Journal of Educational Technology and Innovation
Homepage: http://jurnal.unipar.ac.id/index.php/jeti/
P - ISSN 2621-2137 *E - ISSN* 2621-2080
Vol 9 No 1, 2026, pp 84-101
DOI: https://doi.org/10.31537/jeti.v9i1.3109

Analysis of the Alignment of the A1-Level Arabic Language Assessment on the AlifBee App with the CEFR Levels

Alyaa Adhinna Putri[1], Asep Sopian[2], Hikmah Maulani[3]

[123]Arabic Language Education, Faculty of Language and Literature Education, Indonesia University of education, Bandung, Indonesia
Email: alyaaadhin@upi.edu[1] ; asepsopian@upi.edu [2]; hikmahmaulani@upi.edu[3]

Article Info

Keywords:

assessment; arabic language; CEFR; digital learning; Modern Standard Arabic (MSA)

Article history:
Received: [*April, 30 2026*]
Revised: [*Mei, 26 2026*]
Accepted: [*June, 10 2026*]
Published: [*June, 30 2026*]

Abstract

Introduction: Digital language learning applications increasingly incorporate proficiency level labels to guide learners, yet the validity of such classifications remains underexamined. This study investigates the alignment between A1-level assessments in the AlifBee Arabic learning application and the Common European Framework of Reference for Languages (CEFR) descriptors. **Methods:** A descriptive qualitative approach was employed, using document analysis to systematically compare 48 assessment items from AlifBee's A1 level against CEFR A1 proficiency descriptors. **Results:** Analysis revealed that 33 items (69%) demonstrated alignment with the A1 level, while the remaining 15 items (31%) were more appropriately classified at the Pre-A1 level. Misalignment was predominantly found in vocabulary translation, letter arrangement, and oral repetition tasks, all of which measure isolated linguistic knowledge rather than communicative competence. **Discussion:** These findings indicate that the A1 classification in AlifBee has not been fully substantiated by task demands consistent with CEFR descriptors. The study recommends integrating more communicative-based tasks to strengthen level validity, with implications for developers of digital language learning tools seeking to align their assessments with internationally recognized proficiency frameworks.

Keywords: Assessment; Arabic language; CEFR; Digital learning; Modern Standard Arabic (MSA)

Abstrak

Pendahuluan: Aplikasi pembelajaran bahasa digital semakin banyak menyertakan label tingkat kemahiran untuk memandu pembelajar, namun validitas klasifikasi tersebut masih belum banyak diteliti. Penelitian ini menyelidiki keselarasan antara asesmen tingkat A1 pada aplikasi pembelajaran bahasa Arab AlifBee dengan deskriptor Common European Framework of Reference for Languages (CEFR). **Metode:** Pendekatan kualitatif deskriptif diterapkan dengan menggunakan analisis dokumen untuk membandingkan secara sistematis 48 butir asesmen tingkat A1 AlifBee terhadap deskriptor kemahiran A1 CEFR. **Hasil:** Hasil analisis menunjukkan bahwa 33 butir (69%) selaras dengan tingkat A1, sementara 15 butir sisanya (31%) lebih tepat diklasifikasikan pada tingkat Pre-A1. Ketidakselarasan terutama ditemukan pada tugas penerjemahan kosakata, penyusunan huruf, dan pengulangan lisan, yang mana seluruhnya

hanya mengukur pengetahuan linguistik yang terisolasi alih-alih kompetensi komunikatif. **Diskusi:** Temuan ini mengindikasikan bahwa klasifikasi A1 pada AlifBee belum sepenuhnya didukung oleh tuntutan tugas yang konsisten dengan deskriptor CEFR. Penelitian ini merekomendasikan integrasi tugas berbasis komunikatif untuk memperkuat validitas level, serta memberikan implikasi bagi pengembang aplikasi pembelajaran bahasa digital dalam menyelaraskan asesmen dengan kerangka kemahiran bahasa yang diakui secara internasional.

Kata kunci: Asesmen; Bahasa Arab; CEFR; Modern Standard Arabic (MSA); Pembelajaran digital.

Corresponding Author:
Alyaa Adhinna Putri
Arabic Language Education, Faculty of Language and Literature Education, Indonesia University of education, Bandung, Indonesia
Affiliation ; alyaaadhin@upi.edu

1. INTRODUCTION

The development of digital-based language learning over the past decade has significantly transformed the landscape of language assessment (Aslan, 2025). Language learning applications no longer function merely as platforms for content delivery, but also as self-assessment systems that determine learners' levels, progress, and competency achievement (Gou, 2023). While this phenomenon expands access to learning, it simultaneously raises a fundamental concern: to what extent do assessment systems embedded in digital applications demonstrate validity, comparability, and alignment with internationally recognized language proficiency standards (Muawanah et al., 2024). In the context of educational globalization, proficiency claims without clear reference to a common framework risk creating ambiguity in the interpretation of learners' abilities (Shen et al., 2022).

As a universal framework, the Common European Framework of Reference for Languages (CEFR) serves as a primary benchmark for the development of instructional materials, syllabus design, and the standardization of foreign language assessment systems (Kamal et al., 2025). Grounded in a communicative competence framework, CEFR defines language proficiency through can-do descriptors across six levels (A1–C2), covering linguistic, pragmatic, and sociolinguistic competencies (Council of Europe, 2020). CEFR has been widely recognized as an international reference in curriculum development, test design, and the reporting of language learning outcomes (Lowie et al., 2010). However, linking assessments to CEFR is not merely a matter of assigning level labels; rather, it requires careful analysis of task constructs, linguistic complexity, and empirical evidence to support score interpretation (Bachman, 2004). This process, often referred to as standard-setting or linking, demands systematic alignment procedures involving expert judgment, construct mapping, and empirical validation (Council of Europe, 2020). Without systematic alignment procedures, the use of CEFR labels risks becoming symbolic rather than substantive (Wisniewski, 2018).

This issue of alignment becomes increasingly critical within the context of educational technology and digital assessment systems. A study by Benedetto et al. (2025) demonstrates that even large language models still face challenges in accurately encoding and applying CEFR descriptors. Similarly, Fulcher (2014) argues that the interpretive validity of proficiency levels is fundamentally dependent on the quality of construct specification embedded in tasks. This finding underscores that the implementation of CEFR in digital systems is not an automatic process, but rather requires a deep conceptual understanding of

competency descriptors. On the other hand, the advancement of technology-based assessment, including computer-assisted language assessment, offers opportunities for innovation while simultaneously demanding stronger validity and accountability (Winke & Isbell, 2017).

