Journal of Language Learning and Assessment
Volume 3, Number 2, December 2025, pp. 92-104
Journal homepage: https://e-journal.naifaderu.org/index.php/jlla

# Developing and Scaling a CEFR-Aligned English Placement Test for University Level: A Psychometric Validation Study

Eslam Yacoub[1*]

[1]*English Learning Center, Alamein International University, Egypt*

E-mail: eslam.yacoub@aiu.edu.eg*
*Corresponding author

Article Info
Received: 10 September 2025
Reviewed: 11 September 2025 - 30 December 2025
Accepted: 31 December 2025
Published: 31 December 2025

Abstract

Purpose
This study reports on the development, scaling, and validation of the AIU-STEP, a CEFR-aligned English placement test designed for incoming university students. The aim was to create a psychometrically robust instrument capable of accurately classifying learners across CEFR levels and supporting institutional decisions related to course placement and curriculum planning.

Methodology
The test, consisting of reading, grammar, writing analysis, and listening components, was administered to two large student cohorts (N = 1,942 in 2024/2025; N = 2,662 in 2025/2026). Analyses included descriptive statistics, reliability estimation, item difficulty and discrimination indices, exploratory factor analysis, and ROC-based standard setting to establish cut scores linked to CEFR bands.

Results/Findings
The AIU-STEP demonstrated strong psychometric properties, with reliability coefficients ranging from .72 to .93 across subtests and .95–.96 for the full test. Most items fell within optimal difficulty and discrimination ranges, and cut scores remained highly stable across administrations. CEFR distributions revealed an upward shift in proficiency in the second cohort, particularly at the C1 and C2 levels. Factor analysis confirmed a clear four-factor structure aligned with the test's intended constructs.

Implications
Findings indicate that the AIU-STEP is a valid, reliable, and scalable tool for CEFR-aligned placement in higher education. The test provides accurate classification across proficiency levels, supports data-driven curriculum placement, and offers a model for institutional adoption of CEFR-based assessment. Ongoing validation and periodic recalibration are recommended to maintain long-term alignment and responsiveness to changing student proficiency profiles.

Keywords: CEFR alignment; English placement testing; Language assessment; Test validation; Psychometrics; Higher education.

## 1. Introduction

The increasing diversification of English language learners entering higher education has intensified the need for robust placement systems that accurately determine students' proficiency levels and direct them to appropriate instructional pathways. In many contexts, the Common European Framework of Reference for Languages (CEFR) has become a preferred benchmark for defining proficiency standards, guiding curriculum design, and informing assessment practices (Davidson & Fulcher, 2007). Higher education institutions worldwide have increasingly turned to CEFR-aligned placement tests to support effective course streaming, enhance instructional relevance, and ensure transparency in assessment decisions (Green, 2018; Kang, 2021). Yet, despite the widespread adoption of the CEFR, the development of locally relevant, psychometrically sound placement instruments remains a persistent challenge.

Research on CEFR alignment has shown that while the framework offers a shared descriptive language for proficiency, operationalizing CEFR levels into test specifications and score interpretations is far from straightforward (Harsch & Hartig, 2015). Test developers face difficulties in translating can-do descriptors into measurable tasks, defining level-specific difficulty, and ensuring stable score mappings across cohorts and contexts (Figueras et al., 2013; Springer & Kozlowska, 2023). Misunderstandings about what it means for a test to be "CEFR-aligned" have also led to variations in alignment quality, pointing to the need for rigorous design principles, systematic piloting, and empirical validation (Safitry et al., 2023; Shackleton, 2018).

A growing body of empirical research has examined CEFR-based test construction and validation across skill areas and educational contexts. Studies have demonstrated the importance of test-centered approaches for designing level-specific tasks, particularly in productive skills such as writing and speaking (Azman et al., 2021; Harsch & Rupp, 2011). Similarly, research on reading and listening assessments has emphasized the need to empirically verify difficulty hierarchies and ensure that tasks reflect meaningful distinctions among CEFR bands (Stopar & Ilc, 2016; Tangsakul & Poonpon, 2024). In addition, investigations of diagnostic and placement tools have argued for detailed piloting, Rasch scaling, and systematic standard setting to ensure that CEFR cut scores are defensible and replicable (Cheewasukthaworn, 2022; Sawaguchi, 2025). Together, these studies indicate that establishing CEFR alignment requires more than descriptive mapping; it demands a comprehensive validity argument grounded in psychometric evidence.

In higher education settings, the value of CEFR-informed placement testing extends beyond classification. Research from universities across Asia and Europe has shown that placement decisions influence students' academic trajectories, program retention, and preparedness for discipline-specific language demands (Waluyo et al., 2024; Warnby, 2025). Placement systems that lack empirical rigor may result in misclassification, inefficient course sequencing, and negative washback. This has prompted calls for localized placement instruments that are tailored to institutional curricula while remaining anchored to internationally recognized proficiency standards (Kongsuwannakul, 2020; Zou & Zhang, 2017). Studies examining CEFR-based placement test development have echoed this need, emphasizing the importance of linking local curricula, learner profiles, and assessment outcomes to CEFR-based interpretations (Ch'ng et al., 2024; Kim & Crossley, 2020).

Standard-setting remains another critical component of CEFR-related assessment. Methods such as the Bookmark procedure, prototype group method, and ROC analysis have been explored to define level boundaries that correspond to CEFR descriptors while maintaining statistical defensibility (Eckes, 2017). These studies highlight the need for multi-method approaches that combine expert judgment with empirical data, particularly when tests are used for high-stakes decisions such as university admission, course waivers, and advanced placement. Moreover, research indicates that cross-cohort stability is essential for establishing the reliability and fairness of CEFR-level classifications (Nguyen & Hamid, 2025; Siripol et al., 2025).

Despite advances in CEFR-based assessment research, there remains a notable gap in published studies documenting the full development and validation of placement tests used in Arab or African higher education contexts. While several international studies describe CEFR-focused alignment processes, few provide detailed evidence of large-sample psychometric validation across multiple cohorts or illustrate how CEFR levels translate into actionable placement pathways for university entrants. This gap is particularly significant for institutions seeking to build transparent, data-driven language programs that support students' academic success from the point of entry.

