AN NABIGHOH

Journal homepage: http://e-journal.metrouniv.ac.id/an-nabighoh

# DEVELOPMENT OF ARABIC READING SKILLS TEST ITEMS BASED ON COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES THEORY

**Safira Aina Najiyah [a], Ihwan Mahmudi [b,*], Muhammad Ismail [c], Latif Fatus Sa'diyah [d]**

[a, b, c] Master of Arabic Teaching, Postgraduate, Universitas Darussalam Gontor, Ponorogo, Indonesia
[d] Department of Arabic Studies, Faculty of Islamic Studies, Al-Azhar University, Cairo, Egypt

## Article Info

***Corresponding Author:**
Name:
Ihwan Mahmudi
Email:
ihwanm@unida.gontor.ac.id

**Article History:**
Received: September 9, 2025
Revised: February 15, 2026
Accepted: March 26, 2026
Published: April 9, 2026

## Abstract

**Introduction:** The lack of standarized assessment instruments aligned with the CEFR remains a significant challenge in evaluating student's Arabic reading proficiency within the Indonesian educational context. In particular, there is a critical need for instruments that bridge local curricula with the Common European Framework of Reference for Languages (CEFR). **Research Objectives:** This study aims to develop valid and reliable Arabic reading skills test items based on the CEFR proficiency descriptors. **Methodology:** To address this gap, this research employed a Research and Development approach using the 4D model, which consists of define, design, develop, and disseminate stages. However, the implementation of this study was limited to the first three stages: defining needs, designing the test blueprint, and developing test items. The participants of this study were 259 fifth-grade students of Pondok Modern Darussalam Gontor Putri 1, selected using the Slovin formula. The initial blueprint comprised 119 test items distributed across six CEFR proficiency levels, ranging from A1 to C2. The test items were constructed in various formats, including multiple-choice, true-false, and matching questions. To ensure the quality of the instrument, expert validation and limited field trials were conducted. The data obtained were analyzed to determine content validity, empirical validity, reliability, item difficulty level, and discrimination power. **Results:** The results revealed that 72 items met the validity criteria and were suitable for use. The reliability analysis showed a coefficient of 0.99, indicating a very high level of reliability. Most of the valid items were categorized as easy, although several items required revision to improve their discrimination power and better differentiate between high- and low-performing students. **Unique Contribution:** This study contributes practically by providing Arabic language teachers in Indonesia with a standardized, CEFR-based assessment tool that can support accurate measurement of student's reading competencies and enhance the effectiveness of Arabic reading instruction. **Conclusion:** Overall, the findings demonstrate thet the developed instrument is valid, reliable, and appropiate for assessing student's reading profiency based on CEFR standards. **Recommendations:** Further research is recommended to implement the disseminate stage and expand the use of this instrument in broader contexts.



**Keywords:** Arabic Test; CEFR; Online Test; Reading Skills.

Please cite this article as Safira Aina Najiyah, Ihwan Mahmudi, Muhammad Ismail, and Latif Fatus Sa'diyah, "Development of Arabic Reading Skills Test Items Based on Common European Framework of Reference for Languages Theory," *An Nabighoh* 28, no. 1 (2026): 47-70, DOI: https://doi.org/10.32332/an-nabighoh.v28i1.47-70.

## Introduction

The availability of Arabic reading assessment instruments aligned with the Common European Framework of Reference for Languages (CEFR) in the Indonesian context remains limited. Most existing tests are not systematically linked to international proficiency standards, resulting in less accurate measurement of student's reading competence across levels. Therefore, a valid and reliable Arabic reading assessment instrument based on CEFR is urgently needed to ensure standarized and comparable evaluation of learner reading proficiency.

Reading skills (Maharah Qira'ah) are fundamental to Arabic language learning, enabling students to comprehend text and gain cultural insight.[1] Nevertheless, mastering these skills remains a significant challenge, particulary for non-native Arabic speakers in many educational contexts.[2] In the learning process, the evaluation of reading skills plays a strategic role in measuring students' comprehension levels, a view strongly supported by classical and contemporary figures in Arabic psychometrics.[3] However, the evaluation instruments currently used in Indonesia often lack reference to established international standards. Consequently, many existing reading skills tests fail to adequately meet the criteria for validity, reliability, and appropriate aligment with students' needs and competency levels, thereby necessitating the development of a standardized and reliable evaluation tool.[4]

In the digital era, the demand for accurate and standardized reading skills test aligned with international competency frameworks like the CEFR is increasingly urgent.[5] Such standardized instruments offer the advantage of providing consistent and comparable measurements across diverse learning contexts.[6] Nevertheless, the development of a CEFR-based Arabic reading evaluation instrument, complete with

[1] Buhori Muslim et al., 'Taṭwîr Kitāb Al-Qirā'at Al-Rasyîdah Li Tarqiyyah Mahārah Al-Qirā'ah 'Inda Al-Ṭālibah Bi Istikhdām Al-Kitāb Al-Elektrûny Al-Tafā'Uliy Fi Al-Madrasah Al-Mutawassiṭah Insān Qur'āny Aceh Besar', *Jurnal Ilmiah Islam Futura* 23, no. 2 (2023): 347, https://doi.org/10.22373/jiif.v23i2.19489.

[2] Hadjer Mokhtari, 'Asālīb Ta'līm Al-Lughah Al-'Arabiyah Li-Ghairi Al-Nāṭiqīn Bihā', *HuRuf Journal : International Journal of Arabic Applied Linguistic* 2, no. 2 (2023): 156, https://doi.org/10.30983/huruf.v2i2.5956.

[3] Maulia Yasminah Zakkiyah et al., 'Assessment Design and Analysis of Arabic Reading Skills Instructional Materials', *IJIE International Journal of Islamic Education* 3, no. 1 (2024): 31–46, https://doi.org/10.35719/ijie.v3i1.2000.

[4] Fathi Hidayah, 'Crosswalking as a Tool to Decide Arabic Language Standard in Madrasa Tsanawiyah: From Arabic Curriculum to ACTFL and CEFR', *International Conference on Humanity Education and Society (ICHES)* 3, no. 1 (2024), https://proceedingsiches.com/index.php/ojs/article/view/265.

[5] Lijun Shi et al., 'A Systematic Literature Review of Current Studies on Comparison Between the CEFR and CSE', *International Journal of Social Science Research* 12, no. 2 (2024): 18, https://doi.org/10.5296/ijssr.v12i2.21627.

[6] Mirdawati Razida et al., 'Blending Technology and Pedagogy: Optimizing Maharah al-Qirā'ah through the Alef Education Platform', *Al-Irfan : Journal of Arabic Literature and Islamic Studies* 8, no. 2 (2025): 211–25, https://doi.org/10.58223/al-irfan.v8i2.424.

rigorous validation and reliability testing, remains rare in the Indonesian context, which highlights the necessity of this study.[7]

The problem-solving approach offered in this study demonstrates clear advantages and novelty over previous research concerning the development of Arabic reading skills test, both domestically and internationally.[8] While previous studies have addressed this area, such as Halim and Alwi's research on Higher Order Thinking Skills (HOTS)-based tests and work developing CEFR-based materials for the B1 level, their scope remains limited. On a international scale, research focusing on the comprehensive development of a CEFR-aligned Arabic reading test instrument (A1-C2) through rigorous psychometric validation remains scarce.[9] Crucially, most existing studies, both local and international, have not specifically utilized the CEFR framework in its entirety (from A1-C2) nor have they developed the instruments in the form of a standardized, ready-to-use digital test.[10] This study, therefore, fills a significant gap by developing a comprehensive, validated, and reliable Arabic reading skills test instrument spanning all CEFR levels (A1-C2), representing a distinct contribution to standardized Arabic evaluation.

The novelty of this research lies in the application of CEFR frameworks which covers the full spectrum of proficiency levels, from A1 to C2, as well as through a comprehensive validation and evaluation process.[11] With this approach, the study provides a more structured, competency-based framework for test development compared to traditional assessments that refer exclusively to local needs without strong theoretical guidance.[12] In addition, this instrument is designed to be able to be used flexibly at different levels of learning, making it more adaptive compared to

[7] Umi Mahmudah and Tulus Musthofa, 'Reading Skills Learning in the "Arabic-Online.Net" Application by Saudi Electronic University Based on the Common European Framework of Reference for Languages (CEFR)', *Scaffolding: Jurnal Pendidikan Islam Dan Multikulturalisme* 5, no. 3 (2023): 370–85, https://doi.org/10.37680/scaffolding.v5i3.3377.

