structure, and Reading—mapped to CEFR levels A2 to C1. Phase 1 involved focus group consultation with English language specialists to design a randomized test blueprint. Phase 2 assessed psychometric properties using confirmatory factor analysis and reliability testing with 300 undergraduates. The test demonstrated excellent model fit, high internal consistency (α = .806–.894), and a clear factorial structure, with Usage and Functional Language emerging as the strongest predictor of overall proficiency. The algorithm ensured thematic balance, avoided item repetition, and upheld difficulty calibration—overcoming common challenges in manual test construction. These results support the model's feasibility and relevance as a

E-ISSN: 2287-0024

	scalable solution for modernizing English language assessment in higher education.
Keywords	Algorithm-based test generation, SWU-SET, CEFR, psychometric properties, computerized test system
APA citation:	Deeprom, J. (2025/2026). Development and validation of an algorithmic randomization model for a CEFR-aligned English proficiency test for college students [Special Issue]. *PASAA, 73*, 166–201.

1. Introduction

Globally, education systems are placing increasing emphasis on outcomes-based learning and standardized assessments to ensure that students develop competencies aligned with international benchmarks (UNESCO, 2023). Among these, English language proficiency has emerged as a critical skill in higher education due to its significant role in academic achievement, professional advancement, and cross-cultural communication (Goodman et al., 2024; Zhou et al., 2024). Reflecting this global trend, Thailand's ongoing educational reform prioritizes the enhancement of learning quality and the holistic development of human potential (Office of the Education Council, 2023). A central objective is to raise the English proficiency of university students, thereby preparing graduates to compete and communicate effectively in the global arena (Li et al., 2022).

In alignment with these national objectives, Srinakharinwirot University has introduced institutional policies aimed at strengthening the quality of teaching and learning (Office of the Education Council, 2023). These initiatives are part of a broader global movement toward the professionalization of academic staff, emphasizing curriculum innovation, evidence-based instructional practices, and rigorous assessment strategies. Faculty members are increasingly expected to play a central role in developing 21st-century skills, fostering lifelong learning, and enhancing national and regional competitiveness—particularly within the ASEAN context (Office of the Higher Education Commission, 2024; Trilling & Fadel, 2009).

To advance these institutional priorities, the university established the Srinakharinwirot University Standardized English Test (SWU-SET). Administered by the Language and Academic Services Center under the International College for Sustainability Studies, the SWU-SET functions as a standardized tool for evaluating students' English language proficiency. Importantly, it is aligned with the Common European Framework of Reference for Languages (CEFR), which provides a widely accepted benchmark for language competence in higher education (Athiworakun & Wudthayagorn, 2018). While the SWU-SET plays a critical role in academic quality assurance, its current development process remains largely manual—relying on item writers and lacking integration with modern technological advancements. This hinders its scalability and suitability for digital testing platforms.

To address these limitations, recent innovations in educational measurement—such as Automatic Item Generation (AIG)—offer promising alternatives (Circi et al., 2023). AIG utilizes computer algorithms to generate test items from structured templates based on psychometric and cognitive models. Grounded in Assessment Engineering (AE), AIG systematically incorporates validity, reliability, and content balance into test design (Embretson & Yang, 2007; Gierl & Haladyna, 2013). The process typically involves three stages: (1) defining the cognitive domains to be assessed, (2) constructing item models with defined difficulty and content levels, and (3) generating items using algorithmic methods (Gierl et al., 2008). This approach allows for rapid development of diverse, high-quality items, aligning well with frameworks like the CEFR (Gierl & Lai, 2013).

Empirical research supports the viability of this approach. For example, Rafatbakhsh et al. (2020) developed and validated an automatic item generation (AIG) system for English idioms using corpus-based semantic modeling. The system demonstrated high validity, with expert judgments and Rasch analysis indicating that approximately 67.5% of generated items were acceptable without revision, and there was strong alignment in item difficulty ratings between the

system and expert evaluations. These findings highlight the potential of algorithmic methods to enhance the efficiency and psychometric quality of language test construction.

Despite these international advancements, the current version of the SWU-SET has not yet integrated Automatic Item Generation (AIG) into its development. Continued reliance on manual item writing restricts the test's responsiveness to increasing demands for scalable, computer-based testing and limits its ability to ensure consistent psychometric quality. Introducing an algorithmic randomization model has the potential to address these challenges by enabling the generation of a broader and more balanced item pool, while ensuring alignment with CEFR proficiency levels. Research by Falcão et al. (2023) reinforces this potential, demonstrating that items generated through AIG are comparable in structure and quality to those developed manually by expert item writers. Their findings suggest that AIG not only maintains psychometric integrity but also enhances the efficiency, consistency, and scalability of item development—key advantages for modern language assessment systems such as the SWU-SET.

Beyond full AIG systems, several empirical studies have examined algorithm-based test assembly and structured random item selection models within computerized assessment contexts (Jatobá et al., 2020; Rice et al., 2022). These studies have indicated that blueprint-driven randomization can reduce item exposure, ensure content balance, and maintain measurement stability when implemented within validated item banks. However, many institutional testing programs continue to rely on fixed-form examinations or partially digitized systems without comprehensive empirical validation of their randomization mechanisms (Kirsten et al., 2026; Van Wijk et al., 2024). Furthermore, relatively few studies have integrated CEFR alignment, structured randomization, and confirmatory factor analysis within a unified development and validation framework tailored for institutional English proficiency testing.

In light of these considerations, the present study aimed to develop and validate an algorithm-based randomization model for constructing CEFR-aligned English proficiency test items within a structured item bank system. To enhance clarity and guide the structure of the investigation, the study was driven by the following research questions: (1) How can a rule-based algorithmic model be designed to assemble CEFR-aligned English proficiency test forms through structured randomization within predefined item pools? and (2) Do the algorithmically assembled test forms demonstrate satisfactory psychometric properties, including reliability, construct validity, and appropriate item difficulty alignment?

Accordingly, the research has two primary objectives: (1) to design a structured algorithm-based randomization model integrated with test item bank software, and (2) to examine the psychometric properties of the assembled English proficiency test during trial administration. By focusing on enhancing the SWU-SET for use among university students, this research sought to improve test assembly efficiency, strengthen psychometric rigor, and support the transition toward computer-based testing (CBT). The outcomes of this study are expected to contribute to institutional quality assurance, reinforce national educational reform efforts, and promote fair, accurate, and sustainable language assessment practices within Thailand's higher education system.