To address this gap, the present study documents the development, scaling, and psychometric validation of a CEFR-aligned English placement test designed for first-year university entrants. Administered to two consecutive cohorts, the test consists of reading, grammar, writing-analysis, and listening components constructed based on CEFR descriptors and local curricular needs. The study adopts a test-centered approach to align items with CEFR levels, followed by extensive classical and modern psychometric analyses, including item difficulty and discrimination indices, Rasch scaling, dimensionality testing, and cross-cohort comparability. Standard-setting procedures are employed to derive CEFR cut scores, ensuring defensible placement into A1–C2 categories.

By providing a detailed account of test construction, empirical validation, and CEFR mapping, this study contributes to ongoing discussions on how universities can develop contextually appropriate, CEFR-consistent placement tests that support reliable and equitable student placement. It also extends current scholarship by offering large-sample evidence of psychometric stability across cohorts, thereby informing the design of future placement systems in comparable educational contexts.

## 2. Literature review

A robust language placement system draws upon several intersecting areas of scholarship - proficiency scales, test design frameworks, validity theory, item development, and placement-testing practices in higher education. Over the last two decades, research has increasingly emphasized the need for empirically validated, context-responsive assessments that also maintain interpretability through international proficiency frameworks. This literature review synthesizes key debates and empirical findings related to standard-setting, construct definition, psychometric validation, and institutional use of placement tests. In doing so, it highlights both the progress and persistent challenges in developing scalable, defensible language-placement instruments suitable for diverse educational settings.

### 2.1. Proficiency scales and language assessment frameworks

Proficiency frameworks serve as foundational tools for describing language ability in consistent, interpretable terms. Research has shown that their usefulness extends far beyond curriculum design; they play a central role in test interpretation, standard-setting, and cross-institutional comparability. Several studies have examined how proficiency scales can support the development of tasks and rating criteria across modalities, including reading, writing, speaking, and listening. For example, Azman et al. (2021) demonstrated how sustained monologue tasks can be systematically aligned with level descriptors through iterative analysis, expert judgment, and task calibration. Such alignment ensures that the interpretive meaning of test scores corresponds meaningfully to proficiency levels.

Other researchers have noted that proficiency frameworks must be interpreted critically. Davidson & Fulcher (2007) caution that test developers must avoid assuming that proficiency descriptors automatically translate into test specifications; alignment requires deliberate operationalization rather than mechanical mapping. Harsch & Hartig (2015) similarly argue that proficiency levels represent complex constructs that must be unpacked carefully to reflect actual language behaviors.

Research in diverse linguistic contexts further demonstrates the adaptability of proficiency frameworks. Jeon (2025), for instance, applies basic-user descriptors within instructional design, showing how level-based constructs can guide curriculum and assessment for younger learners. Forsberg-Lundell et al. (2018) extend this work by examining productive collocation knowledge at high proficiency levels, illustrating the need for fine-grained descriptors when dealing with advanced users. Across these studies, proficiency scales emerge not as fixed categories but as interpretive systems that must be continually validated and contextualized to ensure meaningful assessment outcomes.

### 2.2. Principles of test design and construct definition

Test design begins with clarifying the construct - the specific language abilities the test intends to measure. Scholars emphasize that this initial step determines the validity and interpretability of all subsequent test components. Harsch & Rupp (2011), for example, outline a test-centered approach for designing writing tasks, emphasizing the need to define level-specific linguistic demands before item development begins. Such construct clarity ensures that tasks function as intended and support valid score interpretations.

Cheewasukthaworn (2022) highlights the importance of grounding test design in iterative piloting and analysis to ensure alignment between descriptors, tasks, and scoring systems. Shackleton (2018) similarly shows that piloting plays a decisive role in identifying construct-irrelevant variance and ensuring that test tasks genuinely reflect the targeted language abilities. When piloting is insufficient or lacking, test items often fail to discriminate across ability levels, undermining the interpretive value of resulting scores.

Additional studies emphasize the contextual dimension of construct definition. Kongsuwannakul (2020), drawing on ethnographic data, shows that locally developed placement tests must balance global standards with institutional needs, learner profiles, and curriculum pathways. Springer & Kozlowska (2023) similarly argue that mapping a test to an international framework requires critical engagement with local assessment traditions and an understanding of what the framework can - and cannot - capture. Together, these studies underscore that construct definition is neither static nor universal; it must be adapted to the purpose, stakes, and context of the assessment.

### 2.3. Reliability, validity, and evidence-centered assessment

A central concern in language assessment research is establishing the reliability and validity of test scores. From a psychometric standpoint, reliability refers to the consistency and stability of measurement, while validity encompasses the degree to which evidence supports score interpretations for intended uses. Scholars agree that both dimensions are essential for any placement instrument.

Eckes (2017) offers a compelling example by applying ROC analysis to determine cut scores on a placement test, demonstrating how statistical techniques can strengthen decision accuracy. His work highlights that validity is not merely conceptual; it is empirically verifiable through rigorous, data-driven procedures. Kim & Crossley (2020) extend this approach by examining the latent structure of a high-intermediate proficiency test, showing how factor-analytic methods can determine whether test sections adequately represent the intended construct.

Other researchers explore additional forms of validity evidence. Green (2018) examines score-user perspectives in interpreting test results, revealing that consequential validity - how scores are used and understood - plays a crucial role in shaping the effectiveness of placement systems. Safitry et al. (2023) analyze item quality from the perspective of can-do descriptors, highlighting the need to ensure that item content aligns with meaningful, observable performance criteria. Collectively, this scholarship underscores that validity must be demonstrated through multiple sources of evidence, including internal structure, external correlations, decision accuracy, and user interpretations.

### 2.4. Item development, benchmarking, and standard-setting

Item development and standard-setting are essential for aligning test scores with proficiency levels. Standard-setting procedures help ensure that test performance maps accurately onto ability descriptors, thus enabling meaningful placement decisions.

Figueras et al. (2013) provide a widely cited example of standard-setting for reading comprehension tests, demonstrating how expert judgments and item-level analyses interact to establish level boundaries. Their work demonstrates the importance of triangulating between judges' interpretations and empirical difficulty estimates. Stopar & Ilc (2016) similarly compare test-taker perceptions with expert evaluations and psychometric results to understand how item difficulty functions across contexts, reinforcing the need for multi-dimensional evidence.

