SCRIPT JOURNAL
Journal of Linguistic and English Teaching
http://jurnal.fkip-uwgm.ac.id/index.php/Script
P-ISSN: 2477-1880; E-ISSN: 2502-6623
April 2026, Vol. 11 No. 1

Revisiting the Institutional English Proficiency Test (EPT): Evaluating Its Validity, Reliability, and CEFR Alignment

Moh Syafik[1], Eva Nikmatul Rabbianty[2,1]*, Yazid Basthomi[2], Ling Gan[3], Nurul Hadi[1]
[1]Universitas Islam Negeri Madura, Indonesia
[2]Universitas Negeri Malang, Indonesia
[3]Beijing Technology and Business University, China
* Email Correspondence: eva@iainmadura.ac.id

Abstract

Background:
There is an increasing need for higher education in Indonesia to develop an institutional English Proficiency Test, as internationally recognized high-stakes English Tests are often financially unaffordable and logistically inaccessible for many academic communities. Therefore, the Center for Language Development of UIN Madura developed the EPT (English Proficiency Test) as an institutional proficiency test used to measure individual English proficiency. However, its validity, reliability, and alignment with the CEFR have not yet been empirically established.

Methodology:
This study employed a quantitative approach within a validation framework. To examine content validity, four subject matter experts (SMEs) were invited to evaluate the alignment between test items and their intended objectives using the Item-Objective Congruence (IOC) method. In addition, the study assessed the alignment of the EPT with the Common European Framework of Reference for Languages (CEFR) by examining three key processes: specification, standardization, and standard setting.

Findings:
The results revealed that EPT demonstrated satisfactory CV, as indicated by IOC indices exceeding 0.75 across the three measured language skills. Internal consistency and stability were supported by reliability coefficients meeting the acceptable criteria. However, insufficient empirical evidence indicates any alignment between EPT and CEFR, as the administrator was found not to follow the three validation frameworks: what is assessed (specification), how performance is interpreted (standardization), and how comparison is made (standard setting).

Conclusion:
The results conclude that EPT should be improved by aligning it with CEFR. Administrators should conduct a standard-setting study to map EPT scores onto the CEFR and provide the minimum scores (cut scores) needed to enter each targeted CEFR level.

Originality:
This study addresses a specific gap in the literature by evaluating the CEFR alignment of an institutional English Proficiency Test (EPT) and exploring its relevance as an affordable proficiency-testing option in a specific Indonesian higher education context.

Keywords	:	English Proficiency Test; Item-Objective Congruence; reliability; validity
DOI	:	10.24903/sj.v11i1.2331
Received	:	February 2026
Accepted	:	April 2026
Published	:	April 2026
How to cite this article (APA)	:	Syafik, M., Rabbianty, E. N., Basthomi, Y., Gan, L., & Hadi, N. (2026). Revisiting the institutional English proficiency test (EPT): Evaluating its validity, reliability, and CEFR alignment. *Script Journal: Journal of Linguistic and English Teaching*, *11*(1), 172-199. https://doi.org/10.24903/sj.v11i1.2331

1. INTRODUCTION

High-stakes English proficiency testing refers to standardized assessments designed to measure an individual's ability to use English for academic or professional purposes, with outcomes that have significant consequences for the test taker (Ockey & Gokturk, 2019). Globally, these tests are widely used as objective benchmarks for language competence in both academic and professional settings. In higher education, the results of these tests are often mandatory for university admissions, determining eligibility for graduation, awarding scholarships, and enabling international study and work opportunities. Examples of widely recognized high-stakes English tests include the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL), both of which are accepted by thousands of institutions worldwide, including in Indonesia (Dang & Dang, 2021; Pearson, 2021; Pearson, 2023; Saadatara et al., 2023; Sari & Mualimin, 2021). However, alongside these international tests, many higher-education institutions also develop and use institutional English proficiency tests for practical academic purposes. These tests may support students in evaluating their language proficiency (Dimova, 2017; Wibowo et al., 2024), classroom-based assessment practices (Jung et al., 2019; Mahesar & Jokhio, 2021; Mohseni, 2021), and preparing students for international language examinations (Lu & Hu, 2022; Solovjeva & Baksheev, 2021). Because institutional English proficiency tests increasingly serve academic assessment and administrative functions, their design and implementation need careful attention in higher education contexts.

Many Indonesian higher education institutions have also increasingly developed their own English proficiency tests to address local academic program and administrative needs. These institutional tests are widely used to meet graduation requirements, support scholarship selection, and, in some cases, serve employment-related purposes, particularly when access to international high-stakes tests is limited or unaffordable. Notable examples include the English Language Proficiency Test (ELPT) developed by Universitas Airlangga (Pusat Bahasa UNAIR, 2025), the Test of English as a Foreign Language (TEFL) by Institut Teknologi Sepuluh Nopember (Global Language Center ITS, 2025), and the English Proficiency Institutional Test (EPIT) by UIN Sunan Ampel Surabaya (Pusat Pengembangan Bahasa UIN Sunan Ampel Surabaya, 2023). These tests serve large numbers of test-takers

who need proof of English proficiency for higher education and employment purposes. Given their widespread use and consequential role in academic and professional decisions, institutional English proficiency tests require evidence of validity, reliability, and alignment with external frameworks such as the CEFR to support meaningful score interpretation.

In Indonesia, access to international high-stakes English proficiency tests remains limited, reflecting structural inequalities in testing availability. For example, according to the British Council (2022), IELTS is offered in only 12 official venues with approximately 96 test sessions per month. In contrast, TOEFL ITP is available at just 54 venues, with around 2 sessions per month, and TOEFL iBT is offered at only 5 venues, with 5 sessions per month (International Test Center, 2022). Due to the growing demand for English certification, the limited number of available testing venues and sessions creates unequal access to international testing. As a result, institutional English proficiency tests serve as a clear, practical, and accessible alternative for those needing proof of English proficiency for academic or professional reasons.

Besides the limited availability of the official TOEFL and IELTS test venues, the cost of taking these international tests also remains a major barrier for many Indonesian test takers. Official data indicate that the fees are approximately 200 USD for IELTS (British Council, 2022), 39 USD for TOEFL ITP (IIEF, 2022), and 230 USD for TOEFL IBT (International Test Center, 2022). For many Indonesians, these costs can be financially prohibitive, especially when compared with the average annual net income of Indonesian casual workers, which is estimated at around USD 2,000 (BPS, 2022), in a country classified as middle-income (UNESCO Institute for Lifelong Learning, 2021). In response to these affordability constraints, institutional English proficiency tests developed by Indonesian higher education institutions offer a practical alternative, typically charging lower fees of approximately USD 10–13 and providing more flexible scheduling and testing venues. In East Java, examples include ELPT at UNAIR, TEFL at ITS, EPIT at UINSA, TEP at UNESA, TEP at UIN Maliki, and EPT at UM.

Furthermore, in addition to the above reasons, not all test takers require the level of proficiency or difficulty represented by high-stakes tests such as IELTS and TOEFL, as language assessment should be aligned with local needs, strategic goals, and specific purposes (Chuang & Yan, 2025; Göktürk & Alaca, 2026). This principle of fitness for purpose provides a strong rationale for developing institutional English proficiency tests,

such as the EPT, which are designed to reflect specific academic contexts and intended uses (Hille & Cho, 2020). Additionally, institutional English proficiency tests can provide valuable information for evaluating teaching and learning outcomes, which in turn can inform instructional practices and policy decisions within educational settings (Nordström et al., 2019; Schildkamp et al., 2020; Swiecki et al., 2022). As a key assessment activity, language tests have proven significant for student achievement, teacher instruction, curriculum design, educational interventions, pedagogical initiatives, and other policies related to language development (Pearson, 2020).

