Al-Ta'rib
Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab
IAIN Palangka Raya
Vol. 11, No. 2, December 2023, 293-308
p-ISSN 2354-5887 | e-ISSN 2655-5867
https://doi.org/10.23971/altarib.v11i2.6721

Development of an Arabic Receptive Proficiency Test Instrument Based on the Common European Framework of Reference for Languages

Erfan Gazali[1*], Hasan Saefuloh[2]
[1,2]Institut Agama Islam Negeri Syekh Nurjati Cirebon, Indonesia
**E-mail: erfangazali@iain-syekhnurjati.ac.id*

Abstract

This research develops a Common European Framework of Reference for Languages (CEFR)-based Arabic receptive proficiency test instrument at the Institut Agama Islam Negeri (IAIN) Sheikh Nurjati Cirebon. This study addresses the lack of international standard measures to test Arabic receptive competency, including listening and reading. This work created a CEFR-compliant test instrument that matches Arabic language specifics. The Research and Development (RnD) strategy using the ADDIE model—analysis, design, development, implementation, and evaluation—is applied. The steps include identifying needs, instrument design, feedback-based iterative development, implementation in an authentic setting, and detailed efficacy evaluation. This study found that the instrument is valid, reliable, and effective in testing CEFR-level Arabic receptive proficiency. The instrument considers Arabic's intricacies and complexities to assess receptive performance holistically and contextually. This work enriches Arabic language competency measuring literature and provides a valuable measurement tool for educators and learners, particularly in Islamic education.

Keywords*: Arabic proficiency,* Framework of Reference for Languages (*CEFR), receptive skill, ADDIE*

Abstrak

Fokus penelitian ini adalah untuk mengembangkan instrumen uji kemampuan pemahaman bahasa Arab yang efektif dan terstandar, menurut *Common European Framework of Reference for Languages* (CEFR), di Institut Agama Islam Negeri (IAIN) Syekh Nurjati Cirebon. Pernyataan masalah yang mendasari penelitian ini adalah kekurangan instrumen yang dapat secara akurat mengukur kemahiran reseptif bahasa Arab, yakni keterampilan mendengar dan membaca, sesuai dengan standar internasional. Penelitian ini bertujuan menghasilkan sebuah instrumen tes yang sesuai dengan standar CEFR dan cocok untuk bahasa Arab. Metode penelitian yang digunakan adalah pendekatan *Research and Development* (RnD) dengan model ADDIE, yang meliputi tahapan Analisis, Desain, Pengembangan,

Implementasi, dan Evaluasi. Proses ini melibatkan identifikasi kebutuhan, perancangan awal instrumen, pengembangan iteratif berdasarkan umpan balik, implementasi dalam konteks nyata, dan evaluasi komprehensif terhadap efektivitasnya. Hasil utama dari penelitian ini menunjukkan bahwa instrumen yang telah dikembangkan berhasil memenuhi standar validitas dan reliabilitas, serta terbukti efektif dalam mengukur kemampuan pemahaman bahasa Arab sesuai dengan tingkat CEFR. Instrumen ini menawarkan penilaian yang lebih menyeluruh dan terkait dengan konteks dalam kemampuan pemahaman, memperhatikan detail dan kerumitan dalam penggunaan bahasa Arab. Penelitian ini memberikan kontribusi yang penting dalam memperkaya literatur mengenai pengukuran kemahiran bahasa Arab dan menyediakan alat ukur yang berguna bagi para pendidik dan pembelajar bahasa Arab, terutama di lingkungan pendidikan Islam.
Kata kunci: Kecakapan berbahasa, CEFR, keterampilan reseptif, ADDIE

INTRODUCTION

Critical to global language education is the capacity to quantify and comprehend language proficiency, particularly in the case of Arabic, a language significant in cultural and global contexts. Arabic is an enduring cultural treasure in the heritage of global civilization (Ernst, 2013) and is spoken by over 1.5 billion Muslims worldwide during religious rituals (liturgy) (Bokova, 2012). Arabic language study has grown substantially in popularity in Indonesia, specifically at the Sheikh Nurjati Language Development Centre Cirebon. Nonetheless, the paucity of effective proficiency testing instruments, particularly for receptive skills (listening and reading), still needs to be addressed for Arabic language instruction and evaluation.