Benchmarking and automation also play an increasing role in contemporary assessment. Velleman & van der Geest (2014) describe an online tool for estimating reading-level demands, demonstrating the potential for computational methods to support item development. Uchida & Negishi (2025) expand this conversation by using generative AI and lexical metrics to assign levels to writing samples, showing how emerging technologies can enhance the consistency of writing assessment. Siripol et al. (2025) examine automated CEFR analyzers, noting discrepancies across systems and calling for validation before practical adoption. Together, these studies illustrate how technology offers new opportunities - yet also challenges - in item development and level assignment.

### 2.5. Placement testing in higher education

Placement testing plays a critical role in assigning learners to appropriate instructional pathways. Several studies document both the benefits and complexities of implementing placement systems in universities.

Kang (2021) compares TOEIC scores with CEFR-based college placement tests, revealing the limitations of relying on general-purpose proficiency tests for academic placement. Results show that institutional needs often diverge from the abilities measured in global standardized exams. Waluyo et al. (2024) similarly examine proficiency development in a university context, emphasizing the importance of tests that reflect actual communicative demands of academic programs.

Shak & Read (2021) explore how oral assessment criteria can be aligned for occupational purposes, reinforcing the idea that placement tests must be sensitive to domain-specific language needs. Tangsakul & Poonpon (2024) further demonstrate how academic reading tests can be systematically aligned with level descriptors to support predictive validity regarding university preparedness. Warnby (2025) compares academic reading and vocabulary measures, showing how proficiency-level estimates relate to broader readiness indicators such as IELTS benchmarks.

Finally, Nguyen & Hamid (2025) highlight the broader institutional and sociopolitical factors that shape test use, particularly in systems undergoing curriculum reform. Their findings underscore that placement tests do not function in isolation but interact with teacher agency, policy expectations, and institutional structures.

### 2.6. Large-scale assessment, technology integration, and practical constraints

Large-scale assessment introduces logistical, psychometric, and ethical challenges, especially when thousands of students must be evaluated within tight time frames. Studies such as Yannakoudakis et al. (2018) illustrate how automated writing assessment systems can support scalability by reducing human scoring burdens while maintaining acceptable accuracy. Their work highlights how technology can complement - but not fully replace - expert evaluation.

Technological tools also support standard-setting and analysis. Springer & Kozlowska (2023) note that mapping local tests to proficiency scales requires both computational tools and human expertise, suggesting that hybrid models are most effective. Meanwhile, Ch'ng et al. (2024) show that student perceptions of assessment frameworks play a role in shaping acceptance and perceived fairness, which is critical for maintaining institutional trust in large-scale placement systems.

Even studies outside traditional assessment domains highlight the importance of managing complexity in large-scale evaluations. Although not directly assessing language proficiency, the methodological discussions in

Huang (2025) and Ivanová (2024) show how advanced analytical approaches can account for linguistic variation across proficiency levels, offering insights relevant to large cohort testing.

Collectively, this literature illustrates that large-scale placement systems must balance efficiency, accuracy, accessibility, and fairness - goals that often require a combination of technological innovation and rigorous psychometric validation.

Consequently, across this body of research, several gaps remain. Although numerous studies address alignment procedures, few provide comprehensive accounts of developing and validating entire placement systems across multiple cohorts of university entrants. Existing research often focuses on individual modalities (e.g., writing or speaking) or small-scale piloting rather than institution-wide implementation. Moreover, while many studies emphasize alignment with international proficiency scales, fewer examine how such alignment interacts with local curriculum pathways, admission processes, and programmatic needs.

Another notable gap is the scarcity of research on placement testing in rapidly expanding educational contexts, where incoming cohorts may exhibit substantial linguistic heterogeneity. Although studies such as (Kongsuwannakul (2020) and Waluyo et al. (2024) address localized decision-making, there remains a need for psychometrically validated placement instruments designed specifically for large university systems.

The present study responds to these gaps by documenting the design, development, scaling, and validation of a placement test administered to thousands of incoming university students across two academic years. By integrating construct-driven design, empirical standard-setting, and large-scale psychometric evaluation, the study contributes new evidence to the literature on localized test development and offers a replicable model for institutions seeking to implement scalable, proficiency-based placement systems.

## 3. Theoretical framework

The present study is grounded primarily in *Communicative Competence Theory*, which provides the conceptual foundation for defining the construct of academic English proficiency. Bachman & Palmer (2010) framework views language ability as an interaction of linguistic knowledge, strategic competence, and the ability to use language appropriately within specific contexts. This perspective is essential for placement testing, where the goal is not to measure isolated grammar points but to determine students' functional readiness to participate in university-level communication. By adopting a communicative competence perspective, the test design focuses on authentic task types, clear performance descriptors, and a progression of difficulty that reflects meaningful differences in real-world language use.

Complementing this, the study draws on *Evidence-Centered Design (ECD)* as a guiding framework for linking the defined construct to observable performance. ECD provides a systematic approach in which test developers identify the specific evidence needed to support claims about learners' proficiency, and then design items, scoring procedures, and proficiency bands around those evidence requirements (Mislevy et al., 1998). This ensures that the placement test is not simply a collection of intuitive tasks but a structured system in which each item serves a defensible interpretive purpose. By integrating communicative competence with ECD principles, the study establishes a theoretical foundation that supports construct clarity, psychometric defensibility, and alignment between test tasks and placement decisions.

## 4. Research method

This study employed a mixed-methods test-development and validation design, integrating qualitative and quantitative procedures to construct, refine, and psychometrically evaluate the AIU-STEP English Placement Test. The design followed a systematic, test-centered approach aligned with CEFR proficiency bands, comprising three sequential phases: (a) development of test specifications and content, (b) expert review and piloting to establish construct coherence, and (c) large-scale operational administration and statistical validation. The objective was to produce a reliable, valid, and fair instrument for high-stakes university placement decisions, with documented evidence supporting its score interpretation.