In the context of the Arabic language, this issue presents its own complexities. Arabic is characterized by diglossia, namely the coexistence of Modern Standard Arabic (MSA) and spoken varieties that serve different social and communicative functions. The morphological complexity and root-based derivation system in MSA, as well as the absence of short vowels in written texts, substantially influence task design, cognitive load, and the interpretation of proficiency levels in assessment contexts (Holes, 2004; Ryding, 2005). Furthermore, Najiyah et al. (2026) emphasize that Arabic language testing faces various challenges, ranging from the standardization of constructs and the selection of language varieties to alignment with international frameworks. Inconsistencies in level mapping frequently arise when assessment tasks prioritize isolated linguistic forms over functional communicative use. This concern is particularly salient in digital platforms that rely on gamified or decontextualized item formats. On the other hand, Alanazi (2024) highlights that issues regarding the quality of translation and the adaptation of the CEFR into Arabic directly impact the understanding of descriptors and assessment practices, thereby potentially affecting the accuracy of proficiency level interpretation. These findings are also consistent with Norrbom & Zuboy (2021), who emphasize the importance of conceptual clarity in the process of adapting the CEFR framework to the Arabic language context.

Several studies have attempted to link Arabic language learning with the CEFR framework. For instance, Alrababa'h et al. (2024) examined the suitability of Arabic reading texts for non-native speakers at the University of Jordan from a CEFR perspective and found variations in students' perceptions of text difficulty. In Indonesia, CEFR-based policies in Arabic language teaching have also begun to receive attention, particularly in the context of Islamic higher education (Musthofa, 2022) and the development of Arabic teaching materials within both national and global perspectives (Pransiska et al., 2024). Maulani et al. (2024) demonstrated that the adoption of CEFR in Arabic proficiency testing in Indonesia contributes to providing internationally standardized level descriptors, thereby supporting more objective assessments of language ability that are relevant to academic and professional needs. Furthermore, Dewi et al. (2025) found that Arabic language learning content on the TikTok account "Learn Arabic for Beginners" is largely aligned with A1–A2 competencies, particularly in terms of phonological mastery and the use of basic vocabulary, although aspects of interaction and mediation remain limited. Collectively, these studies demonstrate a growing scholarly interest in CEFR alignment within Arabic language education; however, they predominantly address formal instructional contexts and static digital content, leaving a notable gap in the analysis of assessment constructs embedded within interactive mobile learning applications.

Despite the growing body of research on CEFR-aligned Arabic language assessment, a critical gap remains: few, if any, studies have systematically analyzed the construct alignment of assessment tasks within interactive digital Arabic language learning applications. Existing research has predominantly focused on (1) curriculum and syllabus alignment in formal educational settings (Musthofa, 2022; Pransiska et al., 2024), (2) proficiency testing instruments in institutional contexts (Maulani et al., 2024), and (3) static digital content on social media platforms (Dewi et al., 2025). While these contributions are valuable, they do not address the internal assessment logic of mobile learning applications, which operate through automated item delivery, level placement algorithms, and rapid feedback mechanisms that may diverge from the proficiency constructs they claim to represent (Jurāne-Brēmane, 2023; Gou, 2023). This gap is particularly significant because a growing number of users worldwide rely on these platforms for self-directed language learning, and the validity of level labels directly shapes learners' self-assessment, learning trajectories, and credential interpretation (Bachman & Palmer, 2010). This lack of transparency is compounded by the fact that automated scoring and level placement systems in e-learning platforms frequently do not disclose the measurement parameters they employ, making external alignment analysis both difficult and necessary (Gou, 2023; Newton et al., 2021).

In the context of digital Arabic language learning, AlifBee serves as an example of a platform that utilizes Modern Standard Arabic (MSA) as the foundation for both instruction and assessment. The application offers ten levels of Arabic learning and caters to users across multiple countries. AlifBee accommodates users from ten countries, including Indonesia, Malaysia, France, Turkey, Spain, and the United Kingdom, which adopt the CEFR framework, as well as users from Arab countries, Uzbekistan, Kiswahili-speaking regions, and Urdu-speaking contexts that rely on MSA as a reference. This multilingual user base makes AlifBee a particularly relevant site for examining construct alignment, as the platform implicitly positions its levels as comparable across different proficiency frameworks without providing transparent empirical evidence for such claims. In the literature, level assignment and the interpretation of achievement require evidence of task construct alignment as well as assessment-based argumentation to ensure validity and comparability (Asli et al., 2024; Pellegrino et al., 2016).

In efforts to improve the quality of Arabic language teaching and learning, international frameworks such as the CEFR have been identified as promising instruments for standardizing and strengthening Arabic language education across various contexts (Musthofa, 2022; Salam et al., 2025). The application of CEFR in Arabic language learning is considered relevant, as the framework is designed based on learners' needs and takes into account the social contexts in which the language is learned (Soliman & Familiar, 2023).

Addressing this gap, this study conducts a systematic CEFR alignment analysis of A1-level assessment tasks in the AlifBee Arabic learning application. The choice of A1 is theoretically motivated: as the foundational entry point in the CEFR continuum, accurate construct alignment at this level is essential for the validity of all subsequent level classifications and for learners' initial self-placement (Jeon, 2025; Norris, 2018). Moreover, by focusing on the most basic proficiency level, this study establishes a benchmark for evaluating whether the application maintains appropriate cognitive load and linguistic complexity for beginner learners without prematurely escalating task demands (Abu-Zhaya & Arnon, 2024). The objectives of this study are to examine the extent to which assessment tasks in AlifBee align with CEFR A1 descriptors and to provide an evidence-based reference for developers, educators, and learners in understanding language proficiency levels (Costantino & Martin, 2025; Newton et al., 2021). Specifically, this study identifies the correspondence between task demands and CEFR descriptors so that non-native Arabic learners can obtain a more objective and standardized understanding of their proficiency. Overall, this research not only focuses on level classification but also contributes to improving transparency, consistency, and accountability in digital Arabic language assessment within the evolving landscape of educational technology.

2. METHODS

Research Design

This study employed a descriptive qualitative approach using document analysis to examine the alignment of A1-level Arabic language assessment in the AlifBee application with the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2020). A qualitative approach was selected because the aim of the study was not to test instructional effectiveness, but rather to identify the degree of equivalence between assessment task demands and language proficiency descriptors through systematic documentary examination (Romero-Yesa et al., 2023). The analytical procedure was carried out in four sequential steps: (1) collecting and inventorying all assessment items from AlifBee's A1 level; (2) identifying the linguistic and communicative demands of each item based on task type, skill focus, and response format; (3) mapping each item against the corresponding CEFR A1 can-do descriptors across relevant competency categories; and (4) categorizing each item as aligned with CEFR A1 or more appropriately mapped to Pre-A1 based on the degree of correspondence between task demands and descriptor criteria. This systematic procedure enabled the researcher to interpret the meaning, structure, and characteristics of the assessment in depth without involving research participants, thereby maintaining analytical focus on the substance of assessment content and language proficiency descriptors (Bowen, 2009).