Over the past five years (2018-2023), observations at the Sheikh Nurjati Cirebon Language Development Centre in Indonesia have revealed a significant limitation in administering the Test of Arabic as a Foreign Language (TOAFL). This limitation lies in excluding writing and speaking assessments in evaluating Arabic language proficiency. Currently, the focus is solely on listening, reading, and grammar.

The root of this issue, as indicated by Mr. Khasan Aedi, the leader of the language development institution, is the lack of standardized tools at the Language Development Centre for evaluating Arabic language proficiency. This deficiency predominantly affects the assessment of productive abilities – speaking and writing. One of the significant challenges in this context is the limited availability of native Arabic speakers who can evaluate speaking skills. Additionally, standardized criteria are absent for assessing overall language proficiency, leading to an incomplete evaluation of a student's command of the Arabic language.

This article aims to rectify these deficiencies by creating an Arabic receptive proficiency assessment tool that aligns with the Common European Framework of Reference for Languages (CEFR). The CEFR is a highly acknowledged framework used to describe levels of language acquisition and communicative ability in Europe and globally (Aleksandrova & Pouliot, 2023; Galantomos, 2021; Musthofa, 2022; Nurdianto et al., 2022; Prajapati, 2022; Subekti et al., 2023; Vani et al., 2022).

By incorporating the CEFR, the proposed instrument will establish standardized and thorough evaluation criteria and facilitate the comparison of Arabic language ability with other languages that adhere to the same framework.

The CEFR-based Arabic receptive proficiency test instrument is essential for several reasons. The tool improves the Arabic language evaluation system in Indonesia, as shown at IAIN Sheikh Nurjati Cirebon's Language Development Center and standardizes language competence assessment worldwide. Adopting the CEFR allows this instrument to compare Arabic language ability with other languages using the same framework, improve teaching and evaluation, and expand multilingual education. It helps teachers and educational institutions establish more effective and focused teaching methods, contributes to research and curriculum development, and enriches educational materials and intercultural communication bridges. Thus, this instrument is crucial to Indonesian language education and promoting Arabic as a worldwide language with standardized and quantitative teaching and assessment methods.

In 1978, Roushdy Ahmad Toiemah (1978) developed a standardized test to measure language proficiency in Arabic for foreign speakers among students studying Arabic in several American universities. It was one of the first studies to measure language proficiency in Arabic for foreign speakers. Consequently, research about the standardization of language proficiency examinations emerged and became the primary focus of Arabic linguists' research (Ben Khiroun et al., 2014; de Graaf, 2021; Masrai & Milton, 2019; Rifaie et al., 2021; Winke & Aquil, 2014).

Assessment of a language's proficiency has been around as long as teaching itself (Farhady, 2018). Black and Williams (2010) and Kennedy et al. (2008) found that standardized assessment and evaluation boosted student knowledge retention and comprehension. Highlighting areas of strength and improvement can aid students in their education. David Boud argues that doing assessments helps students learn and enhance their abilities (Boud, 1990).

Language testing procedures are fundamentally distinct from those employed in most other disciplines. It is due to the fact that teachers of foreign languages have a broad range of assessment tools from which to choose for their students (Brown & Hudson, 1998). Experts refer to the ability to comprehend, speak, read, and write proficiently in a given language as "language proficiency"(Bachman, 2000; Richards & Schmidt, 2010). These four abilities have been identified by educators in the field of linguistics as essential for language learners. Receptive skills include reading and listening, whereas productive skills include speaking and writing (Laufer & Goldstein, 2004; Masduqi, 2016; Sreena & Ilankumaran, 2018).