[8] Reşat Alatlı et al., 'Examination of the Reading Comprehension Skills of Good and Poor Readers in the Dimension of Reading Components Developed by a Reading Skills Assessment Tool', *Education and Science* 47, no. 211 (2022): 273–95, https://doi.org/10.15390/EB.2022.11080.

[9] Wan Alisa Hanis Wan Abdul Halim and Nik Aloesnita Nik Mohd Alwi, 'Cefr-Aligned Language Tests: A Systematic Scoping Review', SSRN Scholarly Paper no. 4244716 (Social Science Research Network, 11 October 2022), https://papers.ssrn.com/abstract=4244716.

[10] Nik Aloesnita binti Nik Mohd Alwi and Wan Alisa Hanis binti Wan Abdul Halim, 'Variations and Methodological Components in CEFR-Aligned Language Tests: A Systematic Review', *Journal of Creative Practices in Language Learning and Teaching* 12, no. 1 (2024), https://journal.uitm.edu.my/ojs/index.php/CPLT/article/view/2708.

[11] Habibur Rohman and Faiq Ilham Rosyadi, 'Development of Arabic Teaching Materials Based on the Common European Framework of Reference (CEFR) to Improve Students' Arabic Language Skills', *Al Mahāra: Jurnal Pendidikan Bahasa Arab* 7, no. 2 (2021): 163–83, https://doi.org/10.14421/almahara.2021.072-01.

[12] Erfan Gazali and Hasan Saefuloh, 'Development of an Arabic Receptive Proficiency Test Instrument Based on the Common European Framework of Reference for Languages', *Al-Ta'rib : Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab IAIN Palangka Raya* 11, no. 2 (2023): 293–308, https://doi.org/10.23971/altarib.v11i2.6721.

other evaluation instruments that tend to be uniform and less contextual.[13] For teachers, the instrument allows for precise diagnostic assessment, enabling them to accurately map each student's current proficiency and tailor instructional materials to specific CEFR levels.[14] For students, this standardized assessment provides clear, measurable learning goals, making their progress transparent and globally comparable, which ultimately enhances the quality and effectiveness of Arabic language learning in Indonesia.

The participants in this study consisted of 5th-grade students at Pondok Modern Darussalam Gontor Putri 1. Grade 5 was chosen because they generally have maturity in Arabic language skills, including the ability to read texts critically and analytically. However, preliminary observations indicated that some students still had difficulty in analyzing Arabic texts in depth. Therefore, the purpose of this research is not to replace the learning methods that have been used, but to measure the mastery of reading skills (Maharah Qira'ah) that have been learned and provide standardized evaluation instruments. Based on this description, the formulation of the problem in this study is: 1) What is the process for preparing CEFR-based Arabic reading skills test items? 2) What are the results of the validity and reliability of the CEFR-based Arabic reading skills test items?

The purpose of this research is to develop a valid and reliable CEFR-based Arabic reading skills test item. The theoretical and practical implications of this study are expected to assist teachers in providing standardized evaluation instruments, become a reference for researchers in the development of Arabic language tests, and provide useful measurement tools for Arabic learners to find out their level of reading ability.

This study contributes to the field of Arabic Language education by developing and empirically validating an Arabic reading test based on CEFR Theory. It provides a novel standarized assessment model that integrates international proficiency benchmarks with local educational needs, thereby expanding the research area of CEFR based assessment development in Arabic language learning.

## Method

This research employs the Research and Development (R&D) method utilizing the 4D model developed by Thiagarajan, which comprises four stages: define, design, develop, and disseminate.[15] However, the scope of this article is intentionally limited to discussing the first three initial stages (define, design, and develop).[16] This focus is necessary because the primary objective of this specific publication is to

[13] Florence Martin et al., 'Systematic Review of Adaptive Learning Research Designs, Context, Strategies, and Technologies from 2009 to 2018', *Educational Technology Research and Development* 68, no. 4 (2020): 1903–29, https://doi.org/10.1007/s11423-020-09793-2.

[14] Noriko Nagai et al., 'Integrating Learning, Teaching, and Assessment', in *CEFR-Informed Learning, Teaching and Assessment*, Springer Texts in Education (Springer Singapore, 2020), https://doi.org/10.1007/978-981-15-5894-8_5.

[15] Rosita Budi Indaryanti et al., '4D Research and Development Model: Trends, Challenges, and Opportunities Review', *Jurnal Kajian Ilmiah* 25, no. 1 (2025): 91–98, https://doi.org/10.31599/na7deq07.

[16] Jan Van Den Akker et al., eds, *Educational Design Research* (Routledge, 2006), https://doi.org/10.4324/9780203088364.

detail the preparation, comprehensive psychometric validation, and reliability testing of the CEFR-based Arabic reading skills test instrument. The dissemination stage (including large-scale testing and implementation) falls outside the scope of this foundational instrument development paper and will be addressed in a subsequent publication.

The define stage involved a multi-faceted needs analysis to identify pedagogical gaps and evaluate extant reading assessment practices.[17] This analysis was carried out through a literature study related to Arabic reading skills and the application of CEFR in the development of test instruments, interviews with Arabic teachers to explore students' linguistic difficulties in understanding and analyzing texts, and observation of learning in grade 5 of Pondok Modern Darussalam Gontor Putri 1. The results of the analysis were used to identify the gap between the test instruments that have been used and the learning needs, as well as to establish the basic framework for the development of test items based on the six levels of CEFR (A1–C2).

The design stage includes the design of test instruments starting from the preparation of a blueprint or question grid that contains the CEFR level, reading ability indicators, and the proportion of the number of questions at each level.[18] Furthermore, text genres–including narrative, descriptive, expository, and argumentative formats–were aligned with the specific functional requirements of each CEFR level. In addition, an assessment rubric was prepared that referred to the reading skill indicators at each level. The initial design of this instrument was then validated by Arabic linguists and learning evaluation experts to ensure the suitability of the content, clarity of indicators, and measurability of the question items.[19]

The development stage consisted of writing test items according to the validated blueprint, followed by expert validation using Aiken's V or Content Validity Ratio (CVR) to assess the suitability of the content of the test items.[20] After that, an initial trial was carried out on 30 female students to measure the empirical validity, reliability, level of difficulty, and differentiation of the question items. Reliability was calculated using KR-20 or Alpha Cronbach, and the results of the analysis were used to revise the question items to meet the criteria of a standardized instrument.

---

[17] Chadia Haddad et al., 'Validation of the Arabic Version of the "12-Item Short-Form Health Survey" (SF-12) in a Sample of Lebanese Adults', *Archives of Public Health* 79, no. 1 (2021): 56, https://doi.org/10.1186/s13690-021-00579-3.

[18] J. Charles Alderson et al., '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, no. 1 (2006): 3–30, https://doi.org/10.1207/s15434311laq0301_2.

[19] Ahmad A. Alharbi et al., 'Cross-Cultural Adaptation and Psychometric Properties of the Arabic Version of the Fall Risk Questionnaire', *International Journal of Environmental Research and Public Health* 20, no. 8 (2023): 5606, https://doi.org/10.3390/ijerph20085606.

[20] Lewis R. Aiken, 'Three Coefficients for Analyzing the Reliability and Validity of Ratings', *Educational and Psychological Measurement* 45, no. 1 (1985): 131–42, https://doi.org/10.1177/0013164485451012.

The population of this study was all 5th grade students at Pondok Modern Darussalam Gontor Putri 1 which totals 733 people. The research sample was determined using the Slovin formula with a margin of error of 5%, yielding a sample size of 259 students. The initial test for the validity of test items was carried out on 30 5th grade students who were randomly selected (random sampling) from the population. The inclusion criteria for participants specified students with advanced-level Arabic reading instruction, including those who encounter difficulties in critical text analysis.

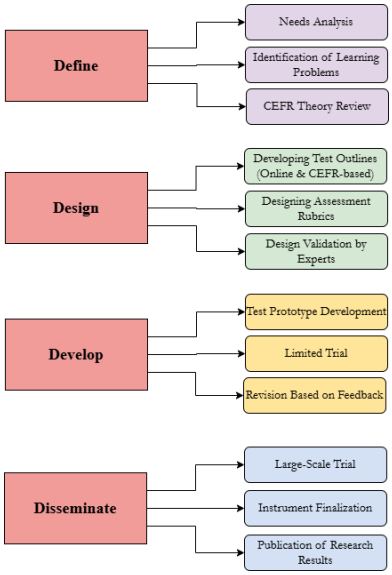


**Figure 1.** Research stages for this study

The instruments used in this study include expert validation sheets to assess the suitability of test items with CEFR indicators, as well as reading skills test kits consisting of reading texts and question items according to six CEFR levels. The data from expert validation results is analyzed using Aiken's V or Content Validity Ratio (CVR) to determine the validity of the content. The data from the reliability test was analyzed using KR-20 or Alpha Cronbach, and the difficulty level and differentiation of the question items were analyzed to ensure the overall quality of the instrument.