In addition to considering the required proficiency level, some test takers—particularly those from rural areas—face challenges related to the time and financial resources needed to travel to test centers. Although some providers have made efforts to improve accessibility, individuals with mobility limitations may still be burdened by the requirement to be physically present at test centers (Isbell & Kremmel, 2020). To address this challenge, UIN Madura—situated in the geographic center of Madura Island and facing limited access to official international testing centers—developed an institutional English Proficiency Test (EPT) to serve both internal and external users. Since its inception in 2016, the EPT at the Center of Language Development, UIN Madura, has been administered to over 2,692 undergraduate and 1,481 postgraduate students, in addition to more than 500 public participants from diverse institutions. Despite the widespread use of the EPT and the external acceptance of its score certificates, there is still limited empirical evidence supporting its validity, reliability, and alignment with external frameworks such as the CEFR. Therefore, a systematic evaluation that specifically addresses these aspects is essential to ensure the appropriate use of results for both internal decision-making and external partnerships (Choi et al., 2023; Renandya et al., 2018; Reynolds et al., 2016; Zhu et al., 2023).

This study focuses specifically on content validity as one dimension of validity evidence in language assessment (Roy et al., 2023). To avoid dispersion across multiple validity domains, it only measures the content validity to ensure a concentrated, academically rigorous investigation. The study further examines the extent to which the test is equivalent to the proficiency scales in major proficiency frameworks for learning, teaching, and foreign-language skills assessment. The researchers adopt the CEFR (Common European Framework of Reference) in this context. The EPT is structurally modeled on the TOEFL® ITP, organized by ETS, which uses the CEFR to map test scores (ETS Global, 2022). CEFR

alignment is examined by evaluating whether the EPT development and score interpretation processes reflect key linking principles, including test specification (what is assessed), standardization (how performance is interpreted), and standard setting (how comparisons are made), as outlined by North (2014). This approach allows for a systematic assessment of whether the EPT demonstrates alignment with CEFR levels.

Numerous studies have investigated institutionally developed English proficiency tests and their alignment with recognized international standards. For instance, Madya et al. (2019) examined the equivalence of the Test of English Proficiency (TOEP), developed by the Indonesian Testing Service Center. They concluded that the test ensures a fair level of difficulty across test-takers (Harsch & Seyferth, 2020), aligns level-specific writing tasks with the CEFR, and that the level-specific approach effectively addresses task difficulty, rating consistency, and learner variation. Similarly, Alderson et al. (2006) and Natova (2021) evaluated the CEFR grids in designing reading and listening assessments and concluded they were useful but required refinement. Additionally, Hulstijn et al. (2012) assessed Dutch learners' speaking proficiency in relation to CEFR scales and identified inconsistencies between lexical knowledge and CEFR levels. Additionally, Sridhanyarat et al. (2021) validated the STEP (Srinakharinwirot University Test of English Proficiency) against the CEFR and confirmed its reliability and alignment with the CEFR.

Despite these studies, there is a noticeable lack of research on the validity and alignment with the CEFR of English proficiency tests developed by Indonesian institutions. While previous research has largely focused on equivalence or alignment to international frameworks, few have explored the test's contextual development within regional or Islamic university contexts. Therefore, the present study contributes to institutional language testing research by providing empirical evidence on the validity, reliability, and CEFR alignment of an institutionally developed English Proficiency Test developed by the Center of Language Development at Universitas Islam Negeri Madura. It demonstrates how an institutionally developed test can be systematically validated and aligned with international proficiency frameworks. Unlike previous studies, this research emphasizes not only affordability and accessibility but also the contextual validity of a localized EPT designed for non-urban Indonesian learners within a religious higher educational context.

The current study aims to answer two research questions:

1. Is the EPT developed by UIN Madura valid and reliable for assessing English proficiency?

2. To what extent does the EPT align with CEFR proficiency levels?

2. METHODOLOGY

2.1. Research Design

This study adopts a validation-focused design to establish validity evidence, reliability, and CEFR-aligned scoring for the English Proficiency Test (EPT) developed by the Center of Language Development at UIN Madura. In language testing, a validation-focused design treats validation as a continuous process rather than a one-time procedure, integrating it throughout the test development cycle to justify the interpretation and use of test scores (Chapelle & Voss, 2013). While consequential perspectives in language testing emphasize the impact of test use in educational and social contexts (Jin, 2022), the present study focuses on establishing measurement quality and alignment with international frameworks as prerequisites for such investigations and to support future research on EPT development. In this study, two of the researchers were directly involved as test administrators and developers. To minimize potential bias associated with the dual role of the researchers, the study relied on objective and standardized quantitative procedures, including predefined scoring criteria, statistical validity and reliability tests, and external expert judgments through Item-Objective Congruence (IOC) to evaluate content validity.

Furthermore, in this context, the researchers were not only passively involved but also actively intervened or modified the test. However, all the data reported in this study were derived exclusively from the original version of the EPT, prior to any modifications. Any modifications to the test were made only after data analysis was completed and were not included in the study results; rather, they served as a basis for subsequent test development. This approach enabled the researchers to be involved in the design and implementation of the intervention (Penuel et al., 2007) and evaluate those interventions in the design (Abma, 2005). This study adopted a collaborative and practice-oriented research design in which test development and evaluation were closely integrated (Walski, 2014). The findings are expected to provide a foundation for future research and continued refinement of the EPT at the Center for Language Development of UIN Madura.

2.2. The Data Collection Procedures

The data were obtained from the back-end test results of 590 test-takers from 2021 and 399 from 2022. These data were collected from the regular English Proficiency Test (EPT)

administered at UIN Madura (Previously IAIN Madura). The EPT consists of 70 items assessing listening comprehension (25 questions), structure and written expression (20 questions), and reading comprehension (25 questions). The test was developed systematically over 2021 and 2022, with items refined to better align with high-stakes tests such as TOEFL® ITP. These years were selected because the test design was more structured and aligned with established international proficiency standards. Although the data were collected from 2021-2022, they remain relevant for the 2025 context, as the core components of the test, including test blueprint, item specification, scoring procedures, tests' structure and content, and mode of administration, have remained unchanged. This consistency ensures that the findings are applicable in understanding the EPT's validity and its alignment with CEFR levels.

A random sample of 100 test-taker score records was selected from the EPT administration database for the 2021 and 2022 testing periods. All 989 records from those years formed the sampling frame. A simple random sampling procedure was used to ensure each record had an equal chance of selection. The same 100 records were used for all validity and reliability analyses, as research indicates that samples of this size provide stable reliability coefficients. Previous research suggests that samples of at least 100 participants can provide meaningful and stable reliability coefficients (Kline, as quoted by Kennedy, 2021). Furthermore, based on Fisher's as quoted by Kennedy (2022) research, a sample size of 100 ensures that the r value falls within the 95% confidence interval for reliability estimates (0.73-0.87). Since these data were secondary (previously collected as part of routine test administration), no additional invitations were sent to participants. All test-takers had previously consented to use their test data for institutional purposes, including research.

2.3. The Data Analysis

Content Validity. Firstly, this study assessed the content validity (CV). Content Validity refers to the extent to which an instrument consists of appropriate items for the underlying constructs intended to be measured. It is usually assessed based on related literature and experts' judgments (Gregori-Giralt & Menéndez-Varela, 2021). The CV was selected as the minimum quality requirement in developing the instruments at the item development stage (Almohanna et al., 2022; Halek et al., 2017). The procedure used to determine content validity was based on the Item-Objective Congruent (IOC) index developed by Rovinelli & Hambleton (1977). This statistical method was used in the test development to provide evidence of content validity at the item development stage. An

evaluation using the IOC index was a process in which content experts rated individual items based on the degree to which they measured the specific objectives prepared in the test developers' blueprints (Turner & Carlson, 2003). The IOC calculation used in this context was not the original calculation by Rovinelli and Hambleton, as it was limited to the assessment of unidimensional items. Instead, a mathematical extension of IOC developed by Turner & Carlson (2003) was used, as it was applicable to the "possibility" of the existence of a multidimensional case in the test, such as one item possibly measuring multiple combinations of skills in the EPT. The steps for the CV evaluation of the EPT by using IOC are as follows:

Expert Judgment. To calculate the index of IOC, a group of subject matter experts (SMEs) was involved. The experts were purposively selected based on their relevant academic backgrounds, professional experience in English language teaching, diverse geographical regions, and expertise in language assessment and test development. Four representative experts in the subject area were invited to judge the appropriateness of the test and the accuracy of the test items. The number of representative experts falls within the commonly recommended range of four experts for content validity assessment using IOC, as suggested by Rovinelli & Hambleton (1977) and Hambleton et al. (1978). The logic is that with four experts, the probability of agreement purely by chance is low. If 3 out of 4 agree, the IOC is 0.75, which exceeds the common psychometric cut-off of 0.50 or 0.70 for item inclusion. Below are brief demographic details for the SMEs.