Awamleh (2004) proposes testing grammar, literary style, and cultural sensitivity in addition to the four basic competencies. Language, culture, and education should inform knowledge assessments. Brown & Hudson (1998) identify three main types of language proficiency tests: (a) selected-response tests, which include true-false, matching, and multiple-choice tests; (b) constructed-response tests, which require language learners to write, speak, or do something. (c) Personal response tests, fill-in-the-blank, short-answer, and performance evaluations. These exams examine students' abilities during the learning process,

depending on their engagement: self-evaluation and peer evaluation, portfolios, and debates.

METHOD

Research Design

The research and development method is used in this study. The ADDIE study model is used for research and development (R&D). The ADDIE model is a way to plan how to build something. It has five stages: analysis, design, development, implementation, and evaluation (Branch, 2009). Development study aims to make things and test how well they work. Objectivity, reliability, and validity are the three primary indicators, also known as core quality criteria, that have become the standard for evaluating the quality of a test instrument (Ebel & Frisbie, 1991; Linn, 2011; Miller et al., 2012). Objectivity is required for accurate measurement, and accurate measurement is required for instrument validity (Wess et al., 2021).

Data Collection Technique

In the development research, the product trial design comprises 1) test design, 2) test subjects, 3) data collection techniques and instruments, and 4) data analysis techniques. The test design developed through research on developing Arabic language proficiency test instruments involves expert validation and small-scale trials. The subjects of the expert test consisted of experts or experts, namely Arabic grammarians, Arabic learning evaluation experts, and lecturers teaching Arabic language skills.

Data Analysis Technique

The data analysis technique on the instrument uses test item validity with four types of tests. First, the content validity ratio (CVR) and content validity index (CVI) with the formula proposed by Lawshe (1975) are categorized into three rating scales: (1) essential, (2) useful but not essential, and (3) unnecessary. Second, the reliability test concerning the Kuder and Richardson formula number 20 is done. Third, the test measures the items' difficulty level using the proportion formula (difficulty index), and the fourth is the item discrimination test with the discrimination index.

RESULT AND DISCUSSION

This study led to constructing questions for an Arabic language competency test. With 100 items, the tests are multiple-choice questions with three to four possible answers (a, b, c, and d). The steps of growth in this study are about how the ADDIE model was made and how it works. Five steps in the ADDIE model include analysis, design, development, implementation, and evaluation.

Analysis

The analysis has two stages: needs assessment and front-end analysis. Needs assessment in the form of analysis of the state of the field and participants, as well as the collection of reference tests that will be used as the subject matter in developing test instruments. Field analysis activities were conducted by collecting information about the need for Arabic language proficiency assessment at IAIN

Sheikh Nurjati Cirebon. The results of the information related to the needs of the Arabic language proficiency test are as follows.

Language Development Center (LDC) IAIN Sheikh Nurjati Cirebon does not yet have a comprehensive instrument for measuring Arabic language skills that can be used to assess the four competencies (listening, reading, writing, and speaking skills) of Arabic language students in the IAIN Sheikh Nurjati Cirebon environment. In the meantime, the LDC only possesses a test instrument for Arabic listening and reading skills but not for writing and speaking skills. At the level of implementation, the test instrument is manual, consisting of a pencil-based answer page and audio media to assist in listening to questions about aspects of listening skills. The Department of Arabic Language and Literature and the Department of Arabic Language Education, on the other hand, still need test instruments to measure the Arabic language abilities of their students.

The LDC's instrument is still generic and does not distinguish between levels of language proficiency, such as novice, intermediate, and advanced. The total score comprises three assessment components: structure and grammar skills, auditory skills, and text comprehension skills. The scoring figures are based on the TOEFL evaluation. The instrument modifies the questions accumulated since 2005 by adjusting a subset of questions as a test variation and not in response to user requirements.