Data Sources

The data for this study were derived from two primary sources. The first consists of assessment documents from AlifBee's A1 level, comprising a total of 48 items drawn specifically from the initial evaluation unit (Exercise 1). These items encompass a range of task formats, including vocabulary recognition, letter arrangement, oral repetition, and translation tasks, and were selected because the initial unit systematically introduces the foundational competencies expected at the A1 stage, namely recognition of basic vocabulary, simple phrases, and isolated linguistic forms (Council of Europe, 2020). Focusing on this unit allowed the study to establish a baseline measure of construct alignment at the most elementary level of the application's assessment sequence, before any increase in linguistic complexity occurs across subsequent units. It is acknowledged as a research delimitation that this study analyzes only the first unit of AlifBee's A1 level; assessment items from subsequent units within the same level were not included in order to maintain analytical focus and manageability within the scope of a single study. Second, official CEFR documents containing proficiency level descriptions, categories of communicative activities, and language user competencies served as the reference framework for mapping (Council of Europe, 2020). The sample selection employed purposive sampling, whereby both document sets were intentionally selected based on their direct relevance to the research objectives (Creswell & Poth, 2018).

Data Collection

Data were collected through documentation study. The researcher identified task types, language skills assessed, contexts of language use, and indicators of language ability present in each assessment item. Subsequently, relevant CEFR descriptors were extracted and both data sources were organized into comparable units of analysis for mapping purposes.

Data Analysis

Data analysis followed the interactive model of Miles, Huberman, and Saldaña (2014): data reduction, data display, and conclusion drawing/verification. In this study, the CEFR alignment framework was primarily operationalized during the data reduction stage and served as the basis for structuring the data presentation. During data reduction, two main procedures from the CEFR alignment handbook were applied: familiarization and specification. In the familiarization stage, the researcher reviewed CEFR A1 and Pre-A1 descriptors, identified relevant levels, and determined the most appropriate scales for the AlifBee assessment context. In the specification stage, each assessment item was analyzed based on task type, targeted skills, communicative context, and language performance demands, and then mapped to the most appropriate CEFR descriptor (British Council et al., 2022). The mapping results were presented in domain-based analysis tables and an overall summary table.

During the data display stage, the researcher adopted an instrument specification table referring to CEFR-informed Learning, Teaching and Assessment (Nagai et al., 2020). At this stage, only items mapped to the Pre-A1 level were described in detail, as these items directly represent potential gaps between the application's level labels and CEFR descriptors. Items aligned with A1 were presented in aggregate form within the summary table without detailed item-by-item description, as the primary objective of the study was to identify misalignment rather than to provide a full descriptive inventory of all items. Finally, in the conclusion drawing/verification stage, the researcher reviewed the consistency of mapping decisions in relation to CEFR descriptors and the adequacy of evidence supporting the analysis. Verification was conducted through rechecking coding results and ensuring alignment between evidence and level descriptors before reporting the final conclusions.

3. RESULTS

3.1. General Description of the Assessment Instrument

The focus of analysis in this study is directed at the first evaluation (Exercise 1) in the AlifBee application. Structurally, this evaluation consists of three competency domains: vocabulary (25 items), composition and sentences (17 items), and dialogue (6 items), resulting in a total of 48 assessment items analyzed in this study.One general characteristic identified is that all items are untimed, meaning they are not constrained by a specific time limit. This condition aligns with pedagogical principles for beginner-level learning; as noted by Oyebode and Nicholls (2021), time pressure in assessment can hinder reflective processing among novice learners and induce language anxiety that distorts authentic performance. Therefore, the time flexibility in AlifBee evaluations has the potential to support more accurate articulation and optimal vocabulary internalization.

Furthermore, to measure the level of alignment between assessment items and the learning materials presented in the preceding unit, the percentage of content alignment was calculated. This calculation employed a formula based on Bloom's taxonomy as follows:

$$P = \frac{n}{\mathrm{N}} \times 100$$

n: the number of learning content indicators represented in the assessment items.

N: the total number of competency indicators targeted in the learning objectives (based on operational verbs in the cognitive domain of Bloom's taxonomy) (Gunawan & Palupi, 2016).

The calculated percentage results were then classified into several categories of alignment levels, namely:

Table 1. Criteria for Item Alignment

Percentage	Category
81 - 100%	Highly aligned items
60 - 79%	Aligned items
40 - 59%	Less aligned items
20 - 39%	Not aligned items

The criteria in Table 1 are used to determine the degree to which the assessment items and the instructional materials in the AlifBee application are aligned with the intended indicators (Utomo, 2019). The results of the data analysis for the items in Exercise 1 of the AlifBee application are presented in the form of an analytical matrix in the following table:

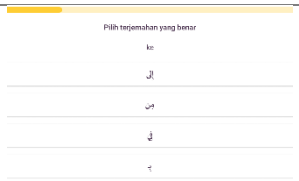
Table 2. Exercise 1 – Vocabulary Section:

no	Assessment type	Alignment of Assessment Items with Previous Learning Material (%)	Timing assessment	Description of test items	Descriptor CEFR	CEFR level
1.		33,3%	-	Listening: - Speaking: - Writing: - Reading: ✓	Has a very basic range of isolated words and phrases relating to personal details and specific concrete situations	Pre -A1

In the first item, the assessment instrument requires learners to directly translate an image into Arabic. This task involves simple visual identification without incorporating sentence context or communicative interaction. The competency assessed in this item falls under the Vocabulary Range scale at the Pre-A1 level, as it only requires the ability to name an object based on a visual stimulus. At this stage, learners have not yet reached A1 competency, which requires understanding short texts or simple phrases.

The alignment percentage for this item is 33.3%, which falls into the category of less aligned items. This indicates that the task primarily demands the ability to recall and recognize a previously introduced image (C1 & C2), rather than to comprehend or use language in a communicative context. Therefore, it does not sufficiently represent the expected level of independent language use at the A1 level.

Table 3. Exercise 1 – Vocabulary Section:

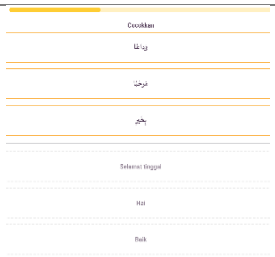
no	Assessment type	Alignment of Assessment Items with Previous Learning Material (%)	Timing assessment	Description of test items	Descriptor CEFR	CEFR level
2.		16,6%	-	Listening: - Speaking: - Writing: - Reading: ✓	can recognise familiar names, words and very simple sentences, for example on notices and posters or in catalogues.	Pre-A1

In the second item, the task requires learners to translate a preposition from Indonesian into Arabic in isolation. There is no supporting sentence or contextual usage provided, which places the competency assessed within basic linguistic units and vocabulary mastery (Vocabulary Range) at the Pre-A1 level. By isolating the word “to” without embedding it in a phrase (such as “to school” or “to home”), the item falls below the A1 standard. From a CEFR perspective, A1 requires comprehension of “a single phrase at a time,” whereas this item merely tests memorization of a single functional word.