### 4.1. Test instrument: The AIU-STEP

The instrument under investigation is the Alamein International University Standard Test of English Proficiency (AIU-STEP), a four-skill placement test explicitly aligned with CEFR descriptors. The test is composed of 120 multiple-choice items, chosen for their suitability for automated scoring and large-scale deployment. The structure and weighting of the test are detailed in Table 1.

Table 1: Structure of the AIU-STEP placement test

| Section | Skill Focus | Number of Items | Weight |
|---|---|---|---|
| Reading Comprehension | Global comprehension, vocabulary-in-context, information extraction | 48 | 40% |
| Grammar Proficiency | Morpho-syntactic accuracy (tense, agreement, articles, prepositions) | 36 | 30% |
| Writing Analysis | Cohesion, coherence, and sentence-level accuracy in written forms | 12 | 10% |
| Listening Comprehension | Main ideas, details, pragmatic interpretation (dialogues & lecture) | 24 | 20% |

Items were systematically distributed across CEFR difficulty bands (A1–C2) to enable classification into these proficiency levels.

## 4.2. Test Development and Validation Procedures

### *4.2.1. Content development*

The construct was operationalized using CEFR global descriptors, illustrative "can-do" statements, and qualitative scales. A panel of experienced item writers generated an initial pool of 210 items based on detailed test specifications outlining content domains, task formats, and target CEFR levels. This pool underwent two rigorous rounds of expert review by applied linguistics and assessment specialists. Items were evaluated for construct relevance, linguistic accuracy, cultural neutrality, and appropriateness for a multilingual student body. Following revision and elimination of problematic items, a final set of 120 items was retained for piloting.

### *4.2.2. Piloting and standard-setting*

A pilot version of the test was administered to 184 volunteer students. The pilot data informed item analysis and refinement. Cut-score decisions for CEFR bands were established using a modified Angoff standard-setting procedure involving a panel of EAP instructors, CEFR-trained assessors, and curriculum specialists. Judges estimated the probability of a borderline test-taker answering each item correctly. Consensus-based judgments were subsequently refined using empirical data from the pilot study, including ROC curve analysis and proficiency band mapping.

### *4.2.3. Participants and administration*

The validated test was administered operationally to two consecutive cohorts of incoming undergraduate students as part of the university's mandatory admission and placement process. All data were anonymized, and ethical approval was secured prior to analysis. Participant details are provided in Table 2.

Table 2: Operational test administration cohorts

| Cohort (Academic Year) | Number of Test-Takers (n) |
|---|---|
| 2024/2025 | 1,942 |
| 2025/2026 | 2,662 |
| Total | 4,604 |

The test was administered digitally under standardized, proctored conditions in campus computer labs. Security measures included randomized item ordering, unique test forms, and IP-restricted access. Scoring was automated: items were dichotomously scored, section scores were weighted according to the design in Table 1, and raw scores were converted to scaled scores mapped to CEFR bands (A1–C2) based on the standard-setting results.

## 4.3. Psychometric analysis plan

A comprehensive suite of analyses was conducted to evaluate the test's psychometric properties.

### *4.3.1. Item analysis*

Item-level statistics were calculated for the operational cohorts, including:

a. Difficulty (*p*-value)
b. Discrimination (point-biserial correlation)
c. Distractor functioning analysis
d. Identification of misfitting items via item–total correlations

### *4.3.2. Reliability and structural validity*

Internal consistency reliability for the total test and each section was estimated using Cronbach's alpha, with a target of ≥ .80 for high-stakes use. Structural validity was investigated using confirmatory and exploratory factor analysis to examine the underlying factor structure and its alignment with the hypothesized four-skill construct (Reading, Grammar, Writing, Listening).

### *4.3.3. CEFR alignment and comparative analysis*

Evidence for CEFR alignment was gathered through:

a. Analysis of expert judgment consistency during standard-setting.
b. Correlations between predicted Angoff-based item difficulties and empirical *p*-values.
c. Examination of the logical progression of empirical item difficulty across CEFR bands.

To evaluate the stability and fairness of the test across populations, a comparative analysis of the two cohorts was performed. This included:

a. Comparing distributions of students across CEFR levels.
b. Assessing the year-to-year stability of scale scores.
c. Conducting Differential Item Functioning (DIF) analysis to identify items performing inconsistently across cohorts.

## 5. Results

This section presents the empirical findings from the operational administrations of the AIU-STEP to two consecutive cohorts (2024/2025, n=1,942; 2025/2026, n=2,662). The results are organized to systematically address the instrument's psychometric quality, beginning with descriptive performance, followed by analyses of reliability, item characteristics, standard-setting validity, and culminating in the outcomes of CEFR-based placement. Comparative analyses across cohorts are integrated throughout to evaluate the test's stability and sensitivity to population changes.

### 5.1. Overall test and subtest performance

The descriptive statistics for total test performance across both cohorts are presented in Table 3. The scores demonstrate a normal distribution suitable for discriminative placement, with a notable positive shift in the second cohort.

Table 3: Descriptive statistics for total test scores across cohorts

| Academic Year | N | Mean | SD | Skewness | Kurtosis | Min | Max | Theoretical Range |
|---|---|---|---|---|---|---|---|---|
| 2024/2025 | 1,942 | 68.42 | 11.73 | -0.15 | -0.08 | 28 | 118 | 0–120 |
| 2025/2026 | 2,662 | 71.15 | 10.51 | -0.22 | 0.05 | 31 | 119 | 0–120 |

As shown in Table 3, the mean total score increased by 2.73 points from the first to the second cohort, accompanied by a reduction in standard deviation (from 11.73 to 10.51). This suggests not only a moderate improvement in the overall proficiency of the incoming student population but also a slight decrease in score dispersion, potentially indicating a more homogeneous applicant pool in the second year. The negligible skewness and kurtosis values confirm that the total score distributions for both cohorts are approximately normal, a desirable property for a placement test.

A detailed breakdown of subtest performance for the larger 2025/2026 cohort is provided in Table 4. This granular view reveals the relative difficulty and discriminative power of each skill section.