2. Literature Review

This literature review explores key areas relevant to the present study, including CEFR-based language assessment, principles of test validity and reliability, item randomization models, and the role of technology in language testing. The review aims to situate the current research within existing scholarship and identify gaps that the present study sought to address.

2.1 The Development of an Algorithm-based Item Randomization Model from a Test Item Bank for Generating Standardized English Proficiency Tests

The development of standardized English proficiency assessments increasingly incorporates technological innovation, particularly through algorithm-based item randomization and computerized item banking. A test item bank is a structured repository containing a large collection of calibrated test items designed to assess specific language competencies across defined proficiency levels. The establishment of a robust item bank is a critical component in modern test development and administration, allowing for dynamic test construction, efficient item management, and adaptive testing capabilities (Weiss, 2013). Globally, large-scale testing systems have transitioned from static test forms toward algorithm-supported assembly procedures to enhance scalability, test security, and psychometric consistency (Song et al., 2025).

The design and implementation of an item bank involve several systematic steps: analyzing the needs of end users, developing test specifications, generating and reviewing test items, storing items with appropriate metadata, conducting pilot testing, and refining items based on psychometric evidence. Once finalized, the item bank supports efficient item selection through algorithmic randomization, promoting content validity, item balance, and security in repeated testing contexts (Haladyna & Rodriguez, 2013). International empirical studies have demonstrated the effectiveness of algorithm-based test assembly systems in maintaining blueprint fidelity while minimizing item exposure (Luo, 2020; Proietti et al., 2020). In many global contexts, structured randomization models have been validated using Item Response Theory (IRT) or Rasch modeling to ensure item calibration and scale invariance across administrations (Dunn, 2024; Eghan et al., 2026).

From a theoretical standpoint, the literature on educational measurement and language assessment underscores the importance of aligning test content with established standards. The present study employed CEFR as a guiding framework for test item development. The CEFR is widely recognized for its

structured proficiency levels (A1 to C2) and its comprehensive descriptors that guide the evaluation of the four core language skills: listening, speaking, reading, and writing (Council of Europe, 2020). Empirical studies conducted across Europe and Asia have adopted CEFR alignment procedures in test development, often combining blueprint-driven assembly with psychometric validation using confirmatory factor analysis (CFA) or Rasch analysis (Kim & Crossley, 2020; Schnoor et al., 2023). These studies have highlighted the importance of ensuring both content alignment and structural validity when implementing CEFR-based assessments.

In regional contexts, particularly within ASEAN and East Asia higher education systems, institutions have increasingly implemented CEFR-aligned English proficiency tests to support educational reform and international benchmarking (Khamboonruang, 2025; Khan et al., 2023). However, validation approaches vary considerably. Some studies have relied primarily on classical test theory indices, such as reliability coefficients and item discrimination (Zhu et al., 2023), while others have adopted the Rasch model (Anggia & Habók, 2023; Mohd Noh & Mohd Matore, 2022). Despite these advances, relatively few empirical studies in the regional context have explicitly examined structured algorithm-based item randomization systems integrated with comprehensive structural validation models.

At the local level in Thailand, English proficiency assessment reforms have emphasized CEFR alignment and computerized testing platforms. Existing empirical studies have examined test validity, reliability, and alignment with national educational standards (Cheewasukthaworn, 2022; Waluyo et al., 2024). Nevertheless, many institutional assessments continue to rely on manually assembled test forms or static item pools. While psychometric validation using classical test theory has been reported, systematic integration of algorithm-based randomization within a predefined item bank—combined with confirmatory structural validation—remains limited in the Thai higher education context.

In addition to test content and structural alignment, psychometric validation plays a critical role in evaluating the quality and fairness of test items. Fundamental metrics such as item difficulty, discrimination index, reliability coefficients, and construct validity are central to psychometric analyses (AERA et al., 2014). These indices inform test developers about how well the items function in differentiating learners across ability levels, thus supporting the refinement of both the item bank and the randomization model used in test delivery. Contemporary validation frameworks increasingly recommend multi-level validation procedures, including content validation, structural validation, and internal consistency analysis, to ensure comprehensive evidence of test quality.

Moreover, studies in automatic item generation (AIG) have highlighted the utility of algorithm-based models for efficiently generating and selecting items from item banks. AIG relies on templates and rule-based frameworks grounded in cognitive models and language constructs to generate items systematically (Gierl & Lai, 2013). Although automatic item generation (AIG) focuses primarily on item creation through cognitive models and template-based design (Kıyak & Kononowicz, 2025), algorithm-based randomization models operate through structured selection procedures that assemble test forms from pre-validated item banks containing calibrated psychometric information (Eghan et al., 2026). Empirical evidence across computer-based assessment contexts suggests that such algorithmic assembly enhances scalability and reduces test construction bias while preserving measurement consistency (Tomasik et al., 2018). However, validation research emphasizes that assembled test forms must undergo structural and construct validation to ensure that randomization does not compromise construct representation or score interpretability (Fuchimoto & Songmuang, 2026).

3. Methodology

This study employed a two-phase design integrating qualitative and quantitative methodologies to develop and validate an algorithm-based item

randomization and test assembly model for constructing CEFR-aligned assessments from a pre-validated item bank.

3.1 Phase 1: Designing the Algorithm-based Item Randomization Model for the English Proficiency Test

This phase involved the design and development of an algorithm-based item randomization model guided by documentary research and qualitative input from domain experts. The literature review focused on existing models of item randomization, test item banking, and the CEFR alignment. The results informed the formulation of operational definitions, a conceptual model, and a test item randomization framework compatible with CEFR proficiency levels.

3.1.1 Target Group

In the qualitative phase, participants comprised ten English language instructors from autonomous public universities in Thailand, purposively selected for their expertise in English instruction and language assessment. Participants were required to meet at least one of the following criteria: (1) CEFR level B2 or higher on a standardized English proficiency test within the past two years; (2) authorship of publications related to English language instruction; or (3) at least five years of teaching experience with involvement in English proficiency test design. These criteria ensured inclusion of information-rich experts capable of contributing to the development of a CEFR-aligned algorithm-based item randomization model.

The selection of key informants was guided by qualitative research principles emphasizing purposive expert sampling and data saturation rather than statistical representativeness. Prior methodological research suggests that thematic saturation in relatively homogeneous expert groups is typically achieved with nine to 17 participants (Hennink & Kaiser, 2022). As discussion themes stabilized and no new insights emerged in later stages, ten participants were deemed sufficient to provide robust expert input for model development.