The alignment percentage for this item is 16.6%, which falls into the category of not aligned items. This is because the task only engages a single cognitive domain, namely C1 (remembering). Learners simply retrieve a lexical form from memory without needing to understand context or use language communicatively. Since the task is limited to single-word memorization, it does not meet the characteristics of A1, which begin to require understanding or use of simple phrases for concrete purposes. Therefore, this item is more appropriately classified at the Pre-A1 level.

Table 4. Exercise 1 – Vocabulary Section:

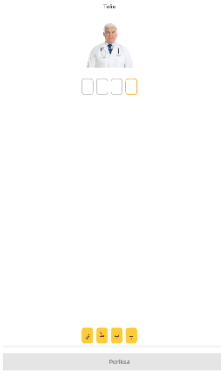
no	Assessment type	Alignment of Assessment Items with Previous Learning Material (%)	Timing assessment	Description of test items	Descriptor CEFR	CEFR level

3.		50%	-	Listening: - Speaking: - Writing: - Reading: ✓	Can use basic greetings and polite forms of greeting and farewell in everyday life.	Pre-A1

In the third item, the assessment instrument focuses on meaning recognition and basic reading comprehension, with primary emphasis on the Linguistic Range aspect (particularly Vocabulary Range). Learners are tested on their ability to recognize meaning and demonstrate basic reading comprehension through a matching task involving greeting expressions and phrases in English with their Indonesian equivalents. This activity is decontextualized, as the vocabulary is presented in isolation without involving more complex dialogic structures. Therefore, the competency measured emphasizes mastery of formulaic expressions and a repertoire of isolated words at the Pre-A1 level.

The alignment percentage for this item is 50%, which falls into the category of less aligned items. This is because the task engages three cognitive domains: C1 (remembering), C2 (understanding), and C3 (applying). Learners are required to recall linguistic forms, understand their meanings, and apply this knowledge to select the correct pairs. However, the task remains limited to direct meaning matching and does not involve dialogic context or authentic social language use. Therefore, although cognitively broader than the previous items, it still does not meet the communicative demands of the A1 level.

Table 5. Exercise 1 – Vocabulary Section:

no	Assessment type	Alignment of Assessment Items with Previous Learning Material (%)	Timing assessment	Description of test items	Descriptor CEFR	CEFR level
4.		50%	-	Listening: - Speaking: - Writing: ✓ Reading: -	Can use a very basic repertoire of isolated words and phrases, and perform basic literacy tasks such as spelling and producing individual letters to form isolated concrete words	Pre-A1

In the fourth item, the assessment task focuses on the earliest stage of literacy and written production. Learners are instructed to arrange scrambled or separated Arabic letters (hijaiyah) into a complete and meaningful word. This type of task reflects basic literacy skills at the Pre-A1 level, particularly related to orthographic control and initial vocabulary recognition. Since the task only requires learners to form a single word without involving sentence construction or grammatical understanding, its level of difficulty remains below the full A1 standard.

The alignment percentage for this item is 50%, which falls into the category of less aligned items. This is because the task is still limited to activities such as recalling, writing, and recognizing previously introduced vocabulary or phrases. It does not yet require learners to understand meaning contextually or to use language in a more complex and communicative manner. Therefore, this item is more appropriately classified at the Pre-A1 level rather than A1.

Table 6. Exercise 1 – Vocabulary Section:

no	Assessment type	Alignment of Assessment Items with Previous Learning Material (%)	Timing assessment	Description of test items	Descriptor CEFR	CEFR level
5.		16,6%	-	Listening: ✓ Speaking: ✓ Writing: - Reading: -	Can reproduce correctly a very limited range of sounds, as well as simple words and phrases from a memorized repertoire.	Pre-A1

In the fifth item, the assessment instrument focuses on listening skills and spoken production. Learners are instructed to listen to a very short self-introduction phrase and then repeat and pronounce it according to the audio provided. This type of task specifically maps onto phonological control and basic spoken production at the Pre-A1 level. The task is purely repetitive (imitation of pronunciation) and does not require learners to engage in two-way communication or construct grammatical structures.

The alignment percentage for this item is 16.6%, which falls into the category of not aligned items. This is because the task only requires learners to recall and recognize previously introduced lexical forms rather than to understand or use language in a communicative context. Therefore, this item is more appropriately classified at the Pre-A1 level rather than A1.

Table 7. Exercise 1 – Vocabulary Section:

no	Assessment type	Alignment of Assessment Items with Previous Learning Material (%)	Timing assessment	Description of test items	Descriptor CEFR	CEFR level
6.		33,3%	-	Listening: - Speaking: - Writing: - Reading: ✓	Can recognize and produce a basic repertoire of isolated words and formulaic expressions related to concrete situations when supported by visual prompts.	Pre-A1

In the sixth item, the assessment task consists solely of identifying the meaning of a visual stimulus. Learners are presented with an image representing a particular situation and are instructed to translate it directly into the appropriate Arabic word or phrase. This instrument falls under the Vocabulary Range competency scale at the Pre-A1 level. Since learners are only required to recall a single lexical unit from

a concrete situation, this item functions as a precise measure of basic vocabulary knowledge to validate early literacy competence among non-Arabic learners.

The alignment percentage for this item is 33.3%, which falls into the category of less aligned items. This is because the task only requires learners to recall and interpret a previously introduced image (C1 & C2), rather than to comprehend or use language within a communicative context. Therefore, it does not adequately represent the expected level of independent language use at the A1 level.

3.2. Results of Mapping by Domain

3.2.1 Vocabulary Domain (25 items)

The analysis of the 25 items in the vocabulary domain indicates that 16 items are aligned with the A1 level, while 9 items are more appropriately classified at the Pre-A1 level. These nine misaligned items consist of: (a) five image-based vocabulary translation items, (b) two items involving the construction of words from hijaiyah letters, and (c) two spoken production items in the form of repetition of heard utterances.

Within the CEFR framework, the five image-to-word translation items are mapped to the Vocabulary Range at the Pre-A1 level, as they only require learners to name or identify a single object based on a visual stimulus without involving sentence context or communicative interaction. In contrast, CEFR A1 requires the ability to understand and use simple phrases, not merely isolated word recognition. The two letter-to-word construction items are mapped to orthographic control (Pre-A1), as they focus solely on forming single words rather than constructing meaningful phrases. Meanwhile, the two spoken repetition items are categorized under phonological control at the Pre-A1 level, as they are purely repetitive and do not require independent communicative responses.

In terms of alignment with instructional content, the Pre-A1 items fall within a range of 16.6% to 50%. A score of 16.6% indicates the involvement of only one cognitive domain (C1: remembering), 33.3% reflects the involvement of two domains (C1–C2: remembering and limited understanding), and 50% involves three domains (C1–C3). However, an increase in the number of cognitive domains does not necessarily correspond to a higher CEFR level. Even items with a 50% alignment remain categorized as Pre-A1 if language use is still isolated and does not engage with a communicative context.