There is a need for instruments founded on generally accepted standards for measuring language proficiency levels. Examples include the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines, the Canadian Language Benchmarks (CLB), the Interagency Language Roundtable (ILR) assessments, and the Common European Framework of Reference for Languages, or CEFR for short. This study's measurement standard is the Common European Framework of Reference for Languages (CEFR), a language standard established by the European Council in 2001 and serves as a guide for describing students' foreign language skills in Europe, particularly in academic settings. There are six levels of language standards: A1, A2, B1, B2, C1, and C2 (Council of Europe, 2001).

The CEFR proficiency levels, which range from basic to professional, require many test instruments to fulfil each level. Therefore, in this study, the development of test instruments focuses on level B1, or the intermediate level, with aspects of measurement on Arabic receptive skills, namely listening skills and understanding texts, which include mastery of grammar and vocabulary.

Planning and design of test instruments

The stage of designing receptive Arabic language proficiency test instruments includes setting test objectives, creating test grids, and creating questions.

Setting Test Objectives

The test objectives are designed to evaluate auditory and reading comprehension of Arabic. The test contains questions based on indicators derived from the Common European Framework of Reference for Languages (CEFR) at the B1 (intermediate) or independent user proficiency level. The equivalence levels between level B1 and other foreign language proficiency standards are as follows:

Table 1
Degree of equivalence of CEFR Level B1 with similar English language measurement standards (Efset.org, 2021)

Test Type	Score equivalent to B1 Level
IELTS	4.0 – 5.0
TOEIC	550-780
TOEFL iBT	42-71

Making question grids

After determining the test's purpose, the researcher compiled a grid of test questions. Arabic receptive skills at level B1 contain the following linguistic competencies:

Table 2
Receptive Proficiency at Level B1-CEFR

level	Aspect	Description of competence
B1	Listening	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, and leisure. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken.
	Reading	Able to understand familiar text or work-related vocabulary. Can understand descriptions of feelings or desires.

The measured aspects of linguistic competence are translated into four indicators, then reduced to ten topics that will be developed into one hundred questions (see Table 7).

Creating a question

After identifying the developed indicators and topics for assessing listening and writing abilities, 100 multiple-choice questions with four possible answers are compiled. The questions are separated into three categories. Listening (30 questions), comprehensive reading (40 questions), and grammar and vocabulary (30 questions) are the specific sections of the exam.

Development

At this stage, errors or deficiencies are rectified by validating the queries that experts have developed. The aim is to measure whether or not the question items achieve the objectives set. The validators comprised eight experts, including Arabic linguists, language learning evaluation professionals, and instructors.

Testing and revising instruments

In order to ascertain the level of validity and reliability of the instrument, eight panelists evaluate the question items that have been developed based on these indicators. Improvements will be made based on the panelists' comments and suggestions. Then, the questions will be tested with a small group online to determine the items' degree of difficulty and their ability to differentiate.

The 100 question items were submitted for validation to eight panelists or experts, who evaluated the content's viability. Using the CVI formulation (Lawshe, 1975), the content validity index of the receptive Arabic language proficiency test was calculated after determining the CVR index for each item of the assessment instrument. As shown in Table 6, the result of the CVI calculation is the mean CVR for all query items. Based on the evaluations of eight experts, the item-CVI index of the test instrument is calculated to be 96.8, and the average scale value of the CVI score (S-CVI) of all items is 0.96 (see Table 4).

Table 3
Content validity index scale and index scale results

Sum of I-CVI	96.88	Sum of UA	86
S-CVI/Ave	**0.97**	S-CVI/UA	**0.86**
Category	**accepted**		**accepted**

The CVR value derived from the calculation was compared to the CVR critical value determined by the number of validators enumerated in Table 5. The item is accepted if its value is equal to or higher than the CVR critical value, and it is rejected if its value is less than the CVR critical value (Ayre & Scally, 2014; Wilson et al., 2012).