3.1.2 Instruments

The primary instrument used in the qualitative phase of this study was a focus group discussion guide, developed to facilitate structured yet open-ended exploration of expert perspectives on algorithm-based item randomization for CEFR-aligned English proficiency tests. The guide included questions designed to elicit insights into the principles, strategies, and challenges of test item development and randomization. It was developed through a three-step process: (1) reviewing relevant literature on item bank development and CEFR alignment (Berger, 2020; Huang et al., 2021; Weiss, 2013); (2) drafting question prompts to explore aspects such as item construction, difficulty distribution, and practical implementation of randomization; and (3) submitting the draft guide to a panel of three experts in language assessment and educational measurement for content validation. Revisions were made based on their feedback to enhance clarity and relevance. A focus group recording form was also employed to document participant responses and discussion flow systematically. Together, these instruments supported the collection of in-depth qualitative data essential for informing the initial design of the algorithm-based item randomization model.

3.1.3 Data Collection

Invitation letters were sent directly to potential participants who met at least one of the inclusion criteria, and those who expressed interest and consented to participate were included in the study. The final focus group consisted of ten instructors with diverse expertise in English language education and assessment. The session was conducted in person at a centrally accessible academic venue and lasted approximately 90 minutes. All discussions were audio-recorded with participant consent, and additional notes were taken using the focus group recording form to ensure accurate data capture for analysis.

3.1.4 Data Analysis

After finalizing the draft of the algorithm-based item randomization model, the research team presented it to a focus group of ten experts for critical feedback

and validation. The qualitative data obtained from this focus group were analyzed using content analysis to systematically identify recurring patterns, themes, and expert insights relevant to the model's functionality, feasibility, and alignment with CEFR standards. Verbatim transcripts of the discussion were reviewed to extract meaningful statements, which were then coded and grouped into thematic categories corresponding to key components of the model.

3.2 Phase 2: Evaluating the Quality of Item Randomization and the Psychometric Properties of the English Proficiency Test

This phase focused on the implementation and evaluation of the algorithm-based item randomization model developed in Phase 1. The evaluation process consisted of two main components: (1) assessing the quality and practicality of the item randomization mechanism, and (2) examining the psychometric properties of the English proficiency test generated using the randomized item model.

3.2.1 Participants

The participants in this phase consisted of 300 undergraduate students who were recruited using convenience sampling from various academic programs at Srinakharinwirot University. The sample size was determined based on statistical considerations for structural equation modeling (Soper, 2025), with an anticipated effect size of 0.25, statistical power of 0.80, and a significance level of .05. Given the model included five latent variables and 20 observed variables, a minimum of 229 participants was required to ensure sufficient analytical power. The final sample of 300 students exceeded this threshold, thereby supporting the robustness of the confirmatory factor analysis and reliability assessment conducted in this study.

3.2.2 Instruments

To evaluate the quality of randomized item generation and the psychometric properties of the English proficiency test, a structured set of instruments was developed through a systematic, multi-step process. These included the CEFR-

aligned randomized test and expert evaluation tools for content validation and practicality assessment.

The development process began with a review of theoretical and empirical literature related to item randomization models and computerized test construction. A test specification table was then created, outlining five core components: Listening, Vocabulary, Usage and Functional Language, Structure, and Reading. Each component was mapped to CEFR proficiency levels (A2, B1, B2, and C1), ensuring balanced item distribution and content relevance.

Test items were drafted to reflect real-world language usage and varied in difficulty according to CEFR standards. The test blueprint guided the proportional distribution of items, with a total of 100 items (See Table 1).

Table 1

Test Blueprint by CEFR Level and Component

Component	A2	B1	B2	C1	Notes
Listening Part	5	5	6	4	General and academic contexts; 4 audio tracks
Vocabulary Part	5	5	5	5	Randomized within level; unique underlined vocabulary
Usage & Functional Language Part	5	5	5	5	4 conversation-based scenarios per set
Structure part	4	6	6	4	Multiple-choice; cloze texts with 5 blanks (80–120 words per passage)
Reading Part	4	6	6	4	Texts ranging from 250–500 words based on CEFR levels
Total Items	**23**	**27**	**28**	**22**	

Content validity was assessed by five experts—three in linguistics and two in educational measurement—using the Index of Item-Objective Congruence (IOC), with a minimum threshold of ≥ .50 (Rovinelli & Hambleton, 1977). The IOC values for all test items ranged from 0.60 to 1.00, indicating acceptable to strong alignment between the test items and the intended constructs across CEFR levels. Additionally, a group of 30 third-year undergraduate students participated in a pilot trial to assess the alignment of the randomized test items with the CEFR framework, ensuring item appropriateness across proficiency levels.

A separate panel of three experts (two linguists and one evaluation specialist) reviewed the test's practicality, including item clarity and interface suitability for computerized delivery. Feedback from both reviews informed item revision to improve clarity, content alignment, and fairness. The finalized instruments served as the foundation for pilot testing and subsequent psychometric evaluations in the second phase of the study.

3.2.3 Data Collection

Recruitment was conducted through announcements posted on university bulletin boards, faculty offices, and digital platforms such as email and university-affiliated social media groups. Students were informed about the voluntary nature of their participation, the purpose of the study, and the confidentiality of their responses. Those who agreed to participate signed an informed consent form prior to taking the test. The test was administered in a controlled environment using the computerized testing platform, and the resulting data were used to evaluate reliability, construct validity, and item-level statistics.

3.2.4 Data Analysis

To examine the construct validity of the randomized English proficiency test, Confirmatory Factor Analysis (CFA) was conducted using both first-order and second-order models. The first-order CFA tested the factorial structure of the five observed components—Listening, Vocabulary, Usage and Functional Language,

Structure, and Reading—each represented by multiple test items. The second-order CFA was performed to determine whether these five domains could be explained by a single underlying latent factor representing overall English proficiency. Model fit was evaluated using indices recommended by Hair et al. (2018), with the following thresholds: Chi-square to degrees of freedom ratio (χ^2/df) < 3, Goodness-of-Fit Index (GFI) > 0.95, Comparative Fit Index (CFI) > 0.95, Root Mean Square Error of Approximation (RMSEA) < 0.05, and Standardized Root Mean Square Residual (SRMR) < 0.05.

Prior to CFA, Pearson correlation coefficients were computed among all subcomponents to examine multicollinearity. All correlation values were below the recommended threshold of 0.85 (Kline, 2016), indicating no multicollinearity and sufficient discriminant validity among the constructs.

To assess the suitability of the dataset for factor analysis (Field, 2009), Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity were performed. The KMO value exceeded the minimum acceptable threshold of 0.70, indicating adequate sampling adequacy. Bartlett's Test was statistically significant ($p < .001$), suggesting that the correlation matrix was factorable and appropriate for CFA.