3.2.2. Composition and Sentence Domain (17 items)

In the composition and sentence domain, 11 items are aligned with the A1 level, while 6 items are mapped to the Pre-A1 level. These six misaligned items consist of four spoken repetition tasks (mapped to phonological control at Pre-A1) and two items involving the arrangement of letters into words (mapped to orthographic control at Pre-A1). Although this domain is labeled “composition and sentences,” the findings indicate that some items do not yet require learners to produce or comprehend complete simple sentences. This pattern is consistent with the findings in the vocabulary domain, where tasks are predominantly oriented toward form reproduction rather than communicative language use

3.2.3. Dialogue Domain (6 items)

In contrast to the two previous domains, all 6 items in the dialogue domain are aligned with the A1 level. In this domain, learners are no longer limited to recognizing words or imitating sounds; instead, they are guided to understand simple dialogues and grasp meaning within real communicative contexts. Language is presented as a more complete unit of meaning rather than as isolated lexical items. Within the CEFR mapping, these dialogue items correspond to A1 receptive and interactional activities, which require understanding very simple everyday expressions in concrete situations. From a cognitive perspective, the tasks extend beyond C1 (remembering) and move toward functional meaning comprehension, a characteristic that qualitatively distinguishes this domain from the previous two.

3.3. Recapitulation of mapping results

Table 8. Recapitulation of Mapping of AlifBee Level A1 Assessment Items to CEFR

Assessment Domain	Number of Items	Aligned to A1	Pre-A1

Vocabulary	25	16 (64%)	9 (36%)
Composition and Sentences	17	11 (65%)	6 (35%)
Dialogue	6	6 (100%)	0 (0%)
Total	**48**	**33 (69%)**	**15 (31%)**

Table 8 presents the overall recapitulation of the alignment of assessment items with CEFR levels. Out of a total of 48 items, 33 items (69%) are aligned with the A1 level, while 15 items (31%) are more appropriately mapped to the Pre-A1 level. The greatest misalignment is found in the vocabulary and composition–sentence domains, whereas all dialogue items are fully aligned with A1. These findings indicate that the first evaluation in AlifBee is not yet fully consistent in representing the A1 level, as approximately one-third of the items still operate below the A1 standard.

Table 9. Distribution of Language Skills in Pre-A1 Items

Skills	Number of Skill Involvements
Listening	6
Reading	5
Speaking	6
Writing	4

The total in Table 9 is not identical to the 15 Pre-A1 items, as some oral items involve more than one skill simultaneously (for example, a single item combining listening and speaking). Therefore, this table should be interpreted as a distribution of skill involvement rather than as a count of mutually exclusive items.

4. DISCUSSION

The findings of this study indicate that out of 48 assessment items analyzed in the first evaluation of AlifBee at the A1 level, 33 items (69%) are aligned with CEFR A1 descriptors, while 15 items (31%) are more appropriately mapped to the Pre-A1 level. This misalignment is not randomly distributed but follows a clear pattern across specific task types: image-based vocabulary translation, arrangement of hijaiyah letters into words, and repetition of spoken utterances. This pattern directly addresses the research objective outlined in the introduction, namely identifying the extent to which assessment demands within a digital platform correspond to CEFR competency descriptors. These findings also confirm concerns raised earlier in the study that proficiency level claims in digital applications may lack sufficient empirical grounding if not supported by systematic alignment procedures (Bachman, 2004; Wisniewski, 2018).

From a pedagogical standpoint, the findings of this study may assist AlifBee users, particularly those who are more familiar with CEFR levels than with the MSA-based leveling system used in the application. By mapping AlifBee assessment items onto CEFR descriptors, this study provides a clearer reference point for understanding learners' actual proficiency levels. The finding that several A1 assessment items are more appropriately categorized as Pre-A1 suggests that learners with limited prior knowledge of Arabic may benefit from beginning at the Pre-A1 stage before progressing to A1-level tasks. For educators, these findings also suggest the need to supplement AlifBee with communicative classroom activities, particularly oral interaction, contextualized sentence production, and simple dialogue practice (Zeng & Huang, 2021).

From a design perspective, the findings suggest that AlifBee's assessment system can still be improved to better align with CEFR levels. Tasks that are less contextual, such as arranging letters and translating isolated vocabulary, may be more suitable as Pre-A1 preparatory exercises rather than A1-level assessment items. Therefore, developers could add more tasks that reflect CEFR A1 competencies, such as answering simple personal questions or choosing appropriate responses in everyday dialogues. In addition, clearer criteria for level placement would help make the application's proficiency claims more valid and increase users' trust in the assessment results (Gou, 2023; Jurāne-Brēmane, 2023).

Interpretation of Findings

Misalignment as a Construct Issue, Not Content

The core finding of this study is that the misalignment between AlifBee's A1 label and CEFR A1 descriptors is not primarily caused by inappropriate content topics, but by the limited communicative orientation of the assessment tasks. Although the topics generally fall within the A1 scope, the misaligned items were classified as Pre-A1 because they required only lexical recall, basic orthographic control, or sound repetition. These forms of performance differ from the communicative demands of CEFR A1, which emphasizes the ability to use language in simple, concrete, and meaningful contexts. This finding reinforces the argument of Bachman (2004) and Lowie et al. (2010) that assessment alignment cannot be determined solely on the basis of content coverage, but must also consider the quality of language performance required from learners. Similarly, Chapelle and Voss (2017) argue that the validity of technology-mediated language assessment depends on whether task demands reflect the target language use domain. In the case of AlifBee, the Pre-A1 items do not fully meet this condition because they limit language use to recognition and reproduction rather than contextualized communication.

Within the CEFR alignment framework, the key distinction between Pre-A1 and A1 lies in the presence of communicative context. CEFR A1 requires learners to understand and use simple phrases in concrete situations and to communicate in a limited way when the interlocutor speaks slowly and clearly (Alexiou & Stathopoulou, 2021). By contrast, tasks such as translating images into single words or arranging hijaiyah letters present language as isolated units rather than as a means of interaction. This supports the view of Fulcher (2014) and Marzuki et al. (2013) that proficiency scales must be anchored in observable communicative behavior, not in the mastery of discrete linguistic sub-skills. Such a discrepancy highlights that CEFR alignment requires analysis at the level of construct, not merely content mapping.

The recurrence of these task types across multiple domains suggests that the issue is not incidental but reflects a structural tendency in the application's assessment design philosophy, wherein task sequencing prioritizes formal linguistic exposure over functional communicative performance (Khezrlou & Stockwell, 2025). As a result, learners may be assessed on their ability to recognize or reproduce isolated forms, even when the level label implies readiness for basic communicative use. This indicates that the A1 label in the first evaluation of AlifBee is not yet fully supported by assessment demands that consistently reflect CEFR A1 descriptors.