Table 5
Simplified Table of CVR critical, Including the Number of Experts Required to Agree on an Item

Panel Size	$N_{critical}$ (Minimum Number of Experts Required to Agree an Item Essential for Inclusion)	Proportion Agreeing Essential	$CVR_{critical}$
5	5	1	1.00
6	6	1	1.00
7	7	1	1.00
8	7	.875	.750
9	8	.889	.778
10	9	.900	.800
11	9	.818	.636
12	10	.833	.667
13	10	.769	.538
14	11	.786	.571
15	12	.800	.600

This result indicates that the devised test instrument has a very high content validity index (CVI) based on the seven-statement assessment instrument used by experts.

The next stage is to revise the questions that need improvement based on the notes and feedback from the experts. Then, it goes to the stage of assembling questions for limited trials in a learning management system based on MOODLE (Modular Object-Oriented Dynamic Learning Environment) version 4.1 by dividing the questions into three groups of questions: grammar, listening, and reading comprehension. The following is a screenshot of the uploaded Arabic exam on the Moodle platform:

Figure 1
The dashboard for the Moodle-based online Arabic Proficiency Test

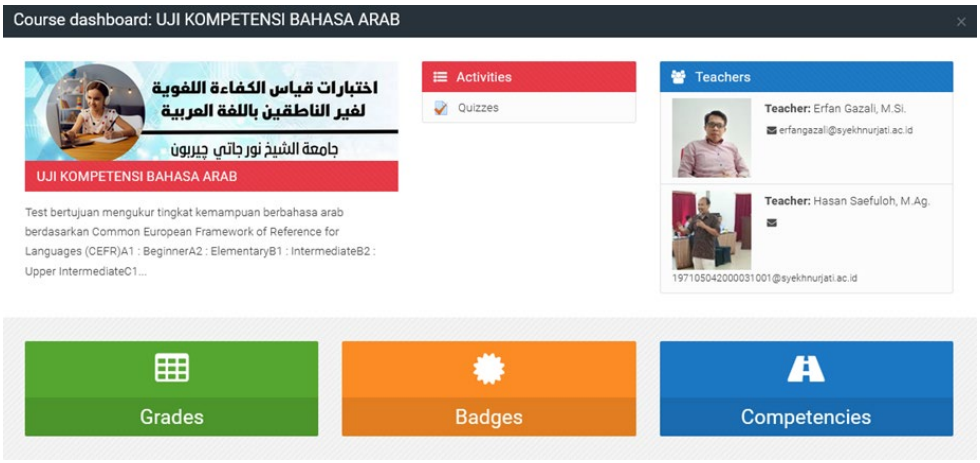

Figure 2
The reading comprehension session question (*fahmu al-maqru*)

Figure 3
Arabic vocabulary and reading comprehension questions

Table 4
Competencies and performance indicators for listening and reading (adapted and processed from the Council of Europe, 2001)

No.	competence	Indicator	Question theme	Number of questions
1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, and leisure. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken.	1. Can understand the main points of clear, standardized input on matters commonly encountered at work, at school, at recreation	1. Personal and professional hopes and dreams for the future	10
2	Able to understand familiar text or work-related vocabulary. Can understand descriptions of feelings or desires.	2. Can handle most situations that are most likely to occur while traveling in the area where the language is spoken.	2. Watching television and favorite shows.	10
		3. Capable of producing straightforward, connected texts on familiar or personal topics.	3. Education and Future Plans	
		4. Possesses the ability to describe experiences, events, dreams, hopes, and aspirations and concisely explain opinions and plans.	4. Favorite music, music or entertainment trends, and plans to attend Arabic drama performances.	10
			5. Healthy practices, diet, lifestyle, and giving and receiving advice in these areas.	10
			6. Meet people on social networks.	10
			7. Going to a restaurant, ordering food, having supper conversation, and paying for dinner.	10
			8. Participate in negotiations	10
			9. Safety concerns, accident reporting, and explanation of regulations.	10
			10. Polite behavior and respond appropriately to impolite behavior.	10