Table 1. Descriptions of the subject matter experts (SMEs)

Demographic information of SME (Variable)		**N**
Affiliation	UIN Madura	2
	Universitas Negeri Malang	1
	University of Queensland	1
Qualification	Senior lecturer of Reading Comprehension	1
	Lecture of TOEFL Preparation class	1
	Doctor of English Teaching and Learning	1
	M.A. in TESOL	1
Teaching and Language Testing Experience	20-29 years	1
	10-19 years	0
	0-9 years	3
Gender	Male	2
	Female	2

Preparing the IOC Rating Sheet. After inviting the SME, the IOC rating sheet was prepared. The test developers provided the set of items and specific objectives to be measured. To keep the evaluators independent, the experts were not told which construct of the individual items was intended to measure. Then, the researchers created a table where each item was put in a "row", and the list of possible objectives was placed in a "column". This list was then distributed to the content experts, who would rate each item for each objective using a prepared rubric. An SME would evaluate each item by rating it with 1 (for a valid objective), -1 (invalid objective), and 0 (the degree to which it measures the content area is unclear).

From the rating of the SME, it was found that the test items measured more than one objective. Therefore, the multidimensional item formula, as adjusted by Crocker and Algina as cited by Turner and Carlson (2003) was applied. The equation for the adjusted index is as follows:

$$I'_{ik} = \frac{(N)\mu_k - (N-p)\mu_l}{2N - p}$$

$I'_{ik}I'_{ik}$ is the index of item-objective congruence for item ***i*** on a set of objectives ***k***

Where **N** represented the number of objectives pp represented the number of valid objectives. $\mu_k\mu_k$ symbolized the judges' mean rating of item ***i*** on a set of objectives ***k***. $\mu_i\mu_i$ indicated the judges' mean rating of the item ***i*** on a set of invalid objectives ***l***. If IOCt > 0.75, this indicates that the test items were valid (Turner & Carlson, 2003).

Reliability. The reliability of the test was then determined, and a sample of the 100 EPT results from the test takers was used as a pilot study. Alderson et al. (2006) state that piloting on suitable samples of test takers is essential to know whether a given item is at the intended difficulty level. Therefore, knowing the proficiency level of the test takers is necessary to judge the adequacy of the items. In this regard, the two types of reliability analysis were employed: (1) internal consistency, estimated using the Kuder–Richardson Formula 20 (KR-20) for dichotomously scored items, and (2) test–retest reliability to assess score stability over time. For the latter, a subset of the 100 test-takers completed the same EPT form a second time after an interval of 2 to 6 weeks, depending on their scheduled testing sessions. Both administrations were conducted under routine testing conditions. Stability was quantified using the Pearson correlation coefficient for total scores, with 95%

confidence intervals. The difficulty level and discrimination level were also tested to support the data.

3. FINDINGS

3.1. Validity and Reliability of EPT (English Proficiency Test)

The validity and reliability of the EPT were provided in two parts. The first part reported the content validity of each section by applying the IOC indices for multidimensional items. Then, the second part presented the reliability of the test by calculating the KR-20 and test-retest formula, which was completed with difficulty and discrimination of the test.

Validity of the Test

Table 2. IOC Indices of items for listening comprehension

Items	Objectives								*IOC*				Results
No.	D	IE	S	A	Pr	I	Pb	T	*p*	µk	µi	N	IOC	
1	**0.50**	-1.00	-1.00	-0.75	-1.00	-0.50	-1.00	-1.00	1	0.5	-0.89	8	0.68	Accepted
2	-1.00	-1.00	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
3	-1.00	-1.00	-1.00	-0.75	-1.00	1.00	-1.00	-1.00	1	1	-0.96	8	0.98	Accepted
4	-1.00	-0.75	-0.75	**1.00**	-0.50	-0.75	-1.00	-1.00	1	1	-0.82	8	0.92	Accepted
5	-1.00	-0.50	-1.00	**0.50**	-0.50	-0.75	-1.00	-1.00	1	0.5	-0.82	8	0.65	Accepted
6	**0.50**	-1.00	-1.00	-0.75	-1.00	-0.25	-1.00	-1.00	1	0.5	-0.86	8	0.67	Accepted
7	-1.00	-1.00	**1.00**	-1.00	-0.50	-1.00	-1.00	-1.00	1	1	-0.93	8	0.97	Accepted
8	**0.50**	**1.00**	-1.00	-1.00	-1.00	-0.75	-1.00	-1.00	2	0.75	-0.96	8	0.84	Accepted
9	-1.00	**0.50**	-1.00	-1.00	-0.25	-0.50	-1.00	-1.00	1	0.5	-0.82	8	0.65	Accepted
10	-1.00	-0.25	-1.00	**0.50**	-0.75	-0.50	-1.00	-1.00	1	0.5	-0.79	8	0.63	Accepted
11	**0.50**	-1.00	-1.00	-0.75	-1.00	-0.50	-1.00	-1.00	1	0.5	-0.89	8	0.68	Accepted
12	-0.50	-1.00	-1.00	-1.00	-1.00	-1.00	-0.75	**0.00**	1	0	-0.89	8	0.42	Unaccepted
13	-1.00	-1.00	**0.50**	-1.00	-1.00	-0.50	-0.25	-0.75	1	0.5	-0.79	8	0.63	Accepted
14	-0.75	-1.00	**1.00**	-1.00	-0.50	-1.00	-1.00	-0.50	1	0.75	-0.82	8	0.78	Accepted
15	-1.00	-1.00	-1.00	**0.75**	-1.00	-0.50	-1.00	-1.00	1	0.75	-0.93	8	0.83	Accepted
16	**0.50**	-1.00	-1.00	-0.75	-1.00	-0.50	-1.00	-1.00	1	0.5	-0.89	8	0.68	Accepted
17	-1.00	-1.00	-1.00	-1.00	-1.00	-0.50	-1.00	**0.50**	1	0.5	-0.93	8	0.70	Accepted
18	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.75	1	1	-0.96	8	0.98	Accepted
19	**1.00**	-1.00	-1.00	-1.00	-0.50	-1.00	-0.75	-1.00	1	1	-0.89	8	0.95	Accepted
20	**0.50**	-1.00	-1.00	-1.00	**0.25**	-1.00	-1.00	-1.00	2	0.375	-1.00	8	0.64	Accepted
21	**1.00**	-1.00	-0.75	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.96	8	0.98	Accepted
22	**0.75**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.50	1	0.75	-0.93	8	0.83	Accepted
23	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
24	-0.50	-1.00	-1.00	-1.00	-1.00	**0.00**	-0.50	-1.00	1	0	-0.86	8	0.40	Unaccepted
25	-0.50	-1.00	-1.00	-1.00	-0.50	-1.00	**0.00**	-1.00	1	0	-0.86	8	0.40	Unaccepted
IOCt													**0.76**	

Table 2 illustrates the IOC indices of item no. 1-25 for the listening comprehension section. The objectives are similar (with some modifications) to those the TOEFL® ITP measures, as 8 of the objectives focus on listening comprehension. In this analysis, the overall IOC or IOCt had a value of 0.76. IOCt > 0.75, indicating that the items for the listening comprehension test were valid (Turner & Carlson, 2003). That is, the items could measure the test takers' listening comprehension. In this analysis, three items (12, 24, and 25) had invalid IOC. The IOCs were less than 0.5. These items were further investigated following consultation with the SMEs. The experts either rated more than one objective or rated different objectives for each item, which was normal. However, since the IOCs still showed positive numbers, those items based on the expert's judgment were subject to modification.