## Result and Discussion

### Test Blueprint

The initial blueprint for the Arabic reading assessment consisted of 119 items spanning six CEFR levels (A1–C2). This initial design was then validated by experts to assess the suitability of the content and the reading skill indicators being measured. Items deemed invalid were revised or deleted, and then a limited pilot test was conducted on 30 female students to assess empirical validity, reliability, difficulty level, and discriminatory power. Based on the synthesis of the expert validation and the limited pilot test, the number of items used in the final stage of the study was reduced to 43.

The substantial reduction from 119 items to 43 items indicates that the validation and pilot testing stages played a crucial role in ensuring the quality of thee

instrument. This process demonstrates that the remaining items possess high discriminatory power and structural validity, adhering to the fundamental principles of standardized psychometric evaluation.

**Table 1.** Distribution of CEFR Levels and Question Types

| CEFR Level | Number of Questions | Question Form |
|---|---|---|
| A1 | 21 | multiple choice, true-false, matching |
| A2 | 24 | multiple choice, true-false, matching |
| B1 | 18 | multiple choice, true-false, matching |
| B2 | 29 | multiple choice, true-false, matching |
| C1 | 12 | multiple choice, true-false, matching |
| C2 | 13 | multiple choice, true-false, matching |

To provide a concrete overview of the developed instruments, the following is an example of reading skill questions for each CEFR level (A1-C2).[21] The examples shown are representative selected from the initial blueprint that has gone through an expert validation process. Each question item is calibrated according to the indicators of reading competence at each CEFR level, from the ability to understand explit information in simple text to the ability to analyze and evaluate arguments in complex text.[22]

**Example of Test Items per CEFR Level**

***Level A1***

At the A1 proficiency level of language proficiency, learners possess an elementary understanding of language that enables them to comprehend very short and simple texts. The CEFR designates this level as one characterized by the ability to recognize familiar names, high-frequency words, and basic phrases, allowing learners to interpret texts with aid from context. This notion is supported by research indicating that A1-level learners often struggle with comprehension unless the texts are both familiar and contextually relevant.[23]

**Table 2.** Sample Test Item of Arabic Reading Skill at A1 Level

| **Question** | **Reading Text** |
|---|---|
| ما الطريق إلى قسم المدرسة؟<br>أ. رقم اثنين ج. رقم أربعة<br>ب. رقم واحد د. رقم ثلاثة | انظر إلى هذه اللوحة ثمّ أجبي!<br>  |

[21] Yan Li et al., 'Validating a Reading Assessment Within the Cognitive Diagnostic Assessment Framework: Q-Matrix Construction and Model Comparisons for Different Primary Grades', *Frontiers in Psychology* 12 (December 2021): 786612, https://doi.org/10.3389/fpsyg.2021.786612.

[22] Ally Oi Kuan Ho et al., 'An Analysis of Macau's Joint Admission Examination–English', *The Journal of AsiaTEFL* 18, no. 1 (2021): 208–22, https://doi.org/10.18823/asiatefl.2021.18.1.12.208.

[23] Glenna Rose Santuya, 'Learners' Level of Reading Comprehension: Basis for Contextualized Reading Materials', *Pantao (International Journal of the Humanities and Social Sciences)* 4, no. 2 (2025), https://doi.org/10.69651/PIJHSS040292.

This item measures learner's ability to extract explicit and direct information from short and simple texts, which corresponds to the CEFR A1 reading descriptor. The correct answer is option (D).

This test item reflects the CEFR A1 reading descriptor, which emphasizes learner's ability to understand familiar words and basic information from simple texts. The use of visual cues (the sign) in the task helps learners identify explicit information without requiring complex linguistic processing. Therefore, the item is appropriate for measuring beginner-level reading comprehension.

***Level A2***

At the A2 proficiency level, learners can manage short and simple texts that relate to familiar and concrete topics, such as everyday life and work-related scenarios. As per CEFR, this stage serves as a foundation for expanding vocabulary and understanding more complex sentence structures compared to A1-level learners. Learners at this stage begin to navigate typical daily situations, such as shopping or making appointments, using basic language principles.[24]

**Table 3.** Sample Test Item of Arabic Reading Skill at A2 Level

| **Reading Text** |
|---|
| **اقرأ النصّ الآتي، ثم أجب عن الأسئلة باختيار الجواب الصحيح!** |
| زارَ وزيرُ التعليمِ مدينةَ نابلس في الضفة الغربية يومَ الثلاثاء، واجتمعَ مع عددٍ من الطلابِ والمعلمينِ في المدارسِ الحكوميةِ. تحدثَ الوزيرُ عن أهميةِ التعليمِ في بناءِ مستقبلِ الشباب، ووعدَ بدعمِ المدارسِ الجديدةِ وتحسينِ المرافقِ التعليميةِ. بعدَ الاجتماعِ، زارَ الوزيرُ مكتبةَ المدينةِ، وقرأَ كتابًا مع الأطفالِ الصغارِ. |
| **Question** |
| **متى زارَ الوزيرُ مدينةَ نابلس؟** |
| أ. يوم الإثنين |
| ب. يوم الثلاثاء |
| ج. يوم الأربعاء |

This item measures learner's ability to understand explicit factual information and spesific details in short, non-fiction texts, such as news or announcements, in line with the CEFR A2 reading descriptor. The correct answer is (B) "يوم الثلاثاء" beacuse the text explicity states that the minister visited the city on Tuesday. Thus, learners are required to identify clearly stated temporal information from a straightforward factual text.

***Level B1***

At the B1 proficiency level, learners are increasingly capable of engaging with simple factual texts relevant to their fields of interest or everyday situations. The CEFR characterizes B1 learners as those who can understand texts primarily composed of high-frequency everyday and job-related language, signaling a

[24] Kendy Suet Cabahug, 'The Implementation of a Validated Contextualized Reading Material to Enhance Decoding Skills of Grade 1 Learners', *Pantao (International Journal of the Humanities and Social Sciences)* 4, no. 2 (2025), https://doi.org/10.69651/PIJHSS040285.

significant progression from earlier levels.[25] This capability allows learners at B1 to interpret information about events, feelings, and wishes in more contextual and personalized texts.

**Table 4.** Sample Test Item of Arabic Reading Skill at B1 Level

| **Reading Text** |
|---|
| **اقرأ رسالة البريد الإلكتروني التالية:**<br>الموضوع: طلب التسجيل في دورة تدريبية<br>السّلام عليكم ورحمة الله وبركاته،<br>تحية طيبة وبعد...<br>أنا الطالبة فاطمة أحمد الهندي، أدرس حاليًّا في قسم تعليم اللغة العربية لغير الناطقين بها، في المرحلة الجامعية. أكتب إليكم هذه الرسالة لأعبّر عن رغبتي في التسجيل في الدورة التدريبية التي ينظّمها مركزكم المحترم، والتي تحمل عنوان: "استراتيجيات تعليم اللغة العربية لغير الناطقين بها باستخدام الوسائل التقنية الحديثة".<br>لقد قرأتُ الإعلان عن هذه الدورة عبر الموقع الرسمي، ولاحظت أن موضوعها يتوافق تمامًا مع اهتماماتي الأكاديمية وتخصصي الحالي. أرغب بشدة في تطوير مهاراتي التعليمية، والتعرّف على أساليب جديدة في توظيف التقنية داخل الصف، مثل استخدام المنصات الإلكترونية، والأدوات التفاعلية، وتطبيقات التعليم الذكي.<br>أعلم أن هذه الدورة ستعزز قدراتي وتمكنني من المشاركة الفاعلة في تعليم اللغة العربية مستقبلاً، خاصة في بيئات متعددة الثقافات. أرجو التكرم بقبول طلبي هذا، وإعلامي بالشروط المطلوبة، والمواعيد المحددة للتسجيل والدفع، إن وُجد. كما أرفق مع هذه الرسالة نسخة من بطاقة الطالب وصورة شخصية، حسب ما طُلب في الإعلان.<br>شاكرة لكم وقتكم واهتمامكم،<br>وتفضلوا بقبول فائق الاحترام والتقدير.<br>مع خالص التحية،<br>**فاطمة أحمد الهندي** |
| **Question** |
| **اختري المعلومة المفهومة من هذه الرسالة!**<br>أ. الطالبة فاطمة ترغب في المشاركة في مسابقة أدبية<br>ب. الطالبة فاطمة تسأل عن شروط السفر إلى الخارج<br>ج. الطالبة فاطمة تريد التسجيل في دورة تدريبية عن استخدام التقنية في تعليم اللغة العربية<br>د. الطالبة فاطمة تعمل في مركز تدريب لغات |

This item measures the ability to discern the main communicative purpose and specific details within a functional text. At this stage, learners must distinguish between the writer's actual intention and similar but unsupported distractors. At this level, learners are expected to grasp both explicit details and the overall communicative purpose of the text, while being able to differentiate closely related ideas to avoid misinterpretation.