Finally, reliability analysis was conducted using Cronbach's alpha coefficient to assess the internal consistency of the test. Each component of the test demonstrated acceptable to high reliability, with Cronbach's alpha values exceeding the commonly accepted cutoff of 0.70, confirming the test's internal coherence across all language skill domains.

4. Results

4.1 Results of Phase 1

To design an algorithm-based randomized question model for the English proficiency test at Srinakharinwirot University (SWU), the researcher reviewed

relevant literature and adapted the existing paper-based format for use with a digital question bank system. The model was aligned with the Common European Framework of Reference for Languages (CEFR) and refined through input from a focus group of ten English language experts. The results are summarized below.

4.1.1 Results of the Randomized Question Model Design

The final model consisted of five core sections: Listening, Vocabulary, Usage and Functional Language, Structure, and Reading. Each section was designed to ensure diversity of content, controlled item difficulty, and fairness through CEFR-level alignment and non-redundant item selection (See Table 2).

Table 2

Design Specifications and Randomization Criteria for CEFR-Aligned English Proficiency Test Sections

Test Section	Themes/Topics	Item Pool/Source	Randomization Rules
1. Listening	14 thematic areas (e.g., health, climate, entertainment); shared with other sections	160 audio tracks across CEFR levels A2–C1	Randomized by CEFR level without repeating themes; fixed question order, shufflable answers
2. Vocabulary	Selected based on CEFR difficulty levels; avoid duplicates or near-synonyms	20 vocabulary items per test	Avoid duplicates or similar words; maintain CEFR alignment
3. Usage and Functional Language	14 themes aligned with Listening; shared with Reading section	160 dialogues classified by CEFR levels	No theme repetition across items; fixed order, shufflable answers
4. Structure	Two subparts: discrete grammar and short passages (5 blanks each)	Items span four CEFR levels with diverse grammar topics	Grammar topics randomized by CEFR level; diverse themes in passages

Test Section	Themes/Topics	Item Pool/Source	Randomization Rules
5. Reading	14 themes consistent with other sections; 160 passages categorized	160 categorized passages by CEFR levels	Randomized by CEFR level; no theme repetition; fixed question order, shufflable answers

4.1.2 Focus Group Results on the Randomized Question Model

After presenting the initial model, the research team conducted a focus group discussion with ten English language experts. Their feedback was instrumental in refining the model prior to implementation (See Table 3).

Table 3

Expert Feedback on Randomized Item Design

Test Section	Expert Recommendations
1. Listening	Ensure variety in conversation types (e.g., academic vs. general); encode contextual diversity to prevent similar scenarios in one test.
2. Vocabulary	Align word difficulty strictly with CEFR levels; exclude near-synonyms to avoid inflated difficulty and redundancy.
3. Usage and Functional Language	Match dialogue length and complexity to CEFR levels; avoid similar scenarios in function or topic.
4. Structure	Avoid repetition of grammar topics within the same CEFR level; ensure vocabulary supports the grammar structure; use distinct themes for cloze passages.
5. Reading	Review thematic distributions holistically; avoid theme repetition; maintain contextual balance for perceived fairness.

In summary, the focus group confirmed the theoretical soundness of the proposed model and provided detailed, actionable recommendations to ensure that randomized item sets are psychometrically balanced and pedagogically valid.

The refined model was then finalized for implementation and subsequent evaluation in Phase 2.

4.2 Results of Phase 2

Prior to conducting CFA, preliminary psychometric analyses were performed to evaluate item quality and assess the suitability of the dataset for factor analysis. Item analysis from the pilot testing indicated that item difficulty indices ranged from 0.28 to 0.72, reflecting levels from moderately difficult to moderately easy. The discrimination indices ranged from 0.25 to 0.62, demonstrating moderate to high discriminative power across items. These findings indicated that all test items met acceptable psychometric criteria and were appropriate for inclusion in subsequent structural analyses.

In addition, preliminary analyses were performed to assess the interrelationships among the five subcomponents of the English proficiency test and to determine the suitability of the dataset for factor analysis. Pearson correlation coefficients between the subcomponents ranged from 0.401 to 0.723, indicating moderate to strong positive correlations. These values suggest sufficient relatedness for factor analysis without indicating multicollinearity. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy was 0.887, which exceeded the recommended minimum threshold of 0.70 and demonstrated the adequacy of the sample for factor analysis. Additionally, Bartlett's Test of Sphericity was statistically significant, $\chi^2(100) = 923.478$, $p < .001$, confirming that the correlation matrix was factorable and appropriate for structure detection through CFA.

4.2.1 Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) was conducted to validate the factorial structure of the English proficiency test, employing both first-order and second-order models. The first-order CFA evaluated the latent constructs underlying the five components—Listening, Vocabulary, Usage and Functional Language,

Structure, and Reading—while the second-order CFA examined whether these domains collectively represented a single latent variable: overall English language proficiency.

The model demonstrated an excellent fit to the data based on widely accepted criteria by Hair et al. (2018). The fit indices were as follows: χ^2 = 130.627, df = 122, p = .280; χ^2/df = 1.07; GFI = 0.96; CFI = 1.00; RMSEA = 0.01; and SRMR = 0.03. All indices were within the recommended thresholds, indicating a highly acceptable model fit.

Factor loadings from the first-order CFA ranged from 0.648 to 0.821 across all CEFR levels within each of the five components. The Listening section had its highest loading at the A2 level (0.811), Vocabulary at B2 (0.789), Usage and Functional Language at B1 (0.784), Structure at C1 (0.821), and Reading at C1 (0.809). These values reflected strong internal structure across levels of language proficiency.

Confirmatory Factor Analysis (CFA) was used for both first-order and second-order model validation of the five test components.

In the second-order CFA, the standardized factor loadings of the five observed domains on the overarching construct of English proficiency ranged from 0.886 to 0.997. The Usage and Functional Language component demonstrated the strongest association with the general proficiency factor (loading = 0.997, R^2 = 0.993), followed by Vocabulary (loading = 0.987, R^2 = 0.974), Reading (loading = 0.970, R^2 = 0.941), Listening (loading = 0.926, R^2 = 0.858), and Structure (loading = 0.886, R^2 = 0.785). These results suggested that all components meaningfully contributed to the overall construct and confirmed the test's robust factorial validity (See Table 4).