Table 3. IOC Indices of items for structure and written expression

Items No.	Objectives															*IOC*					Results
	V	P	N	Adj.	Comp.	Prep.	Conj.	Adv.	Stc + Cl	Pov	Agr.	IVM	Parl.	Red.	Dict.	*p*	µk	µi	N	IOC	
26	-1.00	-1.00	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1.00	-1.00	15	1.000	Accepted
27	**1.00**	-1.00	-0.75	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1.00	-0.98	15	0.991	Accepted
28	-1.00	-1.00	-0.75	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1.00	-0.98	15	0.991	Accepted
29	-0.75	-1.00	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1.00	-0.98	15	0.991	Accepted
30	-0.25	-1.00	**0.50**	-1.00	-1.00	-1.00	-1.00	-1.00	**0.00**	-1.00	**0.00**	-1.00	-1.00	-1.00	-1.00	3	0.17	-0.94	15	0.509	Accepted
31	-1.00	-1.00	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	**0.50**	-1.00	-1.00	-1.00	-1.00	1	0.50	-0.96	15	0.724	Accepted
32	**0.50**	-1.00	-1.00	-1.00	-1.00	-0.50	-1.00	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.50	-0.93	15	0.707	Accepted
33	-0.50	-1.00	-0.50	-0.50	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	**0.50**	-1.00	-1.00	-1.00	-1.00	1	0.50	-0.89	15	0.690	Accepted
34	-1.00	-1.00	**0.50**	-0.50	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	1	0.50	-0.93	15	0.707	Accepted
35	-1.00	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	**1.00**	-1.00	-1.00	1	1.00	-0.96	15	0.983	Accepted
36	-0.50	-1.00	**0.75**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.75	-0.93	15	0.836	Accepted
37	**0.50**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	1	0.50	-0.96	15	0.724	Accepted
38	-0.50	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	**0.25**	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	1	0.25	-0.93	15	0.578	Accepted
39	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.50	-1.00	**0.50**	-1.00	-1.00	1	0.50	-0.96	15	0.724	Accepted
40	**1.00**	-1.00	-0.75	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	1	1.00	-0.95	15	0.974	Accepted
41	-1.00	-1.00	-1.00	-1.00	**0.75**	-0.50	-0.50	-1.00	-1.00	-0.50	-0.50	-1.00	-1.00	-1.00	-1.00	1	0.75	-0.86	15	0.802	Accepted
42	-1.00	-1.00	-1.00	**0.00**	-0.50	-1.00	-1.00	-1.00	-1.00	-0.50	-0.50	-1.00	-1.00	-1.00	-1.00	1	0.00	-0.89	15	0.431	Unaccepted
43	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	**0.50**	-0.50	-1.00	-0.50	-1.00	-0.50	-1.00	-0.50	1	0.50	-0.86	15	0.672	Accepted
44	-1.00	-1.00	-1.00	-0.50	-1.00	-1.00	**0.00**	-0.50	-1.00	-1.00	-0.50	-1.00	-1.00	-1.00	-1.00	1	0.00	-0.89	15	0.431	Unaccepted
45	-1.00	-1.00	**0.50**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.75	**0.00**	-1.00	-1.00	-1.00	-0.50	2	0.25	-0.94	15	0.571	Accepted
IOCt																				**0.752**	

Table 3 describes the IOC indices of item no. 26-45 for the structure and written expression section. There are 15 objectives measured. The finding indicated that IOCt for the structure and written expression section was valid (IOCt > 0.75). The analysis provided evidence of content validity, indicating that items 26–45 are appropriate measures of test-takers' proficiency in structure and written expression. Of 20 questions, there were only two items with IOC less than 0.5, items No. 42 and 44. To address these low scores, the researchers consulted with experts to determine whether the items should be modified or removed. From the analysis, the 4 experts rated different specific objectives for those two items. For example, in item no. 42, two experts rated it as measuring **Adjective (Adj.) or** Agreement (Agr.). The valid point for Adj. was 0.00, and the invalid objective for Agreement & Point of View was -0.50 (Not clear invalid objective -1). However, since the IOCs still demonstrated positive numbers, those items based on the experts' judgment were subject to modification.

Table 4. IOC Indices of items for reading comprehension

Items No.	Objectives								IOC					Results
	Prev.	MI	Voc.	Det.	Infer.	Exc.	Ref.	RtP	*p*	Mk	μi	N	IOC	
46	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
47	-0.75	-1.00	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	1	1	-0.96	8	0.98	Accepted
48	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	**0.50**	-0.50	1	0.5	-0.93	8	0.70	Accepted
49	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
50	**0.00**	-1.00	-1.00	-1.00	-1.00	-0.50	-1.00	-0.50	1	0	-0.86	8	0.40	Unaccepted
51	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
52	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
53	**0.50**	-0.50	-1.00	-1.00	-1.00	-0.50	-1.00	-0.50	1	0.5	-0.79	8	0.63	Accepted
54	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
55	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
56	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
57	-1.00	-1.00	-1.00	-1.00	-1.00	**1.00**	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
58	-0.50	-1.00	-1.00	**0.50**	-1.00	-1.00	-1.00	-1.00	1	0.5	-0.93	8	0.70	Accepted
59	-0.50	-1.00	-1.00	**0.50**	-1.00	-1.00	-1.00	-1.00	1	0.5	-0.93	8	0.70	Accepted
60	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1.00	8	1.00	Accepted
61	-0.50	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	1	1	-0.93	8	0.97	Accepted
62	-1.00	-1.00	**0.00**	**0.00**	-0.25	-1.00	-1.00	**0.00**	3	0	-0.85	8	0.33	Unaccepted
63	-1.00	**1.00**	-1.00	-1.00	-1.00	-0.50	-1.00	-1.00	1	1	-0.93	8	0.97	Accepted
64	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-0.50	1	1	-0.93	8	0.97	Accepted
65	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-0.50	1	1	-0.93	8	0.97	Accepted
66	-1.00	-1.00	-1.00	**0.50**	-1.00	-1.00	-1.00	-0.50	1	0.5	-0.93	8	0.70	Accepted
67	-1.00	-0.50	-1.00	-1.00	-1.00	**0.00**	-1.00	-1.00	1	1	-0.93	8	0.97	Accepted

68	-1.00	-1.00	**1.00**	-0.50	-1.00	-1.00	-1.00	-0.50	1	1	-0.86	8	0.93	Accepted
69	-1.00	-1.00	-1.00	**0.00**	-1.00	-1.00	-1.00	-0.50	1	0	-0.93	8	0.43	Unaccepted
70	-0.50	-1.00	-1.00	**0.50**	-0.75	-1.00	-1.00	-0.50	1	0.5	-0.82	8	0.65	Accepted
IOC_t													0.84	

As shown in Table 4, the data demonstrate acceptable overall content validity for reading comprehension, as indicated by an IOC coefficient of 0.84 (IOC > 0.75), the highest among the three measured skills. However, item-level analysis identified three items (nos. 50, 62, and 69) with IOC values below 0.50. These items were reviewed by experts and will subsequently be modified to improve alignment with the targeted reading comprehension objectives, consistent with standard IOC-guided content validation practices. It is important to note that IOC values below 0.50 are not uncommon when items are designed to assess multiple objectives, and eight objectives were represented across reading comprehension items.