[25] David Neal et al., 'Read and Accepted? Scoping the Cognitive Accessibility of Privacy Policies of Health Apps and Websites in Three European Countries', *DIGITAL HEALTH* 9 (January 2023): 20552076231152162, https://doi.org/10.1177/20552076231152162.

The correct answer is option (C) " الطالبة فاطمة تريد التسجيل في دورة التقنية في تعليم اللغة العربية." Because the messege explicity states Fatimah's intention to enroll in a training course on using technology in teaching Arabic. This requires learners to identify the writer's main communicative purpose and distinguish it from other similar but unsupported options, in line with CEFR level B1 reading competence. This task aligns with the B1 descriptor of identifying global meaning and specific information in formal correspondence.

***Level B2***

At the B2 proficiency level, learners exhibit a heightened degree of autonomy in their reading abilities. The CEFR indicates that these learners can adapt their reading styles and speeds to accommodate a diverse array of texts and purposes, utilizing appropriate reference sources selectively.[26] This capability represents a significant advancement from the previous B1 level, where learners had more limited autonomy in text engagement.

**Table 5.** Sample Test Item of Arabic Reading Skill at B2 Level

| **Reading Text** | |
|---|---|
| **أكمل الفراغ بالكلمة المناسبة من النص!** | |
| يُشير مصطلح "جيل زد" إلى الأفراد المولودين بين منتصف_____ وأواخر العقد الأول من الألفية الثالثة، ويُعرف عنهم اهتمامهم الشديد بالمظهر كوسيلة أساسية للتعبير عن الذات، خاصة مع نشأتهم في ظل ثقافة تركز على الصورة التي تعتبر جزءا أساسيا من_____ في عصر وسائل التواصل الاجتماعي. لكن اللافت للنظر هو شيوع الملابس القبيحة أو ما يطلق عليه Uglycore وكذلك الملابس غير المتناسقة بمقاييس الأجيال السابقة.<br>لكن ما يعتبره البعض فسادا في الذوق يراه أبناء جيل زد موقفا متمردا على الجماليات التقليدية السائدة التي يفرضها المجتمع، وهو اتجاه يعرف ب"Anti-Aesthetic"، أو المناهضة للجمالية، وهو تيار في الفن والموضة يرى أن_____ ليست شرطا للقيمة.<br>وغالبا ما يرتبط بالموضة المضادة، حيث يستخدم ملابس وأساليب غير تقليدية وغير متطابقة لتحويلها إلى شكل جريء من أشكال التعبير عن الذات والاحتفاء بالفردية والتمرد على الذوق المهيمن. كما يتجه الكثير من أبناء جيل زد لاستخدام _____ و_____ في اتجاه يربط الجمالية<br>بالممارسات الأخلاقية والمستدامة. | |
| **Answers** | |
| أ. التسعينيات | د. الهوية الرقمية |
| ب. الجاذبية البصرية | هـ. الملابس المستعملة |
| ج. المعاد تدويرها | |

B2 CEFR Indicator: At this level, learners can read with a high degree of independence, adapting reading style and speed to a variety of text and purposes, and using appropriate reference sources selectively. They utilize wide active reading vocabulary, but may have some difficulty with less commonly used idioms.

[26] Monika M. Połczyńska and Susan Y. Bookheimer, 'General Principles Governing the Amount of Neuroanatomical Overlap between Languages in Bilinguals', *Neuroscience & Biobehavioral Reviews* 130 (November 2021): 1–14, https://doi.org/10.1016/j.neubiorev.2021.08.005.

The item measures the ability to distinguish between nuanced meanings of words according to context, and the connect a word to its implied meaning rather than relying solely on its literal translation. This alighs with the CEFR B2 descriptor. Has a broad active reading vocabulary but many experience some difficulty with low-frequency idioms. At this level, leaarnes are expected to interpret lexical measing beyond surface-level definitions, discerning subtle difference in usage and recognizing implied or figurative meanings within complex texts.

***Level C1***

At the C1 proficiency level, learners can navigate multifaceted texts, employing advanced reading strategies that go beyond basic comprehension. Duke and Cartwright emphasize the relevance of new theories and models in cognitive processes that underlie reading, suggesting that comprehension at this level benefits from a robust understanding of the reading process, guiding practitioners in supporting students' reading development.[27]

**Table 6.** Sample Test Item of Arabic Reading Skill at C1 Level

| **Reading Text** | |
|---|---|
| **بوريل: مرتزقة أميركيون قتلوا ٥٥٠ فلسطينيا في غزة خلال شهر** | |
| قال المسؤول السابق للسياسة الخارجية والأمن للاتحاد الأوروبي جوزيب بوريل إن "مرتزقة أميركيين" قتلوا ٥٥٠ فلسطينيا في غزة خلال شهر واحد، متهما مجلس أوروبا والمفوضية الأوروبية بالتزام الصمت إزاء هذه الأحداث. وذكر في منشور عبر حسابه على منصة إكس أمس الخميس "خلال شهر واحد قُتل ٥٥٠ فلسطينيا يعانون من الجوع على أيدي مرتزقة أميركيين بينما كانوا يحاولون الحصول على الغذاء عند نقاط التوزيع التي حددتها مؤسسة غزة الإنسانية" المدعومة من الولايات المتحدة وإسرائيل.وصف بوريل هذا الفعل بأنه "مروع"، وأرفق حسابي مجلس أوروبا والمفوضية الأوروبية في المنشور، متهما إياهما بعدم الرغبة في التحرك "ضد الجرائم التي ترتكب في غزة". وبعيدا عن إشراف الأمم المتحدة والمنظمات الدولية، بدأت تل أبيب وواشنطن منذ ٢٧ مايو/أيار الماضي تنفيذ خطة لتوزيع مساعدات محدودة عبر ما تُعرف بـ"مؤسسة غزة الإنسانية"، حيث يُجبر الفلسطينيون المجوّعون على المفاضلة بين الموت جوعا أو برصاص الجيش الإسرائيلي. وخلفت هذه الآلية التي باتت تُعرف بـ"مصائد الموت"، حتى ظهر أمس الخميس ٦٥٢ شهيدا وأكثر من ٤ آلاف و٥٣٧ إصابة، وفق آخر تحديث لوزارة الصحة بقطاع غزة. يذكر أن السياسي الإسباني بوريل قد تبنى موقفا وخطابا مختلفا عن إدارة الاتحاد الأوروبي تجاه إسرائيل منذ بدئها حرب الإبادة الجماعية على قطاع غزة في ٧ أكتوبر/تشرين الأول ٢٠٢٣. وكان بوريل قد حاول توحيد موقف الدول الأعضاء في الاتحاد الأوروبي لإدانة انتهاكات إسرائيل للقانون الدولي والدعوة لوقف إطلاق النار، كما انتقد بشكل لاذع المفوضية الأوروبية برئاسة أورسولا فون دير لاين. وقد سلم بوريل منصبه إلى خليفته كايا كالاس في الأول من ديسمبر/كانون الأول ٢٠٢٤. ومنذ ٧ أكتوبر/تشرين الأول ٢٠٢٣، ترتكب إسرائيل بدعم أميركي إبادة جماعية بقطاع غزة، تشمل قتلا وتجويعا وتدميرا وتهجيرا، متجاهلة النداءات الدولية وأوامر لمحكمة العدل الدولية بوقفها. | |
| **Question** | |
| **ما الهدف من خطاب جوزيب بوريل في منشوره؟** | |
| أ. إعلان استقالته من منصبه الرسمي | ج. تحميل الاتحاد الأوروبي مسؤولية الصمت تجاه الجرائم |
| ب. الدفاع عن مؤسسة غزة الإنسانية | د. دعوة الأمم المتحدة لتوسيع المساعدات |

[27] Nell K. Duke and Kelly B. Cartwright, 'The Science of Reading Progresses: Communicating Advances Beyond the Simple View of Reading', *Reading Research Quarterly* 56, no. S1 (2021), https://doi.org/10.1002/rrq.411.