Furthermore, the distribution of misalignment across different language skills suggests that the issue does not lie within a single modality, but in the overall task design. This finding aligns with Schmidgall et al. (2019), who argue that cognitive taxonomies such as Bloom's revised taxonomy describe levels of thinking rather than levels of language use, and therefore cannot replace construct-based proficiency descriptors. Accordingly, increased cognitive complexity based on Bloom's taxonomy does not automatically correspond to a higher CEFR level, because the two frameworks represent different constructs.

Pattern of Misalignment Across Domains

Differences in alignment patterns across domains provide a clearer picture of the root of the issue. The dialogue domain demonstrates full alignment (100%) because it situates language within authentic interactional contexts. In contrast, the vocabulary domain (36% Pre-A1) and the composition sentence domain (35% Pre-A1) are still dominated by tasks that remain within the C1–C3 cognitive levels of Bloom's taxonomy without incorporating communicative context. Notably, an increase in the number of cognitive domains does not automatically correspond to a higher CEFR level; items with a 50% alignment involving three cognitive domains (C1–C3) are still mapped to Pre-A1 when language use remains isolated. This confirms that cognitive alignment with learning material and alignment with CEFR levels are two distinct dimensions that cannot substitute for one another. Xu et al. (2023) similarly caution that task difficulty in language assessment must be evaluated from the perspective of communicative demands rather than cognitive processing load alone.

The distribution of language skills within the Pre-A1 items also shows that misalignment is evenly spread across all modalities: listening (6 items), speaking (6 items), reading (5 items), and writing (4 items). The fact that this issue is not concentrated in a single skill suggests that the root of the problem is systemic, embedded in the overall philosophy of assessment design rather than in the choice of modality. Cummings and Anderson (2025) emphasize that inaccurate CEFR level placement can directly impact learner engagement and motivation, as learners assigned to inappropriate levels risk misinterpreting their actual competence.

The Complexity of MSA Diglossia as a Contextual Factor

In the context of Arabic as the object of study, these findings carry deeper implications. Arabic is characterized by diglossia, namely the coexistence of the standardized variety, Modern Standard Arabic (MSA), and spoken dialects that differ functionally and structurally (Ryding, 2005). As a result, tasks that only require recognition of isolated forms in MSA are not only misaligned with CEFR A1 standards in general, but also fail to represent the use of MSA in authentic communicative situations. This is particularly significant given that AlifBee targets users from both CEFR-adopting countries and MSA reference contexts, meaning that the construct validity of its assessments must simultaneously address two different proficiency frameworks, a dual demand that isolated task formats may not adequately address.

Furthermore, the root-based morphological system of MSA adds another layer of complexity: words that appear simple on the surface may involve more intricate relationships, meaning that mastery of isolated vocabulary does not necessarily reflect functional competence in MSA (Bar-On et al., 2018). Shehata (2021) further notes that the absence of short vowels in standard Arabic orthography introduces an additional processing demand for beginner learners that is not adequately captured by simple vocabulary recognition tasks, thereby widening the gap between assessed performance and actual communicative readiness. In addition, issues related to the quality of adapting CEFR descriptors into Arabic influence how competency constructs are operationalized in MSA-based assessment instruments (Alanazi, 2024; Norrbom & Zuboy, 2021).

Comparison with Previous Research

This pattern is consistent with Gou's (2023) argument that in e-learning ecosystems, level placement algorithms often do not disclose the measurement parameters they employ, resulting in a gap between assigned levels and users' actual competencies that is systemic rather than partial. The 31% misalignment identified in this study can thus be viewed as a concrete empirical manifestation of a phenomenon previously discussed only at a theoretical level. In this sense, the present study makes an original contribution by providing task-level evidence for a problem that prior literature has primarily addressed through policy-level analysis. Furthermore, Newton et al. (2021) emphasize the importance

of evidence-centered design in language assessment, and these findings suggest that the A1 level in AlifBee appears to be determined more by the sequencing of content than by evidence-based argumentation linking task demands explicitly to proficiency descriptors.

This cross-modality misalignment also has direct implications for learners' experiences. Cummings and Anderson (2025) highlight that inaccurate CEFR level placement may affect learner engagement and motivation. In the context of AlifBee, which serves a global audience, this issue becomes particularly important. Self-determination theory further supports this concern, as learners' motivation is closely related to their sense of autonomy and competence. When assessment levels do not accurately reflect learners' actual abilities, learners may experience reduced confidence and a weaker sense of control over their learning progress, which may in turn affect their long-term engagement with the platform (Ryan & Deci, 2000). These findings are also consistent with Dewi et al. (2025), who found that domains involving real communicative interaction tend to align more closely with CEFR than domains focused on form recognition. This pattern directly corresponds to the findings of the present study, in which the dialogue domain shows full alignment with CEFR A1 descriptors, while the vocabulary and composition domains show lower levels of consistency.

This study also differs from previous research in two significant ways. First, while most CEFR alignment studies in the Arabic context focus on formal curricula, textbooks, or institutional assessments (Musthofa, 2022; Pransiska et al., 2024; Alrababa'h et al., 2024), the present study fills a critical gap by examining assessment alignment in a mobile learning platform. Second, the empirically observed Pre-A1 range within the context of MSA suggests that CEFR A1 descriptors may require contextual interpretation when applied to typologically distant languages such as Arabic. This finding resonates with arguments in the broader language testing literature that CEFR descriptors, originally developed in the context of European languages, require systematic empirical validation when applied to typologically distant languages such as Arabic (Alderson et al., 2006; Khalifa & Weir, 2009).

Unlike studies such as Jeon (2025) on European languages, the complexity of MSA's orthographic, phonological, and root-and-pattern morphological systems places beginner non-Arab learners at a pre-communicative stage that is not fully captured within the existing CEFR scale, as also noted by Norrbom and Zuboy (2021) regarding the adaptation of CEFR descriptors into Arabic. This reinforces the perspective of Giraldo (2018) and Winke and Isbell (2017) that transparency of constructs and level verification mechanisms is a critical issue that cannot be overlooked when assessment operates within multilingual digital platforms such as AlifBee.

Theoretical and Practical Implications

This study contributes both theoretically and practically to the field of Arabic language assessment. Theoretically, it demonstrates that CEFR level progression within the context of Modern Standard Arabic (MSA) is not always linear and may include a range of competencies below A1 that can be empirically identified through task analysis. This finding highlights the importance of a construct-based approach in mapping language proficiency levels, particularly in technology-mediated learning environments.

Practically, the results provide important implications for developers of digital language learning applications to design assessment tasks that emphasize communicative language use rather than mere reproduction of linguistic forms. In addition, for non-native Arabic learners, the findings underline the significance of the Pre-A1 stage as a foundational phase before progressing to A1, ensuring a more gradual and pedagogically appropriate learning process aligned with learners' basic needs. Future research could extend this analysis to subsequent levels within AlifBee and to other digital Arabic learning platforms, in order to determine whether the patterns identified here are specific to the A1 unit or reflect broader structural tendencies in the field of technology-mediated Arabic language assessment.