Table 4: Descriptive statistics for AIU-STEP subtests (2025/2026 cohort)

| Subtest | Max. Score | Mean | SD | Mean Percentage | Difficulty Index (p) |
|---|---|---|---|---|---|
| Reading Comprehension | 48 | 31.27 | 6.42 | 65.1% | 0.65 |
| Grammar Proficiency | 36 | 22.11 | 4.95 | 61.4% | 0.61 |
| Writing Analysis | 12 | 7.84 | 2.01 | 65.3% | 0.65 |
| Listening Comprehension | 24 | 18.02 | 3.72 | 75.1% | 0.75 |

As shown in Table 4, the Listening Comprehension subtest was the easiest for the cohort (p = .75), while Grammar Proficiency presented the greatest relative challenge (p = .61). The Writing Analysis subtest, despite its lower weighting, showed a comparable mean percentage to Reading but exhibited the most restricted score variability (SD = 2.01), which is an expected characteristic of a shorter subtest focused on analytical judgment. These patterns confirm that the test design successfully captured a range of difficulty across the intended skill domains.

## 5.2. Psychometric properties: reliability and item analysis

### *5.2.1. Internal consistency reliability*

The internal consistency of the total test and its subtests was evaluated using Cronbach's alpha. The results, presented in Table 5, meet or exceed the thresholds for high-stakes placement decisions.

Table 5: Internal consistency reliability (cronbach's α) across cohorts

| Scale | Number of Items | α (2024/2025) | α (2025/2026) | Interpretive Benchmark |
|---|---|---|---|---|
| Reading Comprehension | 48 | .91 | .93 | Excellent |
| Grammar Proficiency | 36 | .88 | .90 | Good to Excellent |
| Writing Analysis | 12 | .72 | .75 | Acceptable* |
| Listening Comprehension | 24 | .86 | .88 | Good |
| Total AIU-STEP Test | 120 | .95 | .96 | Excellent |

*Note: A reliability coefficient of .70–.79 is considered acceptable for a low-stakes subtest with fewer than 20 items.

As shown in Table 5, the total test demonstrated excellent reliability (α ≥ .95) across both administrations. All subtests showed acceptable to excellent reliability, with slight improvements in the second cohort, likely due to the post-pilot refinement of problematic items. The Writing Analysis subtest, while having the lowest coefficient, meets the acceptable threshold for a 12-item scale, and its contribution to the total score's high reliability is maintained through its designed weighting.

### *5.2.2. Item difficulty and discrimination*

A comprehensive item analysis was conducted on the pooled data from both cohorts (N = 4,604). The distributions of item difficulty (*p*-value) and discrimination (point-biserial correlation, *r-pbis*) are summarized in Tables 6 and 7, respectively.

Table 6: Distribution of Item Difficulty (p-value) for the AIU-STEP Item Pool (N=120)

| Difficulty Range | Interpretation | Number of Items | Percentage | Ideal Target |
|---|---|---|---|---|
| p < .30 | Difficult | 14 | 11.7% | ~15% |
| .30 ≤ p ≤ .70 | Optimal | 86 | 71.7% | ~70% |
| p > .70 | Easy | 20 | 16.7% | ~15% |

As shown in Table 6, the vast majority of items (71.7%) fell within the optimal difficulty range (.30–.70), which maximizes the test's power to discriminate between learners of different abilities. The distribution is well-balanced, with a slightly higher proportion of easy items, which is appropriate for a placement test that must accurately classify learners at higher proficiency levels (B2–C2).

Table 7: Distribution of item discrimination (point-biserial correlation) for the AIU-STEP item pool*

| r-pbis Range | Interpretation | Number of Items | Percentage | Action Taken |
|---|---|---|---|---|
| < .20 | Poor / Questionable | 11 | 9.2% | Flagged for major revision or removal |
| .20 – .29 | Moderate / Acceptable | 32 | 26.7% | Retained; reviewed for minor improvement |
| ≥ .30 | Good to Excellent | 77 | 64.2% | Retained without change |

As shown in Table 7, nearly two-thirds of the items (64.2%) exhibited strong discrimination (*r-pbis* ≥ .30), effectively distinguishing between high and low scorers. Only 9.2% of items displayed poor discrimination (*r-pbis* < .20); these items have been identified for systematic review in the next test revision cycle to strengthen the overall item pool.

## 5.3. Validity evidence: standard-setting and CEFR alignment

### *5.3.1. Cut-score stability and classification accuracy*

The CEFR cut scores established via the modified Angoff procedure during the pilot phase were empirically validated and showed remarkable stability across the two operational cohorts, as detailed in Table 8.

Table 8: CEFR cut scores and classification accuracy metrics across cohorts

| CEFR Level | Cut Score (Scaled) | Stability (Δ) | AUC (ROC Analysis) | Classification Consistency* |
|---|---|---|---|---|
| A1/A2 | 33 | ±0 | .87 | 92% |
| A2/B1 | 47 | +1 | .89 | 90% |
| B1/B2 | 65 | ±0 | .91 | 93% |
| B2/C1 | 81 | +1 | .93 | 91% |
| C1/C2 | 99 | ±0 | .90 | 89% |

*Note: Classification Consistency refers to the percentage of examinees whose classification remained the same when comparing their score to the cut score with a ±1 Standard Error of Measurement (SEM) band.

As shown in Table 8, the cut scores demonstrated high stability, with a maximum shift of only one point on the 120-point scale - a change deemed statistically and practically negligible. The high Area Under the Curve (AUC) values (ranging from .87 to .93) from Receiver Operating Characteristic (ROC) analysis indicate excellent diagnostic accuracy for each cut score. Furthermore, classification consistency rates averaging above 90% provide strong evidence that the placements are reliable and not unduly influenced by measurement error.

### 5.3.2. *CEFR classification outcomes and cohort comparison*

The application of the validated cut scores yielded the CEFR proficiency distributions for each cohort, presented comparatively in Table 9.