This item measures the ability to distinguish between a personal stance and an institutional stance, to understand journalistic metaphors such as "death traps" and relate them to real-world context, and to extract implicit criticism directed at European Institutions. This alighs with the CEFR C1 descriptor. At this level, learners are expected to interpret layered meanings in sophisticated texts, recognize rhetorical devices and figurative language, and critically infer underlying attitudes or critiques embedded within the discourse.

The correct answer is option (C) "تحميل الاتحاد الأوروبي مسؤولية الصمت تجاه الجرائم" because the speech explicity criticizes the European Union for its silince regarding the crimes mentioned in the text. This indicates an implicit critical stance directed a tan institutional actor, which requires learners to interpret evaluative language and underlying criticism, consistent with CEFR level C1 reading competence.

***Level C2***

At the C2 proficiency level, learners possess the ability to decipher intricate texts across diverse genres, including both literary and non-literary works. This degree of comprehension hinges on a sophisticated understanding of language and context. As Li et al. suggest, automated scoring systems, while proficient at evaluating basic linguistic dimensions, struggle to assess more complex advanced writing features, such as argument complexity and clarity.[28] This underscores the significance of human evaluation for high-level texts, as it allows for an appreciation of nuanced arguments and stylistic subtleties that are often lost in machine assessments. The reliability of human insight becomes critical when dealing with the abstract and complex structures that C2 learners encounter.

Furthermore, the ability of C2 learners to critically interpret texts requires not just comprehension but also the application of analytical skills.[29] Maqsood and Anbreen analyze the vocabulary utilized by the CEFR-level writers, revealing how nuanced engagement with vocabulary aligns with proficiency levels.[30] This finding highlights the meticulous attention to language forms that C2 learners must exhibit as they interpret and construct meaning from complex text sources.[31]

[28] Changlin Li et al., 'A Comparative Review of the CEFR and CET4 Writing Assessment with Insights from Task Complexity Theories', *Malaysian Journal of Social Sciences and Humanities (MJSSH)* 10, no. 3 (2025): e003251, https://doi.org/10.47405/mjssh.v10i3.3251.

[29] Özlem Koray and Sercan Çetinkılıç, 'The Use of Critical Reading in Understanding Scientific Texts on Academic Performance and Problem-Solving Skills', *Science Education International* 31, no. 4 (2020), https://www.icaseonline.net/journal/index.php/sei/article/view/239.

[30] Ammara Maqsood and Tanzeela Anbreen, 'Bridging Proficiency and Practice: Aligning Lexical Bundles in Pakistani L2 Learner Writing with CEFR Descriptors', *Journal of Applied Linguistics and TESOL (JALT)* 8, no. 4 (2025), https://jalt.com.pk/index.php/jalt/article/view/1646.

[31] Nuntapat Supunya, 'Towards the CEFR Action-Oriented Approach: Factors Influencing Its Achievement in Thai EFL Classrooms', *3L The Southeast Asian Journal of English Language Studies* 28, no. 2 (2022): 33–48, https://doi.org/10.17576/3L-2022-2802-03.

**Table 7.** Sample Test Item of Arabic Reading Skill at C2 Level

| **Reading Text** | |
|---|---|
| **السعودية وإندونيسيا توقعان صفقات بقيمة ٢٧ مليار دولار** | |
| أبرمت السعودية وإندونيسيا أمس الأربعاء اتفاقيات عدة ومذكرات تفاهم بقيمة إجمالية تبلغ ٢٧ مليار دولا، وذلك خلال زيارة للرئيس الإندونيسي برابوو سوبيانتو إلى السعودية التقى خلالها ولي العهد ورئيس الوزراء محمد بن سلمان. وأكد قادة البلدين الالتزام بتعزيز الشراكة الإستراتيجية في مختلف المجالات، من الطاقة والاقتصاد الرقمي إلى الاستثمارات الخضراء، والخدمات المالية وتطوير الصناعات واللوجستيات والسياحة. كذلك اتفق البلدان على تسريع إبرام اتفاقية التجارة الحرة بين إندونيسيا ومجلس التعاون الخليجي لزيادة حجم التجارة وتدفق الاستثمارات، وجاء في بيان البلدين أنهما "يرحبان بالنتائج الإيجابية لمفاوضات اتفاقية التجارة الحرة بين دول مجلس التعاون الخليجي وإندونيسيا، ويأملان إتمام هذه الاتفاقية في أقرب وقت ممكن". "واتفق الطرفان على تشجيع الابتكار والتكنولوجيا الجديدة مثل الذكاء الاصطناعي في قطاع الطاقة، وتطوير تكنولوجيا الاقتصاد الدائري للكربون والهيدروجين النظيف"، حسب البيان الصادر عن الرئاسة الإندونيسية. وذكرت وكالة الأنباء السعودية أن حجم التبادل التجاري بين البلدين قارب ٣١،٥ مليار دولار خلال السنوات الخمس الماضية، مشيرة إلى أن هذا "يجعل المملكة الشريك التجاري الأول لجمهورية إندونيسيا في المنطقة". ووقعت شركة أكوا باور السعودية اتفاقيات مبدئية لاستكشاف فرص الاستثمار في مشاريع الطاقة المتجددة مع "دانانتارا إندونيسيا" وهو صندوق ثروة سيادي، وشركة الطاقة الحكومية برتامينا، بحسب بيان من دانانتارا. وأضا صندوق دانانتارا السيادي أنه من المتوقع استكشاف استثمارات محتملة تصل قيمتها إلى ١٠ مليارات دولار لتمويل المشاريع. | |
| **Question** | |
| ما الهدف الرئيس من توقيع الاتفاقيات بين السعودية وإندونيسسيا كما ورد في النص؟ | |
| أ. التوقيع على اتفاقيات أمنية وعسكرية | ج. افتتاح منشآت استثمارية جديدة في العاصمة السعودية |
| ب. تعزيز الشراكة الإستراتيجية وتوقيع اتفاقيات اقتصادية متنوعة | د. مناقشة التعاون في المجال الصحي فقط |

The item measures the ability to analyze complex texts of a formal and economic nature, infer relationships between ideas, and identify positions and policies based on information presented within a global context. At this level, learnes are expected not only to grasp explit content but also to critically evaluate nuances arguments, connect abstract ideas across domains, and interpret the broader implications of the text in relation to global contexts.

The correct answer is option (B) "تعزيز الشراكة الإستراتيجية وتوقيع اتفاقيات اقتصادية متنوعة" because the text explicity states that the agreements aim to strengthen the strategic partnership and include various economic cooperation sectors. This requires learners to synthesize key information across the text and infer the overarching policy objective, reflecting advanced reading skills aligned with higher-level CEFR C2 reading competence.

## Validity of Test Items

To determine the extent to which the Arabic reading skills test items can measure the intended ability, an empirical validity test was conducted using item-total correlation (Point-Biserial).[32] A limited trial was conducted with 30 fifth-grade female students at the Modern Islamic Boarding School Darussalam Gontor for Girls 1st Campus. The decision criteria stipulated that an item is considered valid if the

[32] Elsa Albero-Ros et al., 'Development and Initial Validation of the MCL-PRO-CAT: A Computerized Adaptive Test Designed to Measure Multifocal Contact Lens Performance from the Patient's Perspective', *Contact Lens and Anterior Eye* 48, no. 3 (2025): 102378, https://doi.org/10.1016/j.clae.2025.102378.

calculated r-value exceeds the critical r-table value at a 5% significance level (α= 0.05). The results of the item validity test can be seen in Table 8.

**Table 8.** Item Validity Test

| CEFR Level | Total Items | Valid Items | Invalid Items |
|---|---|---|---|
| A1 | 21 | 18 | 3 |
| A2 | 24 | 14 | 10 |
| B1 | 18 | 10 | 8 |
| B2 | 29 | 17 | 12 |
| C1 | 12 | 7 | 5 |
| C2 | 13 | 5 | 8 |

Based on Table 8, the psychometric evaluation reveals that only 72 out of the 119 developed items met the validity criteria, resulting in 47 invalid items. While lower levels (A1 and A2) generally showed higher success rates, the higher proficiency levels (C1 and C2) experienced significant item attrition. Specifically, at the C1 level, only 7 items were valid against 5 invalid ones, and at C2, merely 5 items were valid compared to 8 invalid ones. This trend indicates a challenge in accurately measuring advanced Arabic reading proficiency. The high invalidity rate at the C1–C2 levels is primarily attributed to two factors: the complexity and subtlety of the texts and questions required by CEFR's advanced descriptors, and the potential lack of exposure among the subjects (fifth-grade students) to the highly specialized vocabulary and complex syntactic structures found in C1 and C2 texts.