Table 4

First-Order and Second-Order Confirmatory Factor Analysis of the Algorithmically Generated English Proficiency Test

Variable	Factor Loading		t	R^2
	b (SE)	β		
First-order Factor Loadings				
Listening part				
A2 Level	1.000	.811	-	.658
B1 Level	0.931 (0.070)	.723	13.284*	.522
B2 Level	1.125 (0.083)	.798	13.601*	.636
C1 Level	0.907 (0.071)	.699	12.762*	.488
Vocabulary part				
A2 Level	1.000	.651	-	.424
B1 Level	1.1.81 (0.107)	.725	11.070*	.525
B2 Level	1.184 (0.093)	.789	12.666*	.622
C1 Level	1.243 (0.110)	.745	11.296*	.555
Usage and Functional Language part				
A2 Level	1.000	.743	-	.551
B1 Level	1.134 (0.082)	.784	13.796*	.615
B2 Level	0.948 (0.071)	.760	13.326*	.577
C1 Level	0.743 (0.066)	.648	11.326*	.420
Structure part				
A2 Level	1.000	.736	-	.542
B1 Level	1.296 (0.099)	.799	13.138*	.638

B2 Level	1.240 (0.128)	.769	9.658*	.591
C1 Level	1.181 (0.111)	.821	10.676*	.675
Reading part				
A2 Level	1.000	.716	-	.513
B1 Level	1.069 (0.082)	.789	13.039*	.623
B2 Level	1.043 (0.095)	.670	11.028*	.449
C1 Level	1.106 (0.091)	.809	12.100*	.655
Second-order Factor Loadings				
Listening	1.000	.926	-	.858
Vocabulary	0.920 (0.076)	.987	12.155*	.974
Usage and Functional Language	1.143 (0.087)	.997	13.111*	.993
Structure	0.903 (0.078)	.886	11.649*	.785
Reading	1.053 (0.086)	.970	12.303*	.941

Note: b = unstandardized factor loading; SE = standard error; β = standardized factor loading. * $p < .01$

4.2.2 Reliability

Reliability analysis using Cronbach's alpha revealed that all five components of the English proficiency test demonstrated strong internal consistency, with alpha coefficients ranging from 0.806 to 0.894, exceeding the commonly accepted threshold of 0.70.

5. Discussion

The findings of this study have affirmed the feasibility and psychometric soundness of the algorithm-based randomization model developed for the SWU-

SET English proficiency test. Guided by a rigorous literature review and expert input, the model demonstrated strong alignment with the CEFR framework, effective randomization procedures, and robust validity and reliability indicators.

The qualitative phase was essential in shaping the SWU-SET test by incorporating expert input from language education and assessment specialists. Their feedback ensured that the test's structure, item difficulty, and thematic content were theoretically grounded and practically relevant. Focus group discussions served as content validation, guiding the refinement of item specifications and the randomization model. A key outcome was the alignment of item difficulty with CEFR levels. Experts confirmed that items from A2 to C1 reflected appropriate linguistic and cognitive demands, helping resolve borderline cases and maintain a coherent progression—consistent with established language assessment standards (Al Lawati, 2023).

A key issue addressed in the qualitative phase was the avoidance of thematic redundancy across test sections. Repeated topics, such as similar themes in reading and listening tasks, can introduce construct-irrelevant variance by inflating scores through familiarity rather than actual language ability (Chen et al., 2025). To prevent this, experts recommended embedding thematic controls in the algorithm to eliminate duplication across the Listening, Reading, and Usage and Functional Language sections, thereby enhancing fairness and measurement precision. Experts also guided the alignment of linguistic complexity with CEFR levels to ensure developmental appropriateness and cultural neutrality. For example, B2 grammar items included hypothetical modals, while A2 items focused on the present simple tense—supporting content validity (Council of Europe, 2020; Fulcher & Davidson, 2007). These expert-informed adjustments improved not only fairness but also authenticity. Incorporating diverse themes, language registers, and real-life contexts strengthened the ecological validity of the test (Gan, 2024), aligning with communicative language testing (Bachman & Palmer, 1996; Harding,

2014; Liao et al., 2023), which emphasizes that language proficiency should be assessed through tasks that reflect real-world language use.

The quantitative findings from Phase 2 further supported the model's effectiveness. The high values observed across multiple fit indices in both the first-order and second-order confirmatory factor analyses indicated the strong structural validity of the test. These results confirmed that the five skill-based components—Listening, Vocabulary, Usage and Functional Language, Structure, and Reading—collectively measured a coherent underlying construct of English language proficiency. The highest second-order loading was found for the Usage and Functional Language component, suggesting that communicative competence—defined as the ability to use language appropriately and fluently in real-life contexts—was central to English proficiency among university learners. This finding aligned with influential communicative frameworks in second language acquisition (Aljumah, 2020; Eisenchlas, 2009; Nikolaus & Fourtassi, 2023), which place functional use of language at the heart of proficiency.

While much of the traditional language assessment literature categorizes English proficiency in terms of the four primary macro skills—listening, speaking, reading, and writing—our study offers a more nuanced model by disaggregating these into five subcomponents that better reflect the cognitive and functional dimensions of language use in an academic context. In particular, the separation of Vocabulary and Usage and Functional Language is theoretically grounded in communicative competence frameworks, which distinguish linguistic knowledge from the ability to use language appropriately in context (Bachman & Palmer, 2010; Canale & Swain, 1980). Vocabulary primarily represents lexical knowledge, including breadth and control of word meaning, whereas Usage and Functional Language reflects pragmatic competence—the ability to deploy grammatical structures and expressions to perform communicative functions such as expressing intentions, managing interaction, and conveying meaning appropriately. This distinction is consistent with CEFR descriptors, which differentiate linguistic

competence (e.g., lexical range and accuracy) from pragmatic competence involving functional language use in real-life communication (Council of Europe, 2020).

Rather than treating productive skills as isolated writing or speaking abilities, the Usage and Functional Language component captures integrated productive and pragmatic competencies, including grammar, appropriacy, and idiomatic expression. Such differentiation enables a more granular and communicatively oriented assessment of proficiency, particularly suited to computer-based and algorithm-driven testing environments. In contrast, previous studies such as Aizawa et al. (2023) and Piamsai (2023) have adopted the conventional four-skill framework and reported clustering into receptive (listening and reading) and productive (speaking and writing) domains. The present findings have extended this perspective by demonstrating that vocabulary knowledge and functional language use—often embedded within broader skills—operate as distinct yet complementary constructs when assessment tasks are designed to reflect communicative function rather than formal accuracy alone.