Implementation

Small-Group Testing

The test prototype integrated into the Moodle software was then tested online with 15 students from the Department of Arabic Language and Literature, IAIN Sheikh Nurjati Cirebon. The data from this trial are used to calculate the level of question item reliability, question difficulty, and differentiation and assess the product's usability. Online assessments allow students to conduct product evaluations. After the students completed the online test, the researcher distributed a questionnaire to the respondents to evaluate the test instrument measurement of the receptive Arabic language proficiency based on the application's usability, sound and image quality, and question presentation.

The data analysis technique used to estimate the instrument's reliability used the internal consistency estimation technique with Kuder and Richardson formula number 20, and the reliability score for multiple-choice questions obtained was 0.95, which indicates the level of reliability is in the very high range.

Table 5
Reliability of multiple-choice items on Arabic receptive proficiency

n/n-1	(St2-∑PQ)St2	r
1.01	0.94	0.95

According to Table 6, the difficulty level of the question items does not have a balanced proportion of questions, as the number of medium category questions has a higher proportion of 51 questions (51%). The lower difficult category amounted to 6 questions (6%). For the easy category, as many as 42 questions (42%), while very easy was 1 question (1%), and questions with a very difficult category did not exist.

The distinguishing power of an item depends on the size of the discrimination index value. According to Table 7, the average distractor of 100 questions is classified into five groups. There are three questions (3% in the excellent category), 67 (67% in the good category), 11 (11% in the sufficient category), 12 (11% in the terrible category), and 7 (7% in the very bad category).

The average distraction of 100 queries falls into five categories. Specifically, the class of excellent queries There are three questions (3% of the total) with an average difficulty index of 1, the good category has 67 questions (67% of the total) with an average difficulty index of 0.78, the fair category has 11 questions (11%) with an average difficulty index of 0.60, the bad category has 12 questions (11%) with an average difficulty index of 0. The very bad category has seven questions (7% of the total) with an average difficulty index of -0.22.

Table 6
Item difficulty of multiple-choice questions on Arabic receptive proficiency

N-of Item	Item Number	Average $(\bar{x})$ difficulty index	Category	Percentage (%)
1	37	1	Very easy	1
42	2,4,5,6,7,15,19,20,21,23,25,29,31,32,33,36,41,44,50,51,52,54,58,59,61,63,65,67,69,72,76,78,80,81,82,83,87,90,91,92,95,96	0,78	Easy	42
51	1,3,10,12,13,14,16,17,18,21,24, 26,27,28,30,34,35,38,39,40,42, 43,45,46,47,48,49,53,55,57,60, 62,64,66,70.71,73,74,75,77,798 5,86,88,89,93,94,97,98,99,100	0,60	moderate	51
6	8,9,11,56,68,84	0.28	Difficult	6
0	-	0	Very difficult	0

According to the evaluation results of students' responses to the test instrument product, 74.4% (11) of respondents rated the level of ease in operating the test application as "Good." In comparison, 86.7% (13) of respondents rated the aspect of sound and images presented as "Very Good." The aspect of appearance or layout in the presentation of questions and answers received the category "Very Good" with 12 responses, or 80% of the total.

Evaluation

Product evaluation follows validation and testing. As a small-group trial revision, the product is reviewed. Field evaluations must be addressed if shortcomings are detected. Product evaluation results are approved, changed, or rejected. Accepted question items have an essential or valid correlation, high dependability, moderate difficulty, and are good or very good. If the correlation is valid, reliability is low, difficulty is low, and discriminatory power is low, The item can be approved with corrections. If the correlation is invalid or non-essential, reliability is high, difficulty level is low, and discriminating power is poor or extremely poor, the question must be eliminated due to its poor quality. From the summary of the questions above, it can be concluded that 97 questions are used without revision, three questions (numbers 15, 20, and 23) must be revised, and questions that must be discarded due to poor quality cannot be found.