These results indicate that although the distribution of valid items is relatively sufficient at the beginner (A1–A2) and intermediate levels (B1–B2), the proportion of valid items decreases at the advanced levels (C1–C2). This shows that the development of items at higher CEFR levels is more challenging and requires further revision and refinement to better align with the intended indicators. In general, the 72 valid items obtained are considered feasible to be used in the subsequent stages of analysis, such as difficulty level, discriminatory power, and reliability testing.

**Level of Difficulty**

the validity assessment, an analysis of the item difficulty index (p) was conducted to categorize the questions into easy, medium, and difficult.[33] The calculation was performed on all 119 questions that had been compiled in the initial stage. This analysis is important to see the general picture of the distribution of question difficulty before reduction is carried out based on the validity results.[34] The criteria used were: easy questions if the difficulty index > 0,70; medium question if

[33] Assad Ali Rezigalla et al., 'Item Analysis: The Impact of Distractor Efficiency on the Difficulty Index and Discrimination Power of Multiple-Choice Items', *BMC Medical Education* 24, no. 1 (2024): 445, https://doi.org/10.1186/s12909-024-05433-y.

[34] Helen Zhang et al., 'Developing and Validating the Artificial Intelligence Literacy Concept Inventory: An Instrument to Assess Artificial Intelligence Literacy among Middle School Students', *International Journal of Artificial Intelligence in Education* 35, no. 1 (2025): 398–438, https://doi.org/10.1007/s40593-024-00398-x.

in the range of 0,30-0,70; and difficult question if < 0,30.[35] The results of the difficulty level analysis are shown in Table 9.

**Table 9.** Distribution of Level of Difficulty of Question Items

| Category | p Range | Number of questions | Presentation |
|---|---|---|---|
| Easy | 0,71-1,00 | 75 | 63% |
| Medium | 0,31-0,70 | 33 | 28% |
| Hard | 0,00-0,30 | 11 | 9% |
| Total | | 119 | 100% |

The analysis of item difficulty, as presented in Table 9, indicates a skew towards the lower end of the spectrum, with the majority of items categorized as easy (75 items or 63%). Items classified as medium constituted 33 items (28%), while only 11 items (9%) were deemed hard. This distribution contrasts sharply with established international test construction principles, which generally recommend a bell-curve distribution where medium-difficulty items form the largest proportion (ideally around 60%), ensuring maximum differentiation among test-takers. The current dominance of easy items suggests that the initial test draft may lack sufficient power to discriminate between average and high-ability students. Consequently, these findings highlight a critical need for item bank refinement, specifically by increasing the cognitive load of future items to shift the distribution toward the medium and hard categories. to shift the distribution towards the medium and hard categories. Furthermore, based on the validity test results, subsequent analyses (discriminatory power and reliability) were necessarily limited to the 72 valid items that met the psychometric criteria.

**Discriminatory Power**

The next analysis is the item discrimination test. This test aims to determine the extent to which an item is able to differentiate students with high and low abilities.[36] The calculation is done by comparing the proportion of correct answers between the upper and lower groups. The criteria used are: discrimination power ≥ 0,40 = very good; 0,30-0,39 = good; 0,20-0,29 = sufficient; 0,00-0,19 = poor; and negative values = very poor (the item must be revised or removed). The results of the analysis of the discrimination power of all items are shown in Table 10.

**Table 10.** Distribution of Distinguishing Power of Question Items

| Category | D Range | Number of Questions | Presentation |
|---|---|---|---|
| Very Good | ≥ 0,40 | 6 | 5% |
| Good | 0,30-0,39 | 6 | 5% |
| Sufficient | 0,20-0,29 | 25 | 21% |
| Poor | 0,00-0,19 | 65 | 55% |
| Very Poor | <0,00 | 17 | 14% |
| **Total** | | **119** | **100%** |

[35] Ni Putu Eka Maryani Dewi and Ida Bagus Putrayasa, 'An Application of Difficulty Level Analysis of Question Items in Language Learning Evaluation', *Bulletin of Science Education* 4, no. 3 (2024), https://attractivejournal.com/index.php/bse/article/view/1681.

[36] Sayit Abdul Karim et al., 'Utilizing Test Items Analysis to Examine the Level of Difficulty and Discriminating Power in a Teacher-Made Test', *EduLite: Journal of English Education, Literature and Culture* 6, no. 2 (2021): 256, https://doi.org/10.30659/e.6.2.256-269.

Analysis of the discriminatory power presented in Table 10 reveals a significant psychometric challenge, with the majority of items falling into the poor category (65 items or 55%). Furthermore, 17 items (14%) were categorized as very poor, indicating that over half the items have a low capacity to effectively differentiate between high- and low-ability students. Only a small fraction of the items demonstrated acceptable power, with 25 items (21%) rated sufficient, 6 items (5%) good, and 6 items (5%) very good. This low discriminatory index has serious practical implications for classroom assessment. If used, these items would fail to accurately measure learning differences, potentially leading teachers to misclassify student competence (e.g., classifying a high-ability student as average). To enhance the instrument's quality, targeted improvement is essential. Concrete suggestions for improvement include: revising the distractors in the low-performing items to make them more plausible, simplifying the item stem to avoid ambiguity, or, if revision proves ineffective, discarding the item completely. Only after applying these targeted revisions and focusing exclusively on the 72 valid items with adequate discriminatory power can subsequent analysis, such as reliability testing, provide a meaningful measure of the instrument's overall quality.

**Test Reliability**

After analyzing the validity, difficulty level, and discriminating power, the final step was to test the reliability of the test instrument.[37] The purpose of the reliability test was to determine the internal consistency of the test items, namely the extent to which the items together can measure reading skills stably.[38] The reliability calculation in this study used the KR-20 formula (Kuder Richardson) or Cronbach's Alpha which is appropriate for instruments with multiple-choice questions.[39] Interpretation of reliability values is based on the following criteria: 0,80-1,00 = very high; 0,60-0,79 = high; 0,40-0,59 = sufficient; 0,20-0,39= low; <0,20 = very low. The results of the reliability calculation are shown in Table 11.

**Table 11.** Result of the Reliability Test of the Test Instrument

| Number of Valid Question Items | Reliability Coefficient | Category |
|---|---|---|
| 72 | 0,99 | Very High |

As demonstrated in Table 4, the final instrument exhibits an exceptional reliability coefficient of 0.99, which is categorized as very high. This robust coefficient initially suggests that the developed reading skills test instrument has excellent internal consistency, making it a reliable tool for measuring students' reading ability stably and accurately. However, such an extremely high figure must

[37] Annora Pratama Putri and Joko Sayono, 'Evaluation of Item Quality: Analysis of Difficulty Level and Distinction Power with Quantitative Methods', *Journal of Educational Sciences* 10, no. 1 (2026), https://jes.ejournal.unri.ac.id/index.php/JES/article/view/1246.

[38] Architha Aithal and P. S. Aithal, 'Development and Validation of Survey Questionnaire & Experimental Data – A Systematical Review-Based Statistical Approach', *SSRN Electronic Journal*, ahead of print, 2020, https://doi.org/10.2139/ssrn.3724105.

[39] Simon Ntumi et al., 'Estimating the Psychometric Properties (Item Difficulty, Discrimination and Reliability Indices) of Test Items Using Kuder-Richardson Approach (KR-20)', *Shanlax International Journal of Education* 11, no. 3 (2023): 18–28, https://doi.org/10.34293/education.v11i3.6081.

be interpreted critically. In field of psychometrics, a reliability coefficient approaching 1.00 often indicate redundancy or "overfitting." This implies that the items retained may be too homogeneous or highly correlated, essentially measuring the same narrow skill repeatedly, thereby artificially inflating the coefficient. While the 72 valid and reliable items are technically ready for broader trials and implementation, future studies using this instrument should prioritize exploring item diversity to ensure the test adequately captures the full breadth of reading skills across the CEFR levels, not just a highly uniform subset.

## Conclusion

This study has developed am Arabic reading skill test instrument based on CEFR framework through a systematic R&D approach using the 4D model. The initial blueprint consisted of 119 items covering levels A1-C2. After expert validation and empirical testing, 72 items are declared valid and met the requirements of good test construction. The analysis of item difficulty showed that most items were in the easy category, while the alaysis of discriminating power revealed that only a small proportion of items were categorized good, and verry good, indicating the need for revision in some test items. Despite these areas for revision, the set of validated items demonstrated the stability and reliability necessary for measuring Arabic reading competencies.