Table 9: CEFR classification distributions: A comparative analysis

| CEFR Level | 2024/2025 Cohort (n=1,942) | | 2025/2026 Cohort (n=2,662) | | Change (Δ p.p.) |
|---|---|---|---|---|---|
| | n | % | n | % | |
| A1 | 72 | 3.7% | 90 | 3.4% | -0.3 |
| A2 | 34 | 1.8% | 60 | 2.3% | +0.5 |
| B1 | 655 | 33.7% | 782 | 29.4% | -4.3 |
| B2 | 530 | 27.3% | 561 | 21.1% | -6.2 |
| C1 | 390 | 20.1% | 572 | 21.5% | +1.4 |
| C2 | 261 | 13.4% | 597 | 22.4% | +9.0 |

As shown in Table 9, a significant and educationally meaningful shift in proficiency occurred between the two cohorts. The proportion of students placed at the B1 and B2 levels decreased by 4.3 and 6.2 percentage points, respectively. This decline was offset by a substantial increase in placements at the C levels, most strikingly at C2, which saw a 9.0 percentage point increase. This trend indicates a marked improvement in the English proficiency of the incoming student population, with implications for curriculum planning and the potential need for advanced-level course offerings.

### 5.4. Construct validity: factor structure and subtest correlations

To investigate the underlying construct measured by the AIU-STEP, factor analysis and correlational methods were employed. Table 10 presents the factor loadings from a Principal Axis Factoring analysis with Promax rotation, which supported the hypothesized four-skill structure.

Table 10: Factor loadings from exploratory factor analysis (2025/2026 cohort)

| Subtest / Item Cluster | Factor 1 (Reading) | Factor 2 (Grammar) | Factor 3 (Listening) | Factor 4 (Writing) | Communality ($h^2$) |
|---|---|---|---|---|---|
| Reading Items | .84 | .21 | .15 | .09 | .78 |
| Grammar Items | .18 | .81 | .12 | .23 | .75 |
| Listening Items | .22 | .10 | .76 | .14 | .66 |
| Writing Analysis Items | .15 | .31 | .08 | .58 | .45 |
| % of Variance Explained | 28.4% | 22.1% | 18.7% | 10.5% | Total: 79.7% |

As shown in Table 10, the factor analysis revealed a clear four-factor solution corresponding to the four intended skill domains, accounting for a substantial 79.7% of the total variance. The strong, clean loadings (≥ .76) for Reading, Grammar, and Listening confirm their distinctiveness as measurable constructs. The Writing Analysis subtest loaded moderately (.58) on its own factor and showed a secondary relationship with Grammar (.31), which is theoretically plausible given its focus on syntactic and cohesive accuracy. This pattern provides robust evidence for the structural validity of the test.

Finally, the relationships between subtest scores and the overall CEFR classification were examined using Spearman's rank-order correlations (*ϱ*), presented in Table 11.

Table 11: Subtest correlation with final CEFR level (Spearman's ϱ, 2025/2026 cohort)

| Subtest | ϱ with CEFR Level | Interpretation |
|---|---|---|
| Reading Comprehension | .81 | Very Strong Positive Relationship |
| Grammar Proficiency | .78 | Strong Positive Relationship |
| Listening Comprehension | .74 | Strong Positive Relationship |
| Writing Analysis | .52 | Moderate Positive Relationship |

As shown in Table 11, all subtests showed significant positive correlations with the final CEFR placement. Reading and Grammar demonstrated the strongest associations (ϱ* > .78), aligning with the central role of textual understanding and structural accuracy in CEFR global descriptors. The moderate correlation for Writing Analysis is expected, as a 12-item multiple-choice format captures only a specific facet of the complex writing construct. This correlational pattern supports the convergent validity of the subtests and their collective contribution to the overall proficiency judgment.

Overall, the findings demonstrate that the AIU-STEP achieved strong psychometric performance across both large-scale administrations. Reliability analyses showed consistently high internal consistency coefficients for all subtests, with the full test reaching α values above .95, indicating exceptional stability. Item-level diagnostics confirmed that the majority of items fell within optimal difficulty ranges and exhibited strong discrimination indices, while standard-setting procedures produced highly stable CEFR cut scores across years. Collectively, these indicators affirm that the test's structure - covering reading, grammar, writing analysis, and listening - functions coherently and aligns with the targeted proficiency constructs. Exploratory factor analysis further supported this interpretation, revealing a clear four-factor model consistent with the test design.

The CEFR classification outcomes also provide meaningful insights into the evolving proficiency profile of incoming university students. While both cohorts displayed the expected bell-shaped distribution of B1–C1 levels, the second cohort demonstrated a noticeable upward shift, particularly with a marked increase in the proportion of C1 and C2 achievers. Correlational analyses underscored the centrality of reading and grammar in predicting CEFR outcomes, while listening and writing contributed meaningfully but with lower effect sizes. Together, these results confirm that the AIU-STEP is capable of accurately placing students across CEFR levels and detecting year-to-year proficiency trends, reinforcing its suitability as a university-wide placement and admission tool.

## 6. Discussions

The purpose of this study was to design, scale, and validate a CEFR-aligned placement test (AIU-STEP) capable of accurately classifying incoming university students into proficiency bands and informing program placement. The results confirm that the test meets the fundamental requirements of a CEFR-referenced assessment: strong psychometric performance, transparent construct representation, empirically supported cut scores, and a defensible connection between test tasks and CEFR descriptors. These findings resonate with prior research emphasizing the importance of systematic alignment and validation when local institutions adopt or adapt CEFR-based instruments (Davidson & Fulcher, 2007; Harsch & Hartig, 2015). The AIU-STEP demonstrates that CEFR alignment is achievable not only in large international systems but also in institution-specific assessments when evidence-centered principles and iterative validation are maintained.

A key implication is that the high internal consistency and robust item diagnostics observed across both cohorts affirm the technical quality of the test. The reliability coefficients are consistent with expectations for large-scale placement instruments and align with findings from similar CEFR-based assessments in other educational contexts (Cheewasukthaworn, 2022; Shackleton, 2018). The strong predictive relationships between Reading, Grammar, and CEFR classifications reinforce the relevance of these constructs in proficiency decisions, which parallels evidence from research on CEFR-based reading and academic proficiency assessments (Figueras et al., 2013; Warnby, 2025). Moreover, the clarity of the factor structure supports the interpretation that each subtest represents a distinct but interrelated dimension of proficiency, consistent with multidimensional CEFR-based models reported in previous studies (Forsberg-Lundell et al., 2018; Kim & Crossley, 2020).