5. CONCLUSION

This study concludes that while 69% of the A1 assessment items in the AlifBee application align with CEFR descriptors, a significant 31% are more accurately categorized at the Pre-A1 level, indicating a need for more consistent level representation. This misalignment is primarily driven by task demands that emphasize isolated linguistic forms such as vocabulary memorization and letter sequencing rather than contextual communicative performance, which is more prevalent in dialogue-based tasks. These findings highlight that CEFR alignment in digital platforms should be determined by the quality of communicative performance rather than mere thematic coverage. However, this research is limited by its focus on the initial A1 exercises and its reliance on document analysis without empirical user data. Therefore, future studies should expand the analysis to higher levels such as A2 and the application-specific A3 level, while incorporating field trials and developer perspectives to achieve a more comprehensive understanding of assessment validity in digital Arabic language learning.

Author Contributions:

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Alyaa Adhinna Putri	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓			
Asep Sopian				✓						✓		✓		
Hikmah Maulani				✓						✓		✓		

Penjelasan

C : **C**onceptualization
M :**M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing – **O**riginal Draft
E : Writing – Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

REFERENCES

Abu-Zhaya, R., & Arnon, I. (2024). Does Early Unit Size Impact the Formation of Linguistic Predictions? Grammatical Gender as a Case Study. *Language Learning*, *74*(4), 814–852. https://doi.org/10.1111/lang.12638

Alanazi, M. S. (2024). The use of Modern Standard Arabic and colloquial Arabic in translation tasks: A new perspective. *Cogent Arts & Humanities*, *11*(1), 2366572. https://doi.org/10.1080/23311983.2024.2366572

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

Alexiou, T., & Stathopoulou, M. (2021). The Pre-A1 Level in the Companion Volume of the Common European Framework of Reference for Languages. *Research Papers in Language Teaching and Learning*, *11*((1)), 11–29.

Alrababa'h, I. H., Habashneh, Q. Y., & Rababa, I. A. (2024). Assessing Reading Texts for Non-Native Arabic Speaking Students at the University of Jordan in Light of the Common European Framework of Reference for Languages From the Students' Perspective. *Theory and Practice in Language Studies*, *14*(6), 1818–1827. https://doi.org/10.17507/tpls.1406.23

Aslan, M. (2025). The Evolution of Language Education in The Digital. *Language In The Digital Age*, *118*.

Asli, N. F., Mohd Matore, M. E. E., & Md Yunus, M. (2024). Construct validity of primary trait writing rubrics based on assessment use argument (AUA) validation framework. *Heliyon*, *10*(22), e40053. https://doi.org/10.1016/j.heliyon.2024.e40053

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L. F., & Palmer, A. (2010). Language Assessment in Practice. *Oxford: Oxford University Press.*

Bar-On, A., Shalhoub-Awwad, Y., & Tuma-Basila, R. I. (2018). Contribution of phonological and morphological information in reading Arabic: A developmental perspective. *Applied Psycholinguistics*, *39*(6), 1253–1277. https://doi.org/10.1017/S0142716418000310

Benedetto, L., Gaudeau, G., Caines, A., & Buttery, P. (2025). Assessing how accurately large language models encode and apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence*, *8*, 100353. https://doi.org/10.1016/j.caeai.2024.100353

Bowen, G. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*, 27–40. https://doi.org/10.3316/QRJ0902027

British Council, EALTA, UKALTA, & ALTE. (2022). *Aligning Language Education with the CEFR: A Handbook*. British Council (joint publication). https://rm.coe.int/1680459f97

Chapelle, C. A., & Voss, E. (2017). Utilizing Technology in Language Assessment. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and Assessment* (pp. 149–161). Springer International Publishing. https://doi.org/10.1007/978-3-319-02261-1_10

Costantino, A., & Martin, S. (2025). Exploring sustainable language assessment through a teacher-learner practitioner inquiry. *System*, *134*, 103768. https://doi.org/10.1016/j.system.2025.103768

Council of Europe. (2020). *Common European Framework of Reference for Languages: Companion Volume*. Council of Europe Publishing.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry & research design: Choosing among five approaches* (Fourth edition). SAGE.

Cummings, D., & Anderson, P. (2025). The Impact of CEFR (Common European Framework of Reference for Languages) Level Adjustments on Student Engagement in Emirates School Establishments. *Gulf Education and Social Policy Review (GESPR)*, *6*(2), 130–145. https://doi.org/10.18502/gespr.v6i2.17685

Dewi, D. P., Ramadhani, G. P., & Sopian, A. (2025). Analyzing "Learn Arabic for Beginners" TikTok Content Based on the CEFR: تحليل محتوى تعليم اللغة العربية على تطبيق تيك توك في ضوء الإطار الأوربي المشترك للغات: "Learn Arabic for Beginners" أنموذجا. *LISANIA: Journal of Arabic Education and Literature*, *9*(1), 272–300. https://doi.org/10.18326/lisania.v9i1.272-300

Fulcher, G. (2014). *Testing second language speaking*. Routledge. https://doi.org/10.4324/9781315837376

Giraldo, F. (2018). Language Assessment Literacy: Implications for Language Teachers. *Profile: Issues in Teachers´ Professional Development*, *20*(1), 179–195. https://doi.org/10.15446/profile.v20n1.62089

Gou, P. (2023). Teaching english using mobile applications to improve academic performance and language proficiency of college students. *Education and Information Technologies*, *28*(12), 16935–16949. https://doi.org/10.1007/s10639-023-11864-9

Gunawan, I., & Palupi, A. R. (2016). Taksonomi Bloom – Revisi Ranah Kognitif: Kerangka Landasan Untuk Pembelajaran, Pengajaran, Dan Penilaian. *Premiere Educandum : Jurnal Pendidikan Dasar dan Pembelajaran*, *2*(02). https://doi.org/10.25273/pe.v2i02.50

Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press.

Jeon, J. (2025). The Impact of CEFR Basic User Level Text Complexity on Elementary School Learners' English Comprehension. *Primary English Education*, *31*(1). https://doi.org/10.25231/pee.2025.31.1.143

Jurāne-Brēmane, A. (2023). Digital Assessment in Technology-Enriched Education: Thematic Review. *Education Sciences*, *13*(5), 522. https://doi.org/10.3390/educsci13050522

Kamal, M., Sarip, M., Ilham, A., Jubaedah, S., & Khambali, K. (2025). Compiling E-Learning Kitabah Muqoyyadah Teaching Materials through the CEFR. *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, Dan Budaya Arab*, *8*(1), 21–35. https://doi.org/10.17509/alsuniyat.v8i1.73520

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Khezrlou, S., & Stockwell, G. (2025). A Synthetic Review of MALL Research: Artifact, Environment, and Task. *Digital Studies in Language and Literature*, *2*, 274–301. https://doi.org/10.1515/dsll-2024-0026

Lowie, W. M., Haines, K. B. J., & Jansma, P. N. (2010). Embedding the CEFR in the academic domain: Assessment of language tasks. *Procedia - Social and Behavioral Sciences*, *3*, 152–161. https://doi.org/10.1016/j.sbspro.2010.07.027

Marzuki, E., Ting, S.-H., Jerome, C., Chuah, K.-M., & Misieng, J. (2013). Congruence between Language Proficiency and Communicative Abilities. *Procedia - Social and Behavioral Sciences*, *97*, 448–453. https://doi.org/10.1016/j.sbspro.2013.10.258

Maulani, H., Muthmainah, N., Khalid, S. M., Saleh, N., & Taufik, I. H. (2024). Investigation of the Reference Level Description for Arabic Proficiency Tests in Indonesia. *Jurnal Al Bayan: Jurnal Jurusan Pendidikan Bahasa Arab*, *16*(1), 1. https://doi.org/10.24042/albayan.v16i1.21566

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (Third edition). SAGE Publications, Inc.