3.2. Reliability of the Test

The presentation of the reliability of the EPT, followed by the level of its difficulty and the discrimination, are presented as follows:

Table 5. The reliability, the difficulty, and the discrimination of items for listening comprehension

Dimension	Methods	Indicators	Meanings
Reliability	Kuder-Richardson 20	$KR_{20} = 0.832$	The listening comprehension test is reliable at the significance level of 0.01 level (2-tailed).
	Test-Retest Reliability	$r_{12} = 0.897$	The listening comprehension test is reliable at the significance level of 0.01 level (2-tailed).
Difficulty	Classical Model	p = 0.62	The difficulty level of the items for listening comprehension is moderate level (Kelley, 1939; Vincent & Shanmugam, 2020)
Discrimination	Classical Model	d = 0.447	The discrimination of items for listening comprehension is very good (Ebel & Frisbie, 1991; Kumar, 2023)

Table 5 demonstrated that the items for listening comprehension were of high reliability, either in terms of internal consistency (KR_{20} = 0.832) or the stability of the instrument ($\mathbf{r}_{12}$ = 0.897). Also, it offered a statistically moderate level of difficulty (p = 0.62) and exhibited a very good index of discrimination (d = 0.447).

Table 6. The reliability, the difficulty, and the discrimination of Items for structure and written expression

Dimension	Methods	Indicators	Meanings
Reliability	Kuder-Richardson 20	$KR_{20} = 0.747$	The structure and written expression test are reliable at the significance level 0.01 level (2-tailed).
	Test-Retest Reliability	$r_{12} = 0.761$	The structure and written expression test are reliable at the significance level 0.01 level (2-tailed).
Difficulty	Classical Model	p = 0.50	The difficulty of the items for structure and written expressions is moderate level (Kelley, 1939; Vincent & Shanmugam, 2020)
Discrimination	Classical Model	d = 0.417	The discrimination of items for structure and written expressions

			is very good level (Ebel & Frisbie, 1991; Kumar, 2023)

Table 6 pointed out that the EPT items to the structure and written expressions had statistically significant reliability (KR_{20} = 0.747) and a significant stability ($\mathbf{r}_{12}$ = 0.761). Those numbers were above the acceptable value (0.70 to 0.95) (DeVellis & Thorpe, 2021). It also presented a moderate level of difficulty (p = 0.50) and exhibited a very good level of discrimination (classical model: d = 0.417).

Table 7. The reliability, the difficulty, and the discrimination of items for reading comprehension

Dimension	Methods	Indicators	Meanings
Reliability	Kuder-Richardson 20	KR_{20} = 0.755	The reading comprehension test is reliable at the significance level of 0.01 level (2-tailed).
	Test-Retest Reliability	$\mathbf{r}_{12}$ = 0.746	The reading comprehension test is reliable at the significance level of 0.01 level (2-tailed).
Difficulty	Classical Model	p = 0.45	The difficulty of the items for the reading comprehension test is moderate level (Kelley, 1939; Vincent & Shanmugam, 2020)
Discrimination	Classical Model	d = 0.378	The discrimination of the items for the reading comprehension test is reasonably good but subject to improvement (Ebel & Frisbie, 1991; Kumar, 2023)

Table 7 presented that the reliability of the reading comprehension had statistical significance, both in terms of internal consistency as shown by (KR_{20} = 0.755) and was reliable in terms of stability (Test-Retest Reliability, $\mathbf{r}_{12}$ = 0.746). The difficulty was moderate level (p = 0.45), and unfortunately, its discrimination level was just "reasonably good but subject to improvement" (d = 0.378). Although the index of discrimination was a secondary criterion for selecting items (Ebel & Frisbie, 1991; Kumar, 2023), the researchers consider it in improving the test's quality. Rather than evaluating discrimination item by item, the current revision strategy examines the section as a whole, using the overall discrimination index to inform holistic decisions to improve the reading comprehension component.

To sum up, based on the validity and reliability of testing, the EPT (English Proficiency Test) developed was statistically valid and reliable. The index of Item Objectives Congruence, as the formula used for the content validity of the items that measured the three skills, indicated satisfactory coefficients (all of the $IOC_t > 0.75$). Besides, the EPT was also reliable based on the statistical analysis. The reliability coefficients above the acceptable value (**r > 0.7**) indicate adequate internal consistency and suggest that the test scores are relatively stable.

3.3. Aligning the EPT to the CEFR levels

The EPT was analyzed and found to be valid and reliable. However, further investigation of the EPT instrument's construction is necessary, especially to link the test to an international framework such as the CEFR.

To align the EPT with CEFR levels. Firstly, the researchers analyzed how the EPT administrators "do" with the scores, and found that the administrators reflected the EPT scores through two steps. The first was a PASS/FAIL decision, indicating that the test takers performed satisfactorily on the exam. The minimum score to determine whether they were PASS or FAIL was 450. This number was ruled in accordance with the quality standard provided by the quality assurance system of UIN Madura (previously IAIN Madura) in 2019. The second was interpreting the score results. The EPT administrators interpreted the test takers' scores by leveling them into some categories. The table below shows the proficiency levels of the EPT scores.

Table 8. Proficiency Levels of EPT scores developed by the Center of Language Development of UIN Madura

No	EPT Scores	Proficiency Levels
1	310-420	Elementary
2	420-480	Low Intermediate
3	480-525	High Intermediate
4	525-677	Advanced

From the two steps above, the researchers stated that there was no alignment between the EPT and the CEFR. In deciding the PASS or FAIL decision, the test administrators used the score of 450 without a clear framework. In the CEFR, there are no specific numbers that determine whether test takers pass or fail. The CEFR gives descriptions of all kinds of scores, whether they are low or high. The CEFR then describes the scores on a six-point scale that represent all those achievements. Those six scales are A1, A2, B1, B2, C1, and C2 (Fleckenstein et al., 2020). It can be inferred that the concept of PASS or FAIL decision is not mentioned in the CEFR. The CEFR only provides the specific "threshold level" for each scale and regroups them into three general levels (A1 *& A2 as Basic User*s, B1 & B2 as *Independent User*s, and C1 & C2 as *Proficient Users*) (North, 2014; The CEFR Levels, 2022).

Furthermore, in dividing the proficiency levels of the EPT. The test administrators provided the specific scores and the proficiency levels. However, there is no clear framework for deciding those 4 score ranges and grouping them into four proficiency levels. At this

point, the researchers found no alignment with CEFR levels in deciding the *cut score* and categorizing the specific *proficiency levels*. The test administrator did not follow the commonly recognized procedures for linking language examinations, to the CEFR, as outlined by the Council of Europe in its manual (North, 2014), namely specification of the content of test and examinations (WHAT IS ASSESSED), standardization (HOW PERFORMANCE IS INTERPRETED) or stating criteria to determine the attainment of a learning objective, and standard setting (HOW COMPARISONS ARE MADE) for describing the levels of proficiency in existing tests and examinations, thus enabling comparisons to be made across different systems of qualification.

The findings indicate that the EPT does not demonstrate sufficient alignment with CEFR levels. First is in terms of *specification*. There was no documented expert judgment or systematic procedures to map EPT constructs and test content onto CEFR descriptors. Second, regarding *standardization*. Although performance criteria in the form of proficiency levels were provided by the test administrators, there was no evidence of a formal consensus process involving experts to relate the level of performance of the test and test tasks of the EPT to CEFR levels, nor were specific CEFR-based descriptors explicitly defined—the third concerns standard setting. The *cut scores* and *proficiency levels* appeared to have been determined internally without reference to established CEFR linking procedures, making it difficult to justify comparisons with CEFR levels.

4. DISCUSSION

In developing a language test, it is essential to ensure its validity and reliability and to assess its correspondence with the external framework. Alignment with external frameworks is necessary when test scores are intended to be interpreted or used beyond the local institutional context, as it supports comparability across settings and strengthens the defensibility of inferences drawn from test results in broader programs, jurisdictions, or evaluation contexts (Chen & Flasko, 2020; Folson & Awush, 2024). Test validity represents the extent to which a test accurately measures what it is intended to measure (Halek et al., 2017; Almanasreh et al., 2019), and test reliability indicates the extent to which the measurement occurs without error (Shirali et al., 2018; Ursachi et al., 2015). Furthermore, aligning the test with a common framework can enhance the interpretability and

comparability of test-takers' abilities, thus supporting their potential recognition across different contexts.