Therefore, the development test can be considered a valid and reliable instrument to evaluate Arabic reading skills of students, especially at the intermediate level. To build upon these findings, future research is recommended to implement the test on a larger and more diverse sample, as well to integrate it into digital platforms for broader application in language learning and assessment.

## Acknowledgment

This research was supported by a grant the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia (KEMENDIKBUD RI). We are deeply appreciative of their generous financial support through the research funding program for the year 2025, specifically detailed in: Decree Letter No. 0070/C3/AL.04/2025 dated May 23, 2025; and Grant/Contract Agreements No. 128/C3/DT.05.00/PL/2025 dated May 28, 2025; 062/LL7/DT.05.00/PL/2025 dated May 28, 2025; and 20/UNIDA/E.1-j/PT/XII/1446 dated May 29, 2025.

## Author Contribution Statement

IM conceptualized the study and supervised the research process. SAN conducted the literature review, designed the instrument, and carried out the data analysis. LFS assisted in the detailed psychometric analysis of the item difficulty and discrimination power, and contributed to the initial draft revision. MI contributed to the methodology, validation process, and interpretation of results. All authors contributed to the writing and approved the final version of the manuscript.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the workreported in this paper.

## AI Writing Statement

During the preparation of this manuscript, the authors used Chat GPT to assist idea organization. The authors carefully reviewed and edited the content generated with the assistence of AI and take full responsibility for the final content of this manuscript.

## References

Aiken, Lewis R. 'Three Coefficients for Analyzing the Reliability and Validity of Ratings'. *Educational and Psychological Measurement* 45, no. 1 (1985): 131–42. https://doi.org/10.1177/0013164485451012.

Aithal, Architha, and P. S. Aithal. 'Development and Validation of Survey Questionnaire & Experimental Data – A Systematical Review-Based Statistical Approach'. *SSRN Electronic Journal*, ahead of print, 2020. https://doi.org/10.2139/ssrn.3724105.

Alatlı, Reşat, İsa Güldenoğlu, and Tevhide Kargın. 'Examination of the Reading Comprehension Skills of Good and Poor Readers in the Dimension of Reading Components Developed by a Reading Skills Assessment Tool'. *Education and Science* 47, no. 211 (2022): 273–95. https://doi.org/10.15390/EB.2022.11080.

Albero-Ros, Elsa, Amalia Lorente-Velázquez, David Madrid-Costa, and Mariano González-Pérez. 'Development and Initial Validation of the MCL-PRO-CAT: A Computerized Adaptive Test Designed to Measure Multifocal Contact Lens Performance from the Patient's Perspective'. *Contact Lens and Anterior Eye* 48, no. 3 (2025): 102378. https://doi.org/10.1016/j.clae.2025.102378.

Alderson, J. Charles, Neus Figueras, Henk Kuijper, Guenter Nold, Sauli Takala, and Claire Tardieu. '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, no. 1 (2006): 3–30. https://doi.org/10.1207/s15434311laq0301_2.

Alharbi, Ahmad A., Hamad S. Al Amer, Abdulaziz A. Albalwi, et al. 'Cross-Cultural Adaptation and Psychometric Properties of the Arabic Version of the Fall Risk Questionnaire'. *International Journal of Environmental Research and Public Health* 20, no. 8 (2023): 5606. https://doi.org/10.3390/ijerph20085606.

Alwi, Nik Aloesnita binti Nik Mohd, and Wan Alisa Hanis binti Wan Abdul Halim. 'Variations and Methodological Components in CEFR-Aligned Language Tests: A Systematic Review'. *Journal of Creative Practices in Language Learning and*

*Teaching* 12, no. 1 (2024). https://journal.uitm.edu.my/ojs/index.php/CPLT/article/view/2708.

Cabahug, Kendy Suet. 'The Implementation of a Validated Contextualized Reading Material to Enhance Decoding Skills of Grade 1 Learners'. *Pantao (International Journal of the Humanities and Social Sciences)* 4, no. 2 (2025). https://doi.org/10.69651/PIJHSS040285.

Dewi, Ni Putu Eka Maryani, and Ida Bagus Putrayasa. 'An Application of Difficulty Level Analysis of Question Items in Language Learning Evaluation'. *Bulletin of Science Education* 4, no. 3 (2024). https://attractivejournal.com/index.php/bse/article/view/1681.

Duke, Nell K., and Kelly B. Cartwright. 'The Science of Reading Progresses: Communicating Advances Beyond the Simple View of Reading'. *Reading Research Quarterly* 56, no. S1 (2021). https://doi.org/10.1002/rrq.411.

Gazali, Erfan, and Hasan Saefuloh. 'Development of an Arabic Receptive Proficiency Test Instrument Based on the Common European Framework of Reference for Languages'. *Al-Ta'rib : Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab IAIN Palangka Raya* 11, no. 2 (2023): 293–308. https://doi.org/10.23971/altarib.v11i2.6721.

Haddad, Chadia, Hala Sacre, Sahar Obeid, Pascale Salameh, and Souheil Hallit. 'Validation of the Arabic Version of the "12-Item Short-Form Health Survey" (SF-12) in a Sample of Lebanese Adults'. *Archives of Public Health* 79, no. 1 (2021): 56. https://doi.org/10.1186/s13690-021-00579-3.

Hidayah, Fathi. 'Crosswalking as a Tool to Decide Arabic Language Standard in Madrasa Tsanawiyah: From Arabic Curriculum to ACTFL and CEFR'. *International Conference on Humanity Education and Society (ICHES)* 3, no. 1 (2024). https://proceedingsiches.com/index.php/ojs/article/view/265.

Ho, Ally Oi Kuan, Don Yao, and Antony John Kunnan. 'An Analysis of Macau's Joint Admission Examination–English'. *The Journal of AsiaTEFL* 18, no. 1 (2021): 208–22. https://doi.org/10.18823/asiatefl.2021.18.1.12.208.

Indaryanti, Rosita Budi, Harsono Harsono, Sutama Sutama, Budi Murtiyasa, and Bambang Soemardjoko. '4D Research and Development Model: Trends, Challenges, and Opportunities Review'. *Jurnal Kajian Ilmiah* 25, no. 1 (2025): 91–98. https://doi.org/10.31599/na7deq07.

Karim, Sayit Abdul, Suryo Sudiro, and Syarifah Sakinah. 'Utilizing Test Items Analysis to Examine the Level of Difficulty and Discriminating Power in a Teacher-Made Test'. *EduLite: Journal of English Education, Literature and Culture* 6, no. 2 (2021): 256. https://doi.org/10.30659/e.6.2.256-269.

Koray, Özlem, and Sercan Çetinkılıç. 'The Use of Critical Reading in Understanding Scientific Texts on Academic Performance and Problem-Solving Skills'. *Science Education International* 31, no. 4 (2020). https://www.icaseonline.net/journal/index.php/sei/article/view/239.

Li, Changlin, Nik Aloesnita Nik Mohd Alwi, and Mohammad Musab Azmat Ali. 'A Comparative Review of the CEFR and CET4 Writing Assessment with Insights from Task Complexity Theories'. *Malaysian Journal of Social Sciences and Humanities (MJSSH)* 10, no. 3 (2025): e003251. https://doi.org/10.47405/mjssh.v10i3.3251.

Li, Yan, Miaomiao Zhen, and Jia Liu. 'Validating a Reading Assessment Within the Cognitive Diagnostic Assessment Framework: Q-Matrix Construction and Model Comparisons for Different Primary Grades'. *Frontiers in Psychology* 12 (December 2021): 786612. https://doi.org/10.3389/fpsyg.2021.786612.

Mahmudah, Umi, and Tulus Musthofa. 'Reading Skills Learning in the "Arabic-Online.Net" Application by Saudi Electronic University Based on the Common European Framework of Reference for Languages (CEFR)'. *Scaffolding: Jurnal Pendidikan Islam Dan Multikulturalisme* 5, no. 3 (2023): 370–85. https://doi.org/10.37680/scaffolding.v5i3.3377.

Maqsood, Ammara, and Tanzeela Anbreen. 'Bridging Proficiency and Practice: Aligning Lexical Bundles in Pakistani L2 Learner Writing with CEFR Descriptors'. *Journal of Applied Linguistics and TESOL (JALT)* 8, no. 4 (2025). https://jalt.com.pk/index.php/jalt/article/view/1646.