CONCLUSION

The result of this development research is an instrument for assessing Arabic language proficiency based on the Common European Framework of Reference for Languages (CEFR). The proficiency test instrument must satisfy a limited trial before becoming the final product. However, before testing, the product was validated by eight experts in Arabic grammar, Arabic language skills, and evaluation of Arabic language learning. Based on the validation and limited trial results, expert input was used to refine the product. In terms of validity,

reliability, level of difficulty, and differentiating power, the final product meets the criteria for quality items.

The development of this Arabic language proficiency test instrument has limitations: (1) the development of multiple-choice test instruments is limited to receptive competencies of Arabic language skills, which are listening and reading skills, as well as grammar and vocabulary. (2) Product development trials have yet to reach the stage of large-scale field trials and product testing, and (3) the number of test items has not been altered to interpret the meaning of each indicator theme.

REFERENCES

Aleksandrova, D., & Pouliot, V. (2023). CEFR-based Contextual Lexical Complexity Classifier in English and French. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 518–527. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174487896&partnerID=40&md5=b61588e074905393ecdca6bb78315449

Awamleh, H. (2004). *Mahārāt Ta'līm al-Qirā'aẗ wa al-Kitābaẗ Lilāṭfāl*. Dār Wā'il lilṭbā'aẗ Wa al-Nšar Wa-alTaūzī'.

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Bachman, L. F. (2000). Modern language testing at the turn of the Century: assuring that what we count counts. *Language Testing*, *17*(1), 1–42. https://doi.org/10.1191/026553200675041464

Ben Khiroun, O., Ayed, R., Elayeb, B., Bounhas, I., Ben Saoud, N. B., & Evrard, F. (2014). Towards a new standard Arabic test collection for mono- and cross-language information retrieval. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8455 LNCS*, 168–171. https://doi.org/10.1007/978-3-319-07983-7_23

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Bokova, I. (2012). *History of the Arabic language at UNESCO*. UNESCO. http://www.unesco.org/new/en/unesco/resources/history-of-the-arabic-language-at-unesco/

Boud, D. (1990). Studies in Higher Education Academic Values Assessment and the Promotion of Academic Values. *Studies in Higher Education*, *15*(1), 101–111. https://doi.org/http://dx.doi.org/10.1080/03075079012331377621

Branch, R. M. (2009). *Instructional Design: The ADDIE Approach*. Springer.

Brown, J. D., & Hudson, T. (1998). Alternatives in Language Assessment. *TESOL*

Quarterly, *32*(4), 653–675.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. Language Policy Programme Education Policy Division Education Department Council of Europe. www.coe.int/lang-cefr

De Graaf, A. (2021). Challenges in Developing Standardized Tests for Arabic Reading Comprehension for Secondary Education in the Netherlands. In *Fairness in College Entrance Exams in Japan and the Planned Use of External Tests in English* (hal. 181–189). Springer Singapore. https://doi.org/10.1007/978-981-33-4232-3_13

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5 ed.). Prentice Hall.

Ernst, C. W. (2013). The Global Significance of Arabic Language and Literature. *Religion Compass*, *7*(6), 191–200. https://doi.org/10.1111/rec3.12049

Farhady, H. (2018). History of Language Testing and Assessment. *The TESOL Encyclopedia of English Language Teaching*, pp. 1–7. https://doi.org/10.1002/9781118784235.eelt0343

Galantomos, I. (2021). Developing "conceptual knowledge" descriptors for CEFR-based proficiency levels. *Journal of Second Language Studies*, *4*(1), 96–120. https://doi.org/10.1075/jsls.19029.gal

Kennedy, K. J., Chan, J. K. S., Fok, P. K., & Yu, W. M. (2008). Forms of assessment and their potential for enhancing learning: Conceptual and cultural issues. *Educational Research for Policy and Practice*, *7*(3), 197–207. https://doi.org/10.1007/s10671-008-9052-3

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Linn, R. L. (2011). The Standards for Educational and Psychological Testing: Guidance in Test Development. In T. M. Haladyna & S. M. Downing (Ed.), *Guidance in Test Development* (hal. 27–38). L. Erlbaum.