The first research objective is to analyze the validity and reliability of the EPT. Content validity was evaluated using the Item–Objective Congruence (IOC) method. The results indicate evidence of acceptable content validity for the EPT, with IOC values exceeding the recommended threshold. Specifically, the IOC_t was above 0.75, referring to more than the minimum cutoff value of the accepted value provided by Turner and Carlson (2003). This supports the same results of the previous IOC test for other language testing instruments, such as STEP (Sridhanyarat et al., 2021) and the Srilankan Reading Test (Ismail & Zubairi, 2021). The importance of establishing the validity of the EPT is mainly to provide evidence that the test measures the intended language constructs and that the test items are appropriately aligned with the targeted skills of listening, structure and written expressions, and reading comprehension. However, it is also important to note that the Item-Objective Congruence (IOC) index only reflects the degree of agreement between items and stated objectives based on expert judgment, rather than providing empirical evidence that each item fully represents all of language ability. Thus, IOC serves as a rigorous item-level check of content validity (Dhippayom et al., 2018). The reliability coefficients indicate that the EPT demonstrates acceptable internal consistency, meeting commonly recommended thresholds for reliability coefficients in educational measurement (DeVellis & Thorpe, 2021). This suggests that the EPT yields consistent scores across items and supports the assumption that the true score being measured remains stable over a short time interval (Shou, Sellbom, & Chen, 2022).

The research's second objective is to examine whether EPT aligns with the CEFR levels. In General, the EPT developed by test administrators has not yet undergone formal alignment process with the CEFR. First, regarding decision-making, although the EPT currently employs a pass/fail classification for local institutional use, CEFR-aligned interpretation is necessary for cross-context validity (Chen & Flasko, 2020). However, there is no evidence that the cut scores underlying these decisions are aligned with CEFR descriptors established through recognized linking procedures (Council of Europe, 2011). In this regard, the test administrator should design the test in relation to the CEFR. They can keep the PASS or FAIL decision, but ensure the minimum score is provided with careful consideration, a clear purpose, and a clear context. This is what North (2014) called an

"Illustrative Scale" that test administrators can develop and experiment with, using the CEFR descriptors to suit the context and objectives. This scale is particularly appropriate, as it allows test developers to adapt CEFR descriptors to local or institutional purposes while maintaining conceptual alignment with the international framework. This approach is especially suitable for the EPT as an institutional test designed for a specific population and context, where direct adoption of standardized CEFR may not be fully feasible. However, as mentioned earlier, using the illustrative scale does not mean "free-for-all for users to define as they wish" (North, 2014). The levels should always be applied responsibly, as EPT scores are used to be recognized for certain national or international purposes. Second, in dividing their own *cut scores* and proficiency levels, the test administrators created the EPT score ranges and proficiency levels with unclear standards. Test administrators, therefore, should perform the three possible ways in which the CEFR framework can be used (North, 2014) before deciding on their own *cut scores* and levels of proficiency.

The first step is to determine *what is being assessed* (test specification). Findings revealed that there was no documented expert judgment or systematic procedure for mapping EPT constructs and test content to CEFR descriptors. In this case, experts should be invited to evaluate the alignment between the EPT assessment content and the CEFR descriptor scales, taking into account the curriculum, assessment tasks, and criteria for judging success. The second step is to determine *how performance is interpreted* (test standardization). Although performance criteria in the form of proficiency levels were provided, there was no formal, consensus-based process involving experts to relate test performance and tasks to CEFR levels, nor were CEFR-based descriptors explicitly defined. Therefore, administrators should invite a panel of judges to reach consensus regarding EPT performance levels and test tasks associated with each CEFR level. The third step is to determine *how comparisons can be made* (standard setting). In this context, the EPT *cut scores* and *proficiency levels* appeared to have been determined internally without reference to established CEFR linking procedures, making it difficult to justify comparisons with CEFR levels. To address this, test administrators should determine the proficiency levels of EPT scores and set cut scores between levels in accordance with established CEFR linking procedures.

There are actually five stages in aligning certain tests with the CEFR; they are familiarisation, specification, standardization, standard setting, and validation (Council of Europe, 2011). However, in the EPT case, we do not conduct familiarization because EPT is

administered locally at our institution; we assume test takers are familiar with the questions. As for validation, the CV of this test validity has been proven by statistical analysis written in the first point of this finding. If the process of aligning the EPT with CEFR levels has been completed, the Center of Language Development of UIN Madura will likely find its EPT has specific scales, such as A1, A2, B1, B2, C1, and C2. The scales are then divided based on the proficiency levels (such as basic user, independent user, and proficient user) or probably using their previous terms (elementary, intermediate, and advanced). They are also expected to have their own *cut score* on each scale, and then, finally, they can formulate their own "can do" descriptors.

Previous studies have demonstrated that some language tests have been attempted or already aligned with the CEFR through systematic procedures, including the PTE Academic Test (De Jong et al., 2014), the Duolingo English Test (DET) (Coney & Isbell, 2024; Isaacs et al., 2023), and ETS assessments such as the TOEFL, TOEIC, and TOEIC Bridge Test (Tannenbaum & Wylie, 2008), and the EPT may also be aligned with the CEFR. The CEFR is not a rigid framework; it can serve as guidance on essential test development matters, such as test purpose, response format, time constraints, and topic (Weir, 2005). The CEFR, according to Harsch & Martin (2012), is context and language-independent, enabling test developers to add specific details to the descriptors to fit the purpose and context.

The evidence that there was no alignment between the EPT and CEFR frameworks was that the test's construct should be improved. The process of aligning the EPT with the CEFR should be made immediately to make the scores of EPT more acceptable, thus enabling comparisons to be made across different qualification systems. Providing the systematic alignment of EPT with CEFR will make the EPT more standardized, and the objectives will be clearer and measurable. It will make the EPT have a high degree of relevance for the ability to be measured and finally make the EPT become an alternative test for English proficiency, especially for Indonesian students and society, besides the high stakes such as TOEFL or IELTS.

The test developers should conduct a standard-setting study in the near future to map the EPT scores onto the CEFR, as Tannenbaum & Baron (2011) did for TOEFL® ITP. The result of that study may provide the minimum scores (cut scores) needed to enter each targeted CEFR level. Even though the content validity and reliability of the EPT have been established, it is necessary to test other types of validity, such as construct validity, criterion-

related validity, and face validity, to support the findings of this study. In addition, the test developers should modify items with invalid IOC indices and improve the level of discrimination, especially in reading comprehension. This may also limit the in-depth analysis of why the EPT developers do not use the CEFR as a framework. It is necessary to understand how test developers construct test items.

5. CONCLUSION

The findings revealed that the EPT achieved high content validity and reliability. A high IOC index indicated strong content validity across the three receptive skills; however, some items with low IOC indices need revision. The test also demonstrated reliability, with high internal consistency and stability, as indicated by high reliability coefficients. Regarding difficulty, the EPT was at a moderate level and exhibited good discrimination. Although the EPT (English Proficiency Test) demonstrates strong content validity and reliability, there is no evidence of alignment with the CEFR for decision-making standards, cut scores, or proficiency levels. Key CEFR-alignment procedures (specification, standardization, standard setting) were not documented, limiting interpretability and comparability. To enable both institutional and national use, the EPT must be rebuilt with CEFR mapping. The Center of Language Development at UIN Madura should conduct a formal standard-setting study, define scales and cut scores, and produce *can-do* descriptors aligned with the CEFR. As the current study examines only content validity, future researchers are advised to provide evidence for other validity factors and to conduct additional statistical analyses. Inviting more experts to judge item objectives is also necessary to achieve stronger validity evidence. For test developers, it is crucial to align the EPT with a standardized framework such as the CEFR, which is considered the most appropriate standard. This alignment will ensure that test results are more accountable, accurate, and transparent, and that the EPT can be standardized with common reference levels recognized institutionally, nationally, and internationally. Additionally, it is essential to compare the EPT with similar tests developed by other institutions to identify trends in EPT development across Indonesia.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Center of Language Development at UIN Madura for providing access to the English Proficiency Test (EPT) data.