Martin, Florence, Yan Chen, Robert L. Moore, and Carl D. Westine. 'Systematic Review of Adaptive Learning Research Designs, Context, Strategies, and Technologies from 2009 to 2018'. *Educational Technology Research and Development* 68, no. 4 (2020): 1903–29. https://doi.org/10.1007/s11423-020-09793-2.

Mokhtari, Hadjer. 'Asālīb Ta'līm Al-Lughah Al-'Arabiyah Li-Ghairi Al-Nāṭiqīn Bihā'. *HuRuf Journal : International Journal of Arabic Applied Linguistic* 2, no. 2 (2023): 156. https://doi.org/10.30983/huruf.v2i2.5956.

Muslim, Buhori, Zikrina Zikrina, and Mukhlisah Mukhlisah. 'Taṭwîr Kitāb Al-Qirā'at Al-Rasyîdah Li Tarqiyyah Mahārah Al-Qirā'ah 'Inda Al-Ṭālibah Bi Istikhdām Al-Kitāb Al-Elektrûny Al-Tafā'Uliy Fi Al-Madrasah Al-Mutawassiṭah Insān Qur'āny Aceh Besar'. *Jurnal Ilmiah Islam Futura* 23, no. 2 (2023): 347. https://doi.org/10.22373/jiif.v23i2.19489.

Nagai, Noriko, Gregory C. Birch, Jack V. Bower, and Maria Gabriela Schmidt. 'Integrating Learning, Teaching, and Assessment'. In *CEFR-Informed Learning, Teaching and Assessment*. Springer Texts in Education. Springer Singapore, 2020. https://doi.org/10.1007/978-981-15-5894-8_5.

Neal, David, Sophie Gaber, Phil Joddrell, Anna Brorsson, Karin Dijkstra, and Rose-Marie Dröes. 'Read and Accepted? Scoping the Cognitive Accessibility of Privacy Policies of Health Apps and Websites in Three European Countries'. *DIGITAL HEALTH* 9 (January 2023): 20552076231152162. https://doi.org/10.1177/20552076231152162.

Ntumi, Simon, Sheilla Agbenyo, and Tapela Bulala. 'Estimating the Psychometric Properties (Item Difficulty, Discrimination and Reliability Indices) of Test Items Using Kuder-Richardson Approach (KR-20)'. *Shanlax International Journal of Education* 11, no. 3 (2023): 18–28. https://doi.org/10.34293/education.v11i3.6081.

Połczyńska, Monika M., and Susan Y. Bookheimer. 'General Principles Governing the Amount of Neuroanatomical Overlap between Languages in Bilinguals'. *Neuroscience & Biobehavioral Reviews* 130 (November 2021): 1–14. https://doi.org/10.1016/j.neubiorev.2021.08.005.

Putri, Annora Pratama, and Joko Sayono. 'Evaluation of Item Quality: Analysis of Difficulty Level and Distinction Power with Quantitative Methods'. *Journal of Educational Sciences* 10, no. 1 (2026). https://jes.ejournal.unri.ac.id/index.php/JES/article/view/1246.

Razida, Mirdawati, Nur Hasaniyah, and Abdul Muntaqim Al Anshory. 'Blending Technology and Pedagogy: Optimizing Maharah al-Qirā'ah through the Alef Education Platform'. *Al-Irfan : Journal of Arabic Literature and Islamic Studies* 8, no. 2 (2025): 211–25. https://doi.org/10.58223/al-irfan.v8i2.424.

Rezigalla, Assad Ali, Ali Mohammed Elhassan Seid Ahmed Eleragi, Amar Babikir Elhussein, et al. 'Item Analysis: The Impact of Distractor Efficiency on the Difficulty Index and Discrimination Power of Multiple-Choice Items'. *BMC Medical Education* 24, no. 1 (2024): 445. https://doi.org/10.1186/s12909-024-05433-y.

Rohman, Habibur, and Faiq Ilham Rosyadi. 'Development of Arabic Teaching Materials Based on the Common European Framework of Reference (CEFR) to Improve Students' Arabic Language Skills'. *Al Mahāra: Jurnal Pendidikan Bahasa Arab* 7, no. 2 (2021): 163–83. https://doi.org/10.14421/almahara.2021.072-01.

Santuya, Glenna Rose. 'Learners' Level of Reading Comprehension: Basis for Contextualized Reading Materials'. *Pantao (International Journal of the Humanities and Social Sciences)* 4, no. 2 (2025). https://doi.org/10.69651/PIJHSS040292.

Shi, Lijun, Tuan Sarifah Aini Syed Ahmad, and Anealka Aziz Hussin. 'A Systematic Literature Review of Current Studies on Comparison Between the CEFR and CSE'. *International Journal of Social Science Research* 12, no. 2 (2024): 18. https://doi.org/10.5296/ijssr.v12i2.21627.

Supunya, Nuntapat. 'Towards the CEFR Action-Oriented Approach: Factors Influencing Its Achievement in Thai EFL Classrooms'. *3L The Southeast Asian Journal of English Language Studies* 28, no. 2 (2022): 33–48. https://doi.org/10.17576/3L-2022-2802-03.

Van Den Akker, Jan, Koeno Gravemeijer, Susan McKenney, and Nienke Nieveen, eds. *Educational Design Research*. Routledge, 2006. https://doi.org/10.4324/9780203088364.

Wan Abdul Halim, Wan Alisa Hanis, and Nik Aloesnita Nik Mohd Alwi. 'Cefr-Aligned Language Tests: A Systematic Scoping Review'. SSRN Scholarly Paper No. 4244716. Social Science Research Network, 11 October 2022. https://papers.ssrn.com/abstract=4244716.

Zakkiyah, Maulia Yasminah, Nurriya Maghfirah Fidyahwati, Ahmad Tarajjil Ma'suq, and Novita Anggraini. 'Assessment Design and Analysis of Arabic Reading Skills Instructional Materials'. *IJIE International Journal of Islamic Education* 3, no. 1 (2024): 31–46. https://doi.org/10.35719/ijie.v3i1.2000.

Zhang, Helen, Anthony Perry, and Irene Lee. 'Developing and Validating the Artificial Intelligence Literacy Concept Inventory: An Instrument to Assess Artificial Intelligence Literacy among Middle School Students'. *International Journal of Artificial Intelligence in Education* 35, no. 1 (2025): 398–438. https://doi.org/10.1007/s40593-024-00398-x.

## Biography of Authors

**Safira Aina Najiyah** is a postgraduate student in the Arabic Language Education program at the University of Darussalam Gontor, Indonesia. She earned her Bachelor's degree in Arabic Language Education from the same university and is currently pursuing her Master's degree in Arabic Language Education at UNIDA Gontor. Her academic interests include Arabic language teaching, language testing, and error analysis. She is also involved in research focusing on the development of evaluation instruments and language pedagogy in Islamic educational contexts. She can be contacted via email at safiraaina30@gmail.com.

**Ihwan Mahmudi** is a lecturer at the University of Darussalam Gontor, Indonesia, and currently serves as the Dean of the Faculty of Tarbiyah. He earned his Bachelor's degree from ISID (now UNIDA Gontor), continued his Master's degree in Educational Research and Evaluation at Universitas Negeri Jakarta, and completed his Doctoral degree in the same field at Universitas Negeri Jakarta. His academic interests include educational evaluation, test development, and assessment in learning. He has supervised numerous research projects and actively contributes to academic innovation and institutional quality enhancement at UNIDA Gontor. He can be contacted at ihwanm@unida.gontor.ac.id.

**Muhammad Ismail** is a lecturer in Arabic Language for Non-Native Speakers at the University of Darussalam Gontor, Indonesia. He earned a Bachelor's degree in Arabic Language Education (2007–2011) and a Master's degree in the same field (2011–2014) from UNIDA Gontor. He completed his Doctoral degree at the University of the Holy Quran and Islamic Sciences, Sudan (2018–2022). His expertise spans Arabic testing, translation, research methodology, and teaching strategy. Since 2013, he has been teaching Arabic and is also a certified book editor. He is the developer of the Alikhtibar Learning Management System and has led projects such as the Arabic Online Course and Arabic Adaptive Test. He can be contacted via email at ismail@unida.gontor.ac.id.

**Latif Fatus Sa'diyah** is a student at Al-Azhar University, Cairo, Egypt. She actively participates in various student and social organizations, both in academic and non-academic fields. Her academic interests include Arabic studies, Islamic education, and cross-cultural communication. She can be contacted via email at: latiffatussadiyah20@gmail.com.