Masduqi, H. (2016). Integrating Receptive Skills and Productive Skills into a Reading Lesson. *The 2nd International Conference on Teacher Training and Education*, *2*(1), 507–511. https://jurnal.uns.ac.id/ictte/article/view/7476

Masrai, A., & Milton, J. (2019). How many words do you need to speak Arabic? An Arabic vocabulary size test. *Language Learning Journal*, *47*(5), 519–536. https://doi.org/10.1080/09571736.2016.1258720

Miller, D., Linn, R., & Gronlund, N. (2012). *Measurement and Assessment in Teaching* (11 ed.). Pearson Education Inc.

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. https://doi.org/10.32601/ejal.911545

Nurdianto, T., P, N. J., Fatoni, A., & Kalita, S. (2022). CEFR-Based Beginner Arabic Reading And Writing Curriculum Design In Indonesia. *Journal of Arabic Learning*, *5*(3), 718–738.

Prajapati, M. (2022). Introductory Guide to the Common European Framework of Reference (CEFR) for English Language Teachers. *Integrated Journal for Research in Arts and Humanities*, *2*(6), 291–296. https://doi.org/https://doi.org/10.55544/ijrah.2.6.40

Richards, J. S., & Schmidt, R. (2010). *Longman Dictionary of Language Teaching and Applied Linguistics* (4 ed.). Longman (Pearson Education).

Rifaie, N., Hamza, T. M. A. W., & Elfiky, Y. H. (2021). Standardization of the Revised Arabic Language test for 4-8-year old children. *Egyptian Journal of Ear, Nose, Throat and Allied Sciences*, *22*(22). https://doi.org/10.21608/EJENTAS.2021.92770.1414

Sreena, S., & Ilankumaran, M. (2018). Developing Productive Skills Through Receptive Skills – A Cognitive Approach. *International Journal of Engineering & Technology*, *7*(4.36), 669. https://doi.org/10.14419/ijet.v7i4.36.24220

Subekti, A. S., Widodo, P., & Andriyanti, E. (2023). Indonesian L2 Learners' CEFR-based Listening Proficiency: Interactions with Attitudes towards Teachers' Use of L1. *Acta Paedagogica Vilnensia*, *50*, 37–51. https://doi.org/10.15388/ACTPAED.2023.50.3

Toiemah, R. A. (1978). *The Use of Cloze to Measure the Proficiency of Students of Arabic: As a Second Language in Some Universities in the United States*. University of Minnesota.

Vani, R., Mohan, S., & Ramkumar, E. V. (2022). A Study on Ameliorating Indian Engineering Students' Communication Skills in Relation With CEFR. *Theory and Practice in Language Studies*, *12*(6), 1172–1180. https://doi.org/https://doi.org/10.17507/tpls.1206.17

Wess, R., Klock, H., Siller, H.-S., & Greefrath, G. (2021). *Measuring Professional Competence for the Teaching of Mathematical Modelling*. Springer Nature.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the Critical Values for Lawshe's Content Validity Ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197–210.

Winke, P. M., & Aquil, R. (2014). Issues in developing standardized tests of Arabic language proficiency. In *Handbook for Arabic Language Teaching Professionals in the 21st Century* (hal. 221–233). Taylor and Francis. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118374254&partnerID=40&md5=473574df6fcdf60bd5b605dc60ee2be4

HALAMAN INI SENGAJA DIKOSONGKAN