The test's psychometric properties further support its quality. Cronbach's alpha values for the five subscales ranged from 0.806 to 0.894, exceeding the commonly accepted threshold of 0.70 for internal consistency (Nunnally & Bernstein, 1994). This indicates strong reliability within each skill domain. Moreover, the high correlations observed among subcomponents, in the absence of multicollinearity, suggest that the components are conceptually linked but not redundant—each contributes uniquely to the overall construct of proficiency. Additionally, the item-level results, aligned with CEFR proficiency bands, validate the algorithm's randomization logic and difficulty calibration. Items exhibited clear loadings across A2, B1, and B2 levels, demonstrating that the assessment tool successfully stratifies item difficulty in accordance with international standards (Council of Europe, 2020). This confirms the test's potential for adaptive use in placement, diagnostic, and summative assessments. Together, these findings

reinforce the theoretical and empirical soundness of the proposed model and demonstrate that a five-component structure may offer a more communicatively relevant and psychometrically robust alternative to the traditional four-skill framework, especially in the context of algorithm-based, CEFR-aligned language testing for university learners.

Another significant contribution of this study lies in the innovative use of an algorithm-based item randomization model guided by a structured blueprint and thematic control. Unlike traditional test assembly methods that rely on manual selection or static test forms, this approach employed algorithmic rules to dynamically assemble test forms in real time from a pre-validated item bank. The model ensured that each test administration adhered to predefined content specifications—such as skill domains, CEFR-aligned difficulty levels, and thematic balance—thereby preventing content repetition and maintaining balanced coverage across linguistic components (Circi et al., 2023; Pugh et al., 2020; Westacott et al., 2023). This algorithm-based assembly process supported equitable difficulty calibration, a critical requirement for contemporary assessment systems aligned with international standards such as the CEFR (Council of Europe, 2020).

Furthermore, integrating algorithmic item randomization with a digital item bank enhanced the overall assessment design. By embedding logical constraints within the selection algorithm—such as preventing item duplication, enforcing topic diversity, and rotating item types—the system generated multiple equivalent test forms with minimal human intervention. This approach improved consistency in test assembly while reducing human bias and procedural error commonly associated with manually constructed examinations (Gierl & Haladyna, 2013; Leslie & Gierl, 2023). The proposed rule-based randomized assembly model aligns with emerging trends in technology-enhanced assessment systems that emphasize scalability, fairness, and standardized test construction across administrations (Rausch et al., 2016; Russell et al., 2003).

6. Implications

The findings of this study contribute to the theoretical development of language assessment by extending traditional conceptualizations of English proficiency beyond the conventional four-skill framework. While most CEFR-aligned assessments operationalize proficiency through Listening, Speaking, Reading, and Writing, the present study has demonstrated that Vocabulary and Usage and Functional Language function as distinct yet interrelated constructs within an algorithm-based assessment environment. The strong second-order loading observed for the Usage and Functional Language component highlights communicative competence as a central dimension of language proficiency, supporting communicative language testing theories that emphasize functional language use over isolated grammatical knowledge.

By empirically validating a five-component structure, this study has provided evidence that language proficiency may be more accurately represented as a multidimensional construct integrating cognitive, linguistic, and pragmatic competencies. This reconceptualization contributes to ongoing theoretical discussions in second language assessment regarding how communicative ability should be operationalized in digitally mediated testing contexts. Furthermore, the successful integration of CEFR descriptors into algorithmic test assembly strengthens the theoretical linkage between international proficiency frameworks and technology-enhanced assessment models.

Methodologically, this study has demonstrated the viability of an algorithm-based item randomization model as an alternative to traditional manual test construction. The findings have shown that structured algorithmic assembly from a pre-validated item bank could maintain psychometric integrity while improving scalability and consistency across test administrations. Importantly, the study has illustrated that randomization alone was insufficient; rather, effective algorithm-based testing required blueprint constraints, thematic controls, and CEFR-aligned difficulty calibration to preserve construct representation.

The integration of qualitative expert validation with quantitative psychometric evaluation has provided a replicable mixed methods framework for assessment development. The qualitative phase ensured content validity and contextual relevance, while confirmatory factor analysis verified structural validity and internal consistency. This sequential design offered a methodological model for future researchers developing computerized or randomized assessments, particularly in contexts seeking alignment with international standards.

Additionally, the study has highlighted the importance of validating assembled test forms—not only individual items—when algorithm-based item selection and randomization procedures were employed. This contributed to methodological discussions in educational measurement regarding quality assurance in algorithm-driven test assembly from pre-validated item banks and supported emerging best practices in technology-enhanced assessment design.

7. Limitations and Recommendations

Several limitations should be noted. First, the study was conducted at a single autonomous public university in Thailand using a convenience sample, which may limit generalizability. Institutional factors such as curriculum design, teaching practices, student proficiency profiles, and assessment culture may influence score distributions and parameter estimates, potentially affecting model stability across contexts. Caution is therefore needed when applying the model to other institutions, particularly across ASEAN settings where cultural and pedagogical differences may shape the interpretation and performance of CEFR-aligned assessments. Future research should undertake cross-institutional and cross-cultural validation to examine model transferability and ensure fairness across diverse learner populations.

Second, while the mixed methods design—combining focus groups and psychometric validation—provided strong evidence of content and construct validity, it did not incorporate adaptive testing or real-time analytics. Future

versions could benefit from integrating adaptive algorithms to improve user experience and psychometric quality, especially in large-scale or high-stakes contexts.

Lastly, although the automated item randomization logic reduced redundancy and bias, the system was not tested under live operational conditions with diverse test-taker groups and technical environments. Real-time deployment may present challenges—such as interface usability, response delays, or security risks—not captured in this study. Longitudinal research across multiple test cycles would help assess the tool's reliability, stability, and real-world usability.

8. Conclusion

This study developed and validated an algorithm-based item randomization model for a CEFR-aligned English proficiency test, offering a practical and scalable approach to modernizing language assessment in higher education. The five-component structure—Listening, Vocabulary, Usage and Functional Language, Structure, and Reading—demonstrated strong construct validity, internal consistency, and appropriate difficulty alignment with CEFR levels. Among these, Usage and Functional Language stood out as a key predictor of overall proficiency, emphasizing the importance of communicative competence over isolated skill testing. The integration of blueprint-driven item selection and algorithmic randomization from a pre-validated item bank effectively addressed issues of item repetition, content imbalance, and test scalability. By embedding logical constraints and CEFR guidelines within the algorithm, the model ensures fairness, efficiency, and psychometric robustness, positioning it as a viable foundation for computerized or adaptive testing aligned with educational reform goals.