We also thank the Subject Matter Experts who contributed to the content validation process through Item–Objective Congruence (IOC) evaluation. Our appreciation is also extended to all test-takers whose data made this study possible.

DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used Grammarly Premium to enhance language quality and scite.ai to identify recent, relevant scholarly literature, thus minimizing reliance on outdated sources. After using these tools, the authors thoroughly reviewed and edited all content as needed and take full responsibility for the accuracy, integrity, and originality of the final manuscript.

REFERENCES

Abma, T. A. (2005). Responsive evaluation: Its meaning and special contribution to health promotion. *Evaluation and Program Planning*, *28*(3), 279–289. https://doi.org/10.1016/j.evalprogplan.2005.04.003

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly*, *3*(1), 3–30. https://doi.org/10.1207/s15434311laq0301_2

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Almohanna, A. A. S., Win, K. T., Meedya, S., & Vlahu-Gjorgievska, E. (2022). Design and content validation of an instrument measuring user perception of the persuasive design principles in a breastfeeding mHealth app: A modified Delphi study. *International Journal of Medical Informatics*, *164*, 104789. https://doi.org/10.1016/j.ijmedinf.2022.104789

BPS. (2022). *Statistik Pendapatan*. Badan Pusat Statistik.

British Council. (2022). *Test Dates, Fees, and Locations | British Council Foundation Indonesia*. https://www.britishcouncilfoundation.id/en/exam/ielts/dates-fees-locations

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In *The Companion to Language Assessment* (pp. 1079–1097). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118411360.wbcla110

Chen, M. Y., & Flasko, J. J. (2020). Investigating the alignment between the CELPIP-general reading test and the Canadian language benchmarks: A content validation study.

Canadian Journal of Applied Linguistics, *23*(2). https://doi.org/10.37213/cjal.2020.30649

Choi, Y., Vo, S., & Ockey, G. J. (2023). Investigation into the factor structure of a local English placement test of oral communication. *International Journal of Applied Linguistics*, *34*(1). https://doi.org/10.1111/ijal.12478

Chuang, P.-L., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. *System*, *134*, 103846. https://doi.org/10.1016/j.system.2025.103846

Coney, N., & Isbell, D. R. (2024). *Where the lines are drawn: English language proficiency tests in international student admissions at U.S. research-intensive universities*. Open Science Framework. https://doi.org/10.31219/osf.io/tyn92

Council of Europe. (2011). *Manual for Language Test Development and Examining*. Council of Europe.

Dang, C. N., & Dang, T. N. Y. (2021). The predictive validity of the IELTS test and contribution of IELTS preparation courses to international students' subsequent academic study: Insights from Vietnamese international students in the UK. *RELC Journal*, 0033688220985533. https://doi.org/10.1177/0033688220985533

De Jong, J. H. A. L., Becker, K., Bolt, D., & Goodman, J. (2014). *Aligning PTE Academic test scores to the Common European Framework of Reference for Languages*. Pearson. https://pearsonpte.com/wp-content/ uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage Publications.

Dhippayom, J. P., Trevittaya, P., & Cheng, A. S. K. (2018). Cross-cultural adaptation, validity, and reliability of the patient-rated Michigan Hand Outcomes Questionnaire for Thai patients. *Occupational Therapy International*, *2018*. https://doi.org/10.1155/2018/8319875

Dimova, S. (2017). Life after oral English certification: The consequences of the Test of Oral English Proficiency for Academic Staff for EMI lecturers. *English for Specific Purposes*, *46*, 45–58. https://doi.org/10.1016/j.esp.2016.12.004

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th ed.). Prentice Hall of India.

ETS Global. (2022). *4 reasons why learning English is essential*. https://www.etsglobal.org/pl/en/blog/news/importance-of-learning-english

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43*, 100420. https://doi.org/10.1016/j.asw.2019.100420

Folson, D., & Awush, F. K. (2024). *Assessing Cognitive Alignment in Pre-tertiary TVET Core Mathematics: A Ghanaian case study of curriculum and exit examination*. In Review. https://doi.org/10.21203/rs.3.rs-5711559/v1

Global Language Center ITS. (2025). *TEFL ITS*. https://bahasa.its.ac.id/

Göktürk, N., & Alaca, S. (2026). Open Science in language assessment through the lens of government institutions. *Language Testing*, *43*(1), 79–93. https://doi.org/10.1177/02655322251352540

Gregori-Giralt, E., & Menéndez-Varela, J.-L. (2021). The content aspect of validity in a rubric-based assessment system for course syllabuses. *Studies in Educational Evaluation*, *68*, 100971. https://doi.org/10.1016/j.stueduc.2020.100971

Halek, M., Holle, D., & Bartholomeyczik, S. (2017). Development and evaluation of the content validity, practicability and feasibility of the Innovative dementia-oriented Assessment system for challenging behaviour in residents with dementia. *BMC Health Services Research*, *17*(1), 554. https://doi.org/10.1186/s12913-017-2469-8

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments. *Review of Educational Research*, *48*(1), 1–47. https://doi.org/10.3102/00346543048001001

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*(4), 228–250. https://doi.org/10.1016/j.asw.2012.06.003

Harsch, C., & Seyferth, S. (2020). Marrying achievement with proficiency – Developing and validating a local CEFR-based writing checklist. *Assessing Writing*, *43*, 100433. https://doi.org/10.1016/j.asw.2019.100433

Hille, K., & Cho, Y. (2020). Placement testing: One test, two tests, three tests? How many tests are sufficient? *Language Testing*, *37*(3), 453–471. https://doi.org/10.1177/0265532220912412

Hulstijn, J., Schoonen, R., Jong, N. H. de, Steinel, M. P., & Florijn, A. (2012). *Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR) 1*. https://doi.org/10.1177/0265532211419826

IIEF. (2022). *TOEFL ITP Test Schedule – IIEF*. https://www.iief.or.id/toefl-itp-test-schedule

International Test Center. (2022). *Jadwal Tes – International Test Center*. https://itc-indonesia.com/jadwal-tes/

Isaacs, T., Hu, R., Trenkic, D., & Varga, J. (2023). Examining the predictive validity of the Duolingo English Test: Evidence from a major UK university. *Language Testing*, *40*(3). https://doi.org/10.1177/02655322231158550

Isbell, D. R., & Kremmel, B. (2020). Test Review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Ismail, F. K. M., & Zubairi, A. M. B. (2021). Item Objective Congruence analysis for multidimensional items content validation of a reading test in Sri Lankan University. *English Language Teaching*, *15*(1), 106. https://doi.org/10.5539/elt.v15n1p106

Jin, Y. (2022). Consequential research of accountability testing: The case of the CET. *Language Testing in Asia*, *12*(1), 15. https://doi.org/10.1186/s40468-022-00165-6

Jung, Y. J., Crossley, S., & McNamara, D. (2019). Predicting second language writing proficiency in learner texts using computational tools. *The Journal of AsiaTEFL*, *16*(1), 37–52. https://doi.org/10.18823/asiatefl.2019.16.1.3.37

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*(1), 17–24. https://doi.org/10.1037/h0057123

Kennedy, I. (2021). Sample size determination in test-retest and Cronbach alpha reliability estimates. *Middle East Research Journal of Humanities and Social Sciences*, *1*(1), 16–24. https://doi.org/10.36348/merjhss.2021.v01i01.003