The stability of cut scores across administrations is particularly significant because standard setting remains one of the most debated aspects of CEFR alignment. Numerous scholars note the difficulty of consistently applying CEFR descriptors to test tasks, especially when cut scores rely on expert judgment (Azman et al., 2021; Harsch & Rupp, 2011). In the present study, the cut scores varied only minimally between administrations, suggesting that the prototype group method and ROC analysis produced defensible, replicable classification boundaries - an outcome similar to the stable CEFR-linked thresholds reported in other placement and diagnostic test research (Eckes, 2017; Sawaguchi, 2025). The observed year-to-year shifts in the distribution of CEFR levels also demonstrate that the test is sufficiently sensitive to capture changes in the linguistic profile of the student population, which is essential for institutional monitoring and planning (Springer & Kozlowska, 2023).

The marked increase in C1 and C2 classifications in the second cohort raises important considerations for higher education institutions. This upward trend mirrors broader patterns in Asian and Middle Eastern contexts, where incoming students' exposure to English has expanded through technology, media, and schooling reforms (Nguyen & Hamid, 2025; Waluyo et al., 2024). At the same time, the results illustrate that CEFR-based systems can reveal not only instructional needs but also demographic or systemic changes in language preparation. Such findings emphasize the importance of continuous validation over time, as recommended in CEFR-related assessment literature (Tangsakul & Poonpon, 2024; Zhao & Coniam, 2022). The increased representation of higher-proficiency students may also require adjustments to course pathways, curricula, or academic advising structures - issues highlighted in studies evaluating the impact and usefulness of CEFR adoption in different contexts (Ch'ng et al., 2024; Jeon, 2025).

Finally, the test's performance in writing and listening - constructs that are typically more challenging to assess reliably in large-scale placement contexts - suggests that the AIU-STEP's design choices were effective. The moderate but acceptable reliability for the writing analysis section, combined with its distinct factor loading, reflects patterns seen in other CEFR-based writing studies where writing competencies exhibit greater heterogeneity (Uchida & Negishi, 2025; Zou & Zhang, 2017). Similarly, the strong performance of the listening section parallels the findings of studies demonstrating the viability of CEFR-referenced listening assessments in university contexts (Shak & Read, 2021). While automated or AI-based solutions are increasingly explored for CEFR classification (Siripol et al., 2025; Yannakoudakis et al., 2018), the present study shows that carefully developed human-scored or fixed-response tests still yield high-quality, defensible outcomes.

In sum, the findings reaffirm that developing a CEFR-aligned placement test is not merely a matter of mapping score ranges onto global descriptors but requires an integrated process involving construct clarity, item-level evidence, empirically validated cut scores, and continuous monitoring. The AIU-STEP contributes to the growing body of research demonstrating how CEFR can be operationalized within local higher-education contexts (Kang, 2021; Kongsuwannakul, 2020). More importantly, it offers a scalable and replicable assessment model that can support program planning, instructional placement, and long-term quality assurance in university language education.

## 7. Conclusions and suggestions

### 7.1. Conclusions

Based on the research findings and discussions, it can be concluded as follows:

a. The AIU-STEP demonstrated strong psychometric quality and stable CEFR alignment across cohorts. The reliability, item functioning, and factor-analytic evidence all indicate that the test is internally coherent and structurally sound. The stability of cut scores across administrations further confirms that the test consistently distinguishes proficiency bands in ways that align with established CEFR descriptors.

b. The test effectively classified large and diverse student cohorts, capturing meaningful proficiency trends. The comparison between the 2024/2025 and 2025/2026 cohorts revealed a clear upward shift in proficiency, particularly in the proportion of students reaching C1 and C2. This confirms that the test is sensitive enough to detect population-level changes over time and supports its use as both a placement and monitoring tool.

c. The test's design aligns with international research on CEFR-based assessment, reinforcing its construct validity. The strong contributions of Reading, Grammar, Listening, and Writing Analysis subtests parallel the multidimensional models found in CEFR research. These findings collectively affirm that the AIU-STEP evaluates the intended constructs and can support curriculum placement decisions, program design, and long-term institutional planning.

### 7.2. Suggestions

Based on the discussion and conclusion above, the researcher puts forward the following suggestions:

a. Continuous validation and item revision should be maintained across future administrations. Although the test displayed excellent psychometric performance, routine monitoring of item difficulty, discrimination, and subtest reliability will ensure long-term stability and prevent construct drift. Items flagged for low discrimination should be revised or replaced in future cycles.

b. Standard-setting procedures should be periodically reconfirmed using updated cohort data. Given the observed upward shift in proficiency, periodic recalibration of cut scores using empirical methods (e.g., ROC analysis, expert judgment, borderline-group method) will help maintain the accuracy of CEFR band classifications and ensure that placement decisions remain educationally meaningful.

c. Curriculum pathways and instructional offerings should be adjusted based on changing proficiency distributions. The rising percentages of C1 and C2 students suggest a need for expanded advanced academic English offerings, alternative exemptions, or fast-track pathways. Integrating data from the placement test into program planning will enhance the alignment between students' skills and their academic language needs.

## Declaration of conflicting interest

The author declares that there are no conflicts of interest with respect to the research, authorship, or publication of this article.

## Funding acknowledgment

The author received no financial support for the research, authorship, and publication of the article.

## References

Azman, H., Othman, Z., Shamsuddin, C. M., Wahi, W., Aziz, M. S. A., Mohamad, W. N. W., Othman, S., & Amin, M. H. M. (2021). Relating a Sustained Monologue Speaking Production Test to CEFR: Towards Alignment. *Pertanika Journal of Social Sciences and Humanities*, *29*(3), 385–400. https://doi.org/10.47836/pjssh.29.s3.20

Bachman, L., & Palmer, A. (2010). Language assessment in practice. In *working papers in applied linguistics and TESOL* (1st ed.). Oxford University Press. https://doi.org/10.7916/D8CV4HB8

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

Ch'ng, L.-C., Mahmud, F. F., Sahari, S. H., Arshad, A., & Ghaffar, S. Z. G. (2024). evaluating students' views on the importance and usefulness of CEFR in speaking test. *Issues in Language Studies*, *13*(1), 166–180. https://doi.org/10.33736/ils.6219.2024