9. About the Author

Jamjumrat Deeprom is a lecturer in the Department of English Language for Higher Education, International College for Sustainability Studies, Srinakharinwirot University, Bangkok, Thailand. Her research interests include English language

teaching and intercultural communication. She can be contacted at jamjumrat@g.swu.ac.th

10. Acknowledgement

This research was funded by the International College for Sustainability Studies, Srinakharinwirot University, under the 2023 revenue budget (Contract No. 438/2566) for the development of the Srinakharinwirot University Standardized English Test (SWU-SET).

11. Declaration of AI Use

The author declares that AI tools, including Grammarly, a GPT-based APA citation tool, and ChatGPT, were used exclusively for proofreading, reference formatting, and minor language editing. No AI tools were involved in the analysis or interpretation of the research data. The author remains fully responsible for the content of the manuscript.

12. References

Aera, A. P. A. (2014). Standards for educational and psychological testing. *American Educational Research Association*.

Aizawa, I., Rose, H., Thompson, G., & Curle, S. (2023). Beyond the threshold: Exploring English language proficiency, linguistic challenges, and academic language skills of Japanese students in an English medium instruction programme. *Language Teaching Research, 27*(4), 837–861. https://doi.org/10.1177/1362168820965510

Aljumah, F. H. (2020). Second language acquisition: A framework and historical background on its research. *English Language Teaching, 13*(8), 200–207. https://doi.org/10.5539/elt.v13n8p200

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*, 1–17. https://doi.org/10.1186/s40468-023-00233-5

Anggia, H., & Habók, A. (2023). Textual complexity adjustments to the English reading comprehension test for undergraduate EFL students. *Heliyon*, *9*(1), Article e12891. https://doi.org/10.1016/j.heliyon.2023.e12891

Athiworakun, C., & Wudthayagorn, J. (2018). Mapping Srinakharinwirot University - Standardized English Test (SWU-SET) onto the Common European Framework of Reference (CEFR). *Suranaree Journal of Social Science, 12*(2), 69–84. https://doi.org/10.55766/CTUU4836

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

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

Berger, A. (2020). Specifying progression in academic speaking: A keyword analysis of CEFR-based proficiency descriptors. *Language Assessment Quarterly*, *17*(1), 85–99. https://doi.org/10.1080/15434303.2019.1689981

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47. https://doi.org/10.1093/applin/I.1.1

Cheewasukthaworn, K. (2022). Developing a standardized English proficiency test in alignment with the CEFR. *PASAA*, *63*, 66–92. https://doi.org/10.58837/CHULA.PASAA.63.1.3

Chen, X., Aryadoust, V., & Zhang, W. (2025). A systematic review of differential item functioning in second language assessment. *Language Testing, 42*(2), 193–222. https://doi.org/10.1177/02655322241290188

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education, 8*, Article 858273. https://doi.org/10.3389/feduc.2023.858273

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.

Dunn, K. J. (2024). Random-item Rasch models and explanatory extensions: A worked example using L2 vocabulary test item responses. *Research*

Methods in Applied Linguistics, 3(3), Article 100143. https://doi.org/10.1016/j.rmal.2024.100143

Eghan, R. E., Osei-Sarpong, E., Awashie, G. E., Borkor, R. N., Yaokumah, E., & N'ganomah, A. A. (2026). Item response theory for trait assessment in randomized item pool for computer based test. *Scientific African, 31*, Article e03226. https://doi.org/10.1016/j.sciaf.2026.e03226

Eisenchlas, S. A. (2009). Conceptualizing 'communication' in second language acquisition. *Australian Journal of Linguistics, 29*(1), 45–58. https://doi.org/10.1080/07268600802516376

Embretson, S. E., & Yang, X. (2007). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics*, Volume 26 (pp. 747–768). Elsevier North Holland.

Falcão, F., Pereira, D. M., Gonçalves, N., De Champlain, A., Costa, P., & Pêgo, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of automatic item generation. *Advances in Health Sciences Education, 28*, 1441–1465. https://doi.org/10.1007/s10459-023-10225-y

Field, A. (2009). *Discovering statistics using SPSS: Introducing statistical method* (3rd ed.). Sage.

Fuchimoto, K., & Songmuang, P. (2026). Review of automated parallel test form assembly. *Behaviormetrika*. https://doi.org/10.1007/s41237-026-00293-w

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Gan, Q. (2024). Different registers, different grammars in second language production? The dative alternation in spoken and written Chinese learner English. *Lingua, 309*, Article 103790. https://doi.org/10.1016/j.lingua.2024.103790

Gierl, M. J., & Haladyna, T. (2013). *Automatic item generation: Theory and practice.* Routledge.

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items.

Educational Measurement: Issues and Practice, 32(3), 36–50. https://doi.org/10.1111/emip.12018

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *Journal of Technology, Learning, and Assessment, 7*(2), 1–51.

Goodman, B., Yessenbekova, K., & Curle, S. (2024). English-medium education in Kazakhstan: A multifaceted exploration of student and alumni perceptions on language proficiency, academic performance, and career prospects. *International Journal of Educational Research, 128*, Article 102451. https://doi.org/10.1016/j.ijer.2024.102451

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate data analysis* (8th ed.). Cengage Learning.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, *11*(2), 186–197. https://doi.org/10.1080/15434303.2014.895829

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine*, *292*, Article 114523. https://doi.org/10.1016/j.socscimed.2021.114523

Huang, H. T. D., Hung, S. T. A., Chao, H. Y., Chen, J. H., Lin, T. P., & Shih, C. L. (2021). Developing and validating a computerized adaptive testing system for measuring the English proficiency of Taiwanese EFL university students. *Language Assessment Quarterly*, *19*(2), 162–188. https://doi.org/10.1080/15434303.2021.1984490

Jatobá, V. M. G., Farias, J. S., Freire, V., Ruela, A. S., & Delgado, K. V. (2020). ALICAT: A customized approach to item selection process in computerized adaptive testing, *Journal of the Brazillian Computer Society, 26*, Article 4. https://doi.org/10.1186/s13173-020-00098-z

Khamboonruang, A. (2025). Argument-based validation of Chulalongkorn University Language Institute (CULI) Test: A Rasch-based evidence investigation. *Language Testing in Asia, 15*, Article 10. https://doi.org/10.1186/s40468-025-00346-z

Khan, A., David, A. R., Ahmad, A. H., Ali, A., & Lah, S. C. (2023). Initial insights into CEFR adoption at a language faculty of a public university in Malaysia. *PASAA, 67*, 330–360. https://doi.org/10.58837/CHULA.PASAA.67.1.11