Kennedy, I. (2022). Sample size determination in test-retest and Cronbach alpha reliability estimates. *British Journal of Contemporary Education*, *2*(1), 17–29. https://doi.org/10.52589/BJCE-FY266HK9

Kumar, V. (2023). Using item analysis to evaluate hand hygiene self-assessments at Alberta health services. *American Journal of Infection Control*, *51*(6), 683–686. https://doi.org/10.1016/j.ajic.2022.08.030

Lu, X., & Hu, R. (2022). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods*, *54*(3), 1444–1460. https://doi.org/10.3758/s13428-021-01675-6

Madya, S., Retnawati, H., Purnawan, A., Putro, N. H. P. S., & Apino, E. (2019). The Equivalence of TOEP Forms. *TEFLIN Journal - A Publication on the Teaching and Learning of English*, *30*(1), 88–104. https://doi.org/10.15639/teflinjournal.v30i1/88-104

Mahesar, I. K., & Jokhio, A. A. (2021). Investigating the impact of resilience on learners' motivated behavior of L2 and proficiency in English of university students at Karachi, Pakistan. *Ethical Lingua: Journal of Language Teaching and Literature*, *8*(2), Article 2. https://doi.org/10.30605/25409190.290

Mohseni, A. (2021). The impact of genre-based instruction on Iranian intermediate EFL learners' writing skills. *Vision: Journal for Language and Foreign Language Learning*, *10*(2), 115–132. https://doi.org/10.21580/vjv11i110596

Natova, I. (2021). Estimating CEFR reading comprehension text complexity. *The Language Learning Journal*, *49*(6), 699–710. https://doi.org/10.1080/09571736.2019.1665088

Nordström, T., Andersson, U. B., Fälth, L., & Gustafson, S. (2019). Teacher inquiry of using assessments and recommendations in teaching early reading. *Studies in Educational Evaluation*, *63*, 9–16. https://doi.org/10.1016/j.stueduc.2019.06.006

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching*, *47*(2), 228–249. https://doi.org/10.1017/S0261444811000206

Ockey, G. J., & Gokturk, N. (2019). Standardized language proficiency tests in higher education. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *Second Handbook of Information Technology in Primary and Secondary Education* (pp. 1–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-58542-0_25-1

Pearson, W. S. (2020). Mapping English language proficiency cut-off scores and pre-sessional EAP programmes in UK higher education. *Journal of English for Academic Purposes*, *45*, 100866. https://doi.org/10.1016/j.jeap.2020.100866

Pearson, W. S. (2021). The predictive validity of the Academic IELTS test: a methodological synthesis. *ITL - International Journal of Applied Linguistics*, *172*(1), 85–120. https://doi.org/10.1075/itl.19021.pea

Pearson, W. S. (2023). Test review: High-stakes English language proficiency tests—Enquiry, resit, and retake policies. *Language Testing*, *40*(4). https://doi.org/10.1177/02655322231186706

Penuel, W., Roschelle, J., & Shechtman, N. (2007). Designing formative assessment software with teachers: an analysis of the co-design process. *Research and Practice in Technology Enhanced Learning*, *2*, 51–74. https://doi.org/10.1142/S1793206807000300

Pusat Bahasa UNAIR. (2025). UNAIR'S ELPT. *Pusat Bahasa dan Multibudaya*. https://pusatbahasa.unair.ac.id/unairs-elpt/

Pusat Pengembangan Bahasa UIN Sunan Ampel Surabaya. (2023, February 16). *Layanan P2B - UINSA*. https://uinsa.ac.id/p2b/layanan-p2b

Reynolds, M., Gates, E., Hummelbrunner, R., Marra, M., & Williams, B. (2016). Towards Systemic Evaluation. *Systems Research and Behavioral Science*, *33*(5), 662–673. https://doi.org/10.1002/sres.2423

Renandya, W. A., Hamied, F. A., & Joko, N. (2018). English language proficiency in Indonesia: issues and prospects. *The Journal of Asiatefl*, *15*(3). https://doi.org/10.18823/asiatefl.2018.15.3.618

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.

Roy, R., Sukumar, G. M., Philip, M., & Gopalakrishna, G. (2023). Face, content, criterion and construct validity assessment of a newly developed tool to assess and classify

work–related stress (TAWS– 16). *PLOS ONE*, *18*(1), e0280189. https://doi.org/10.1371/journal.pone.0280189

Saadatara, A., Kiany, G., & Talebzadeh, H. (2023). Bundles to beat the band in high-stakes tests: Pedagogical applications of an exploratory investigation of lexical bundles across band scores of the IELTS writing component. *Journal of English for Academic Purposes*, *61*, 101208. https://doi.org/10.1016/j.jeap.2022.101208

Sari, N. A., & Mualimin, M. (2021). The influence of the pandemic on the motivation of EAP learners in studying IELTS. *E3S Web of Conferences*, *317*, 02031. https://doi.org/10.1051/e3sconf/202131702031

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, *103*, 101602. https://doi.org/10.1016/j.ijer.2020.101602

Shirali, G., Shekari, M., & Angali, K. A. (2018). Assessing reliability and validity of an instrument for measuring resilience safety culture in sociotechnical systems. *Safety and Health at Work*, *9*(3), 296–307. https://doi.org/10.1016/j.shaw.2017.07.010

Shou, Y., Sellbom, M., & Chen, H.-F. (2022). 4.02—Fundamentals of Measurement in Clinical Psychology. In G. J. G. Asmundson (Ed.), *Comprehensive Clinical Psychology (Second Edition)* (pp. 13–35). Elsevier. https://doi.org/10.1016/B978-0-12-818697-8.00110-2

Solovjeva, S. V., & Baksheev, D. P. (2021). Preparation for international language exams as a means of developing intercultural competence. *Proceedings of FIR Conferences - International Relations: History, Theory, Practice*, 425–430. https://elib.bsu.by/handle/123456789/268895

Sridhanyarat, K., Pathong, S., Suranakkharin, T., & Ammaralikit, A. (2021). The development of STEP, the CEFR-based English proficiency test. *English Language Teaching*, *14*(7), 95. https://doi.org/10.5539/elt.v14n7p95

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping TOEFL® ITP scores onto the Common European Framework of Reference*. English Testing Services.

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto the Common European Framework of Reference: an application of standard-setting methodology. *ETS Research Report Series*, *2008*(1), i–75. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

The CEFR Levels. (2022). Common European Framework of Reference for Languages (CEFR). https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Turner, R. C., & Carlson, L. (2003). Indexes of Item-Objective Congruence for Multidimensional Items. *International Journal of Testing*, *3*(2), 163–171. https://doi.org/10.1207/S15327574IJT0302_5

UNESCO Institute for Lifelong Learning. (2021). *GAL Country Profiles as of February 2021*. United Nations Educational, Scientific and Cultural Organization.

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales? external factors with indirect influence on reliability estimators. *Procedia Economics and Finance*, *20*, 679–686. https://doi.org/10.1016/S2212-5671(15)00123-9

Vincent, W., & Shanmugam, S. K. S. (2020). The role of classical test theory to determine the quality of classroom teaching test items. *Pedagogia: Jurnal Pendidikan*, *9*(1), Article 1. https://doi.org/10.21070/pedagogia.v9i1.123

Walski, T. (2014). Consequential Research. *Journal of Water Resources Planning and Management*, *140*, 559–561. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000430

Wibowo, E. P., Lestari, A., Adrian, M., Firdausiyah, K. S., Islamiy, J. H., Putri, R. D., Kemal, T., & Kurniawan, E. (2024). Exploring the ethical dimensions of testing and assessment: an investigation into grade inflation among EFL teachers. *OKARA: Jurnal Bahasa dan Sastra*, *18*(1), 71–97. https://doi.org/10.19105/ojbs.v18i1.12676

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300. https://doi.org/10.1191/0265532205lt309oa

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). https://doi.org/10.1177/21582440231199692