Muawanah, U., Marini, A., & Sarifah, I. (2024). The interconnection between digital literacy, artificial intelligence, and the use of E-learning applications in enhancing the sustainability of Regional Languages: Evidence from Indonesia. *Social Sciences & Humanities Open*, *10*, 101169. https://doi.org/10.1016/j.ssaho.2024.101169

Musthofa, T. (2022). CEFR-Based Policy in Arabic Language Teaching and Cultural Dimension in Indonesian Islamic Higher Education. *Eurasian Journal of Applied Linguistics*, *8*(2), 96–107.

Nagai, N., Birch, G. C., Bower, J. V., & Schmidt, M. G. (2020). *CEFR-informed Learning, Teaching and Assessment: A Practical Guide*. Springer Singapore. https://doi.org/10.1007/978-981-15-5894-8

Najiyah, S. A., Mahmudi, I., Muhammad Ismail, & Sa'diyah, L. F. (2026). Development of Arabic Reading Skills Test Items Based on Common European Framework of Reference for Languages Theory. *An Nabighoh*, *28*(1), 47–70. https://doi.org/10.32332/an-nabighoh.v28i1.47-70

Newton, S., Alemdar, M., Rutstein, D., Edwards, D., Helms, M., Hernandez, D., & Usselman, M. (2021). Utilizing Evidence-Centered Design to Develop Assessments: A High School Introductory Computer Science Course. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.695376

Norrbom, B., & Zuboy, J. (2021). Some Practical Consequences of Quality Issues in CEFR Translations: The Case of Arabic. In B. Lanteigne, C. Coombe, & J. D. Brown (Eds.), *Challenges in Language Testing Around the World: Insights for language test users* (pp. 421–432). Springer. https://doi.org/10.1007/978-981-33-4232-3_30

Norris, J. M. (2018). Task-Based Language Assessment Aligning Designs With Intended Uses and Consequences. *JLTA Journal*, *21*(0), 3–20. https://doi.org/10.20622/jltajournal.21.0_3

Oyebode, B. I., & Nicholls, N. (2021). Does the timing of assessment matter? Circadian mismatch and reflective processing in university students. *International Review of Economics Education*, *38*, 100226. https://doi.org/10.1016/j.iree.2021.100226

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist*, *51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Pransiska, T., Sugiyono, S., Widodo, S. A., & Sayyid, W. A. M. (2024). AL-KUTUB AL-MADRASIYAH AL-'ARABIYAH FĪ INDŪNISIYĀ MIN MAẒŪRI WAṬANĪYIN WA 'ĀLAMĪYIN. *Jurnal Ilmiah Islam Futura*, *24*(2), 488–512. https://doi.org/10.22373/jiif.v24i2.14964

Romero-Yesa, S., Fonseca, D., Aláez, M., & Amo-Filva, D. (2023). Qualitative assessment of a challenge-based learning and teamwork applied in electronics program. *Heliyon*, *9*(12), e22739. https://doi.org/10.1016/j.heliyon.2023.e22739

Ryan, R. M., & Deci, E. L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*, *55*(1), 68. https://doi.org/10.1037/0003-066X.55.1.68

Ryding, K. C. (2005). A Reference Grammar of Modern Standard Arabic. *Cambridge University Press*.

Salam, A. F. A., Firdaus, M. R., Amaliah, T., Azmy, M. U., & Hasanah, U. (2025). Analisis Pengembangan Kurikulum CEFR Bahasa Arab di Eropa dan Internasional. *Al-Hikmah: Jurnal Agama Dan Ilmu Pengetahuan*, *22*(2), 502–513. https://doi.org/10.25299/ajaip.2025.vol22(2).23113

Schmidgall, J., Oliveri, M. E., Duke, T., & Carter Grissom, E. (2019). Justifying the Construct Definition for a New Language Proficiency Assessment: The Redesigned TOEIC Bridge® Tests—Framework Paper. *ETS Research Report Series*, *2019*(1), 1–20. https://doi.org/10.1002/ets2.12267

Shehata, A. (2021). Short Vowels and Context Effects: The Case of English Speakers Reading Arabic. *International Education Studies*, *14*(8), 93. https://doi.org/10.5539/ies.v14n8p93

Shen, W., Xu, X., & Wang, X. (2022). Reconceptualising international academic mobility in the global knowledge system: Towards a new research agenda. *Higher Education*, *84*(6), 1317–1342. https://doi.org/10.1007/s10734-022-00931-8

Soliman, R., & Familiar, L. (2023). Creating a CEFR Arabic Vocabulary Profile: A frequency-based multi-dialectal approach. *Critical Multilingualism Studies*, *11*(1), 266–286.

Utomo, B. (2019). Analisis Validitas Isi Butir Soal sebagai Salah Satu Upaya Peningkatan Kualitas Pembelajaran di Madrasah Berbasis Nilai-Nilai Islam. *JURNAL PENDIDIKAN MATEMATIKA (KUDUS)*, *1*(2). https://doi.org/10.21043/jpm.v1i2.4883

Winke, P. M., & Isbell, D. R. (2017). Computer-Assisted Language Assessment. In *Language, Education and Technology* (pp. 1–13). Springer, Cham. https://doi.org/10.1007/978-3-319-02328-1_25-1

Wisniewski, K. (2018). The Empirical Validity of the Common European Framework of Reference Scales. An Exemplary Study for the Vocabulary and Fluency Scales in a Language Testing Context. *Applied Linguistics*, *39*(6), 933–959. https://doi.org/10.1093/applin/amw057

Xu, T. S., Zhang, L. J., & Gaffney, J. S. (2023). A multidimensional approach to assessing the effects of task complexity on L2 students' argumentative writing. *Assessing Writing*, *55*, 100690. https://doi.org/10.1016/j.asw.2022.100690

Zeng, J., & Huang, L. (2021). Understanding Formative Assessment Practice in the EFL Exam-Oriented Context: An Application of the Theory of Planned Behavior. *Frontiers in Psychology*, *12*, 774159. https://doi.org/10.3389/fpsyg.2021.774159