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. https://doi.org/10.1080/15434303.2020.1775234

Kirsten, K., Greefrath, G., & Emmrich, R. (2026). Technology-based versus paper-pencil: Sources of mode effects in large-scale assessment. *International Journal of Mathematical Education in Science and Technology*, 1–28. https://doi.org/10.1080/0020739X.2025.2584340

Kıyak, Y. S., & Kononowicz, A. A. (2025). Using a hybrid of AI and template-based method in automatic item generation to create multiple-choice questions in medical education: Hybrid AIG. *JMIR Formative Research*, *9*, Article e65726. https://doi.org/10.2196/65726

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Leslie, T., & Gierl, M. J. (2023). Using automatic item generation to create multiple-choice questions for pharmacy assessment. *American Journal of Pharmaceutical Education, 87*(10), Article 100081. https://doi.org/10.1016/j.ajpe.2023.100081

Li, Y., Teng, W., Tsai, L., & Lin, T. M. Y. (2022). Does English proficiency support the economic development of non-English-speaking countries? The case of Asia. *International Journal of Educational Development, 92*, Article 102623. https://doi.org/10.1016/j.ijedudev.2022.102623

Liao, L., Ye, S. X., & Yang, J. (2023). A mini review of communicative language testing. *Frontiers in Psychology, 14*, Article 1058411. https://doi.org/10.3389/fpsyg.2023.1058411

Luo, X. (2020). Automated test assembly with mixed-integer programming: The effects of modeling approaches and solvers. *Journal of Educational Measurement, 57*(4), 547–565. https://doi.org/10.1111/jedm.12262

Nikolaus, M., & Fourtassi, A. (2023). Communicative feedback in language acquisition. *New Ideas in Psychology, 68*, Article 100985. https://doi.org/10.1016/j.newideapsych.2022.100985

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Mohd Noh, M. F., & Mohd Matore, M. E. E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology*, *13*, Article 941084. https://doi.org/10.3389/fpsyg.2022.941084

Office of the Education Council. (2023). *Education in Thailand 2022.* Ministry of Education. https://backoffice.onec.go.th/uploads/Book/2057-file.pdf

Office of the Higher Education Commission. (2024). *Announcement of the Higher Education Standards Committee: Policy on raising English language standards in higher education institutions 2024.* https://www.ops.go.th/th/e-book/edu-standard/download/3253/9625/16

Piamsai, C. (2023). Development and use of CEFR based self-assessment in a Thai tertiary context. *PASAA, 66*(1), 81–126. https://doi.org/10.58837/CHULA.PASAA.66.1.3

Proietti, G. S., Matteucci, M., & Mignani, S. (2020). Automated test assembly for large-scale standardized assessment: Practical issues and possible solutions. *Psych, 2*(4), 315–337. https://doi.org/10.3390/psych2040024

Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Research and Practices in Technology*

Enhanced Learning, 15, Article 12. https://doi.org/10.1186/s41039-020-00134-8

Rafatbakhsh, E., Ahmadi, A., Moloodi, A., & Mehrpour, S. (2020). Development and validation of an automatic item generation system for English idioms. *Educational Measurement: Issues and Practice, 40*(2), 49–59. https://doi.org/10.1111/emip.12401

Rausch, A., Seifried, J., Wuttke, E., Kögler, K., & Brandt, S. (2016). Reliability and validity of a computer-based assessment of cognitive and non-cognitive facets of problem-solving competence in the business domain. *Empirical Research in Vocational Education and Training, 8*, Article 9. https://doi.org/10.1186/s40461-016-0035-y

Rice, N., Pêgo, J. M., Collares, C. F., Kisielewska, J., & Gale, T. (2022). The development and implementation of a computer adaptive progress test across European countries. *Computers and Education: Artificial Intelligence, 3*, Article 100083. https://doi.org/10.1016/j.caeai.2022.100083

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*(2), 49–60.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice, 10*(3), 279–293. https://doi.org/10.1080/0969594032000148145

Schnoor, B., Hartig, J., Klinger, T., Naumann, A., & Usanova, I. (2023). Measuring the development of general language skills in English as a foreign language—Longitudinal invariance of the C-test. *Language Testing*, *40*(3), 796–819. https://doi.org/10.1177/02655322231159829

Song, Y., Du, J., & Zheng, Q. (2025). Automatic item generation for educational assessments: A systematic literature review. *Interactive Learning Environments*, *33*(9), 5386–5405. https://doi.org/10.1080/10494820.2025.2482588

Soper, D.S. (2025). *A-priori sample size calculator for structural equation models* [Software]. https://www.danielsoper.com/statcalc

Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, *9*, Article 2245. https://doi.org/10.3389/fpsyg.2018.02245

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times.* Jossey-Bass/Wiley.

UNESCO. (2023). Global Education Monitoring Report 2023: Technology in education – A tool on whose terms? UNESCO. https://doi.org/10.54676/UZQV8501

Van Wijk, E. V., Donkers, J., De Laat, P. C. J., Meiboom, A. A., Jacobs, B., Ravesloot, J. H., Tio, R. A., Van Der Vleuten, C. P. M., Langers, A. M. J., & Bremers, A. J. A. (2024). Computer adaptive vs. non-adaptive medical progress testing: Feasibility, test performance, and student experiences. *Perspectives on Medical Education*, *13*(1), 406–416. https://doi.org/10.5334/pme.1345

Waluyo, B., Zahabi, A., & Ruangsung, L. (2024). Language assessment at a Thai university: A CEFR-based test of English proficiency development. *rEFLections*, *31*(1) 25–47. https://doi.org/10.61508/refl.v31i1.270418

Weiss, D. J. (2013). Item banking, test development, and test delivery. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 185–200). American Psychological Association. https://doi.org/10.1037/14047-010

Westacott, R., Badger, K., Kluth, D., Gurnell, M., Reed, M. W. R., & Sam, A. H. (2023). Automated Item Generation: Impact of item variants on performance and standard setting. *BMC Medical Education, 23*, Article 659. https://doi.org/10.1186/s12909-023-04457-0

Zhou, R., Samad, A., & Perinpasingam, T. (2024). A systematic review of cross-cultural communicative competence in EFL teaching: Insights from China. *Humanities and Social Sciences Communications, 11*, Article 1750. https://doi.org/10.1057/s41599-024-04071-5

Zhu, A., Mofreh, S. A. M., & Salem, S. (2023). The application of language proficiency scales in education context: A systematic literature review. *Sage Open*, *13*(3). 1–19. https://doi.org/10.1177/21582440231199692