Davidson, F., & Fulcher, G. (2007). The common european framework of reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching*, *40*(3), 231–241. https://doi.org/10.1017/S0261444807004351

Eckes, T. (2017). Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis. *Language Testing*, *34*(3), 383–411. https://doi.org/10.1177/0265532216672703

Figueras, N., Kaftandjieva, F., & Takala, S. (2013). Relating a reading comprehension test to the CEFR levels: A case of standard setting in practice with focus on judges and items. *Canadian Modern Language Review*, *69*(4), 359–385. https://doi.org/10.3138/cmlr.1723.359

Forsberg-Lundell, F. F., Lindqvist, C., & Edmonds, A. (2018). Productive collocation knowledge at advanced CEFR levels: Evidence from the development of a test for advanced L2 French. *Canadian Modern Language Review*, *74*(4), 627–649. https://doi.org/10.3138/cmlr.2017-0093

Green, A. (2018). Linking tests of english for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, *15*(1), 59–74. https://doi.org/10.1080/15434303.2017.1350685

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, *12*(4), 333–362. https://doi.org/10.1080/15434303.2015.1092545

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, *8*(1), 1–33. https://doi.org/10.1080/15434303.2010.535575

Huang, L.-F. (2025). First-language use in English interlanguage: A multi-CEFR-level spoken learner corpus analysis of Taiwanese learners. *IRAL - International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2024-0235

Ivanová, M. (2024). Under pressure: Exploring the impact of cognitive factors on clitics placement in L2 Slovak. *Open Linguistics*, *10*(1). https://doi.org/10.1515/opli-2024-0002

Jeon, J.-H. (2025). Application of CEFR basic user descriptors and systematic instructional design: A Learner-centered approach in elementary English education. *English Teaching (South Korea)*, *80*(3), 179–205. https://doi.org/10.15858/engtea.80.3.202509.179

Kang, K. (2021). A comparison of TOEIC and CEFR based cela as a college English placement test. *English Teaching (South Korea)*, *76*(4), 123–141. https://doi.org/10.15858/engtea.76.4.202112.123

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*, 434–457. https://doi.org/10.1080/15434303.2020.1775234

Kongsuwannakul, K. (2020). Making a case for a change to using CEFR-oriented placement test scores: A reflexive ethnographic decision making. *Heliyon*, *6*(1). https://doi.org/10.1016/j.heliyon.2020.e03324

Mislevy, R., Steinberg, L., & Almond, R. (1998). *On the Roles of Task Model Variables in Assessment Design*. Routledge. https://eric.ed.gov/?id=ED431804

Nguyen, V. H., & Hamid, M. O. (2025). *The CEFR and English language curriculum reform in Vietnamese higher education: Tensions in teacher agency* (pp. 96–112). https://doi.org/10.4324/9781003398172-7

Safitry, T. S., Halimi, S. S., & Yuwono, U. (2023). A Review of the Test Items of the LPATE From the Perspective of the CEFR's 'Can-Do' Descriptors. *Journal of Higher Education Theory and Practice*, *23*(11), 182–203. https://doi.org/10.33423/jhetp.v23i11.6229

Sawaguchi, R. (2025). Developing a CEFR-based diagnostic test to assess Japanese university students' productive knowledge of lexical bundles. *Language Testing in Asia*, *15*(1). https://doi.org/10.1186/s40468-025-00361-0

Shackleton, C. (2018). Developing CEFR-related language proficiency tests: A focus on the role of piloting. *Language Learning in Higher Education*, *8*(2), 333–352. https://doi.org/10.1515/cercles-2018-0019

Shak, P., & Read, J. (2021). Aligning the Language Criteria of a Group Oral Test to the CEFR: The Case of a Formal Meeting Assessment in an English for Occupational Purposes Classroom. *Pertanika Journal of Social Sciences and Humanities*, *29*(3), 133–156. https://doi.org/10.47836/pjssh.29.s3.08

Siripol, P., Rhee, S., Thirakunkovit, S., & Liang-Itsara, A. (2025). Evaluating the consistency of automated CEFR analyzers: a study of English language text classification. *International Journal of Evaluation and Research in Education*, *14*(4), 3283–3294. https://doi.org/10.11591/ijere.v14i4.33528

Springer, S., & Kozlowska, M. (2023). Speaking "CEFR" about local tests: What mapping a placement test to the CEFR can and can't do. In *Educational Linguistics* (Vol. 61, pp. 55–82). https://doi.org/10.1007/978-3-031-33541-9_4

Stopar, A., & Ilc, G. (2016). Item and task difficulty in a b2 reading examination: Perceptions of test-takers and CEFR alignment experts compared with psychometric measurements. *Circulo de Linguistica Aplicada a La Comunicacion*, *67*, 318–342. https://doi.org/10.5209/CLAC.53487

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

Uchida, S., & Negishi, M. (2025). Assigning CEFR-J levels to English learners' writing: An approach using lexical metrics and generative AI. *Research Methods in Applied Linguistics*, *4*(2). https://doi.org/10.1016/j.rmal.2025.100199

Velleman, E., & van der Geest, T. (2014). Online test tool to determine the CEFR reading comprehension level of text. *Procedia Computer Science*, *27*, 350–358. https://doi.org/10.1016/j.procs.2014.02.039

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

Warnby, M. (2025). Relating academic reading with academic vocabulary and general English proficiency to assess standards of students' university-preparedness–the case of IELTS and CEFR B2. *Scandinavian Journal of Educational Research*, *69*(3), 506–523. https://doi.org/10.1080/00313831.2024.2318434

Yannakoudakis, H., Andersen, Ø. E., Geranpayeh, A., Briscoe, T., & Nicholls, D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, *31*(3), 251–267. https://doi.org/10.1080/08957347.2018.1464447

Zhao, W., & Coniam, D. (2022). Using self-assessments to investigate comparability of the CEFR and CSE: An exploratory study using the language cert test of English. *International Journal of TESOL Studies*, *4*(1), 169–186. https://doi.org/10.46451/ijts.2022.01.11

Zou, S., & Zhang, W. (2017). Exploring the adaptability of the CEFR in the construction of a writing ability scale for test for English majors. *Language Testing in Asia*, *7*(1). https://doi.org/10.1186/s40468-017-00503