Eloquence Journal of Foreign Language

DOI: https://doi.org/10.58194/eloquence
Language Development Center, IAIN Sultan Amai Gorontalo
https://ejournal.iaingorontalo.ac.id/index.php/ELOQUENCE

E ISSN 2964-9277
Vol. 5 No. 1, April 2026
pp. 95 - 111

The Quality of Test Items in the Arabic Grammar Olympiad: A Deep Learning Perspective

Faiz Nasyith Hakim[1(✉)] Universitas Pendidikan Indonesia, Indonesia[1]
faiznasyith@upi.edu[1]
Hikmah Maulani[2] Universitas Pendidikan Indonesia, Indonesia[2]
hikmahmaulani@upi.edu[2]
Ilham Sofiyan[3] Al-Azhar University, Egypt[3]
ilhmsopiyan@gmail.com[3]

https://doi.org/10.58194/eloquence.v5i1.3497

Corresponding Author:✉ Faiz Nasyith Hakim

Article History
Received
10-03-2026
Accepted:
15-03-2026
Published:
15-04-2026

ABSTRACT

Background: Test instruments play an important role in measuring abilities, including in competitive events such as the Arabic Grammar Olympiad; therefore, standardized, valid test items that support learning outcomes are required to map participants' ability profiles accurately. The Indonesian government also emphasizes the concept of Deep Learning, which focuses not only on cognitive aspects but also on achieving graduate profile outcomes.
Purpose: This study aims to analyze Olympiad test items in terms of validity, reliability, difficulty level, discrimination power, distractor effectiveness, cognitive level, and their alignment with the eight Deep Learning graduate profiles of the Merdeka Curriculum.
Method: This study uses a mixed-method approach, consisting of a quantitative ex post facto design and a qualitative content analysis design.
Results and Discussion: The findings indicate that the Olympiad test items demonstrate high reliability; However, only 43% of the items are valid, 87% are categorized as easy in terms of difficulty level, 57% have poor discrimination power, and only 40% of the distractors are effective, and the cognitive level of the items is still dominated by the Lower Order Thinking Skills (LOTS) category (67%). Furthermore, the test items do not fully represent the eight Deep Learning graduate profiles, as only three profiles are reflected: independence (100%), faith and devotion to God Almighty (33%), and critical reasoning (20%).
Conclusions and Implications: This study concludes that the Arabic Grammar Olympiad test items still require improvement to function optimally as instruments for measuring participants' abilities while supporting the achievement of Deep Learning graduate profiles.

Keywords: *Deep Learning; Quality Questions; Olympiad; Arabic Grammar; Bloom's Taxonomy.*

ABSTRAK

Latar Belakang: Instrumen tes memiliki peran dalam mengukur kemampuan, termasuk dalam ajang kompetitif seperti Olimpiade Kaidah Bahasa Arab. Untuk mengukur peta kemampuan peserta maka diperlukan soal-soal yang terstandar, valid, dan mendukung capaian pembelajaran. Pemerintah Indonesia juga menekankan konsep pembelajaran mendalam yang tidak hanya menekankan aspek kognitif tapi juga capaian profil lulusan.
Tujuan: Penelitian ini bertujuan untuk menganalisis soal olimpiade ditinjau dari aspek validitas, reliabilitas, tingkat kesukaran, daya pembeda, efektivitas pengecoh, level kognitif, serta kaitannya dengan delapan profil lulusan pembelajaran mendalam dalam Kurikulum Merdeka.
Metode: Penelitian ini menggunakan metode campuran, yaitu kuantitatif desain ex post facto dan kualitatif desain analisis isi.
Hasil dan Pembahasan: Hasil analisis menunjukan bahwa soal olimpiade mempunyai nilai reliabilitas yang tinggi, namun soal yang valid baru mencapai 43%, tingkat kesukaran mudah 87%, daya pembeda jelek 57%, dan pengecoh yang efektif hanya 40%. Kemudian, level kognitif soal masih didominasi oleh kategori LOTS (67%). Selain itu, soal olimpiade ini belum merepresentasikan delapan profil lulusan pembelajaran mendalam, hanya tiga profil yang terepresentasikan yaitu; kemandirian (100%), keimanan dan ketakwaan kepada Tuhan Yang Maha Esa (33%), serta penalaran kritis (20%).
Kesimpulan dan Implikasi: Penelitian ini menegaskan bahwa soal olimpiade ini masih memerlukan perbaikan agar dapat berfungsi optimal sebagai alat ukur kemampuan peserta sekaligus mendukung capaian profil lulusan pembelajaran mendalam.

Kata Kunci: *Pembelajaran Mendalam; Kualitas Soal; Olimpiade; Kaidah Bahasa Arab; Taksonomi Bloom.*

INTRODUCTION

A test is an instrument consisting of a set of questions or statements that must be answered, which functions to measure the level of ability, skills, or certain characteristics of individuals who take the test.[1] One commonly used instrument is an objective test [2], such as multiple-choice questions in academic competitions.[3] In Arabic language studies, one academic competition is the olympiad. The olympiad is not only a competitive event but also serves to map students' abilities.[4] One aspect of the students' ability map referred to is Arabic language grammar. One of the organizers of such an olympiad is the Festival Kampung Arab (FIKAR). According to the guideline book, the Arabic Grammar Olympiad tests Arabic language rules, including *nahwu* and *ṣarf*, at the senior high school/equivalent level, reflecting students' grammatical competence. To assess participants' ability mapping, standardized, valid questions that support learning outcomes are required.[3]

Given the need for standardized questions, various research findings indicate that the quality of evaluation instruments, especially test questions, both regular exam and competition questions, still include items that do not meet ideal quality standards.[5], [6] This becomes a problem because it leads to weak item validity, unstable reliability, low discrimination power, and disproportionate difficulty levels.[3], [7] Instrument quality greatly influences whether the questions truly measure participants' abilities or merely produce meaningless scores.[8] On the other hand, the quality of test instruments is also related to cognitive levels. According to the revised Bloom's Taxonomy, ideal learning aims to direct students toward achieving the HOTS (Higher Order Thinking Skills) domain, and this also applies to test.[9]

In an effort to achieve this domain, the Indonesian government's policy emphasizes deep learning, which focuses not only on cognitive aspects but also on building graduate profiles. These profiles include: (1) Faith and devotion to God Almighty, (2) Citizenship, (3) Creativity, (4) Critical reasoning, (5) Collaboration, (6) Communication, (7) Health, and (8) Independence.[10] According to Biggs (1996), in Lestari and Yusuf (2025), assessment questions must be aligned with the applicable curriculum.[11] Therefore, it is important to examine whether the Arabic Grammar Olympiad questions contain elements that support the achievement of the graduate profile competencies of deep learning.

Preliminary findings from the researcher indicate that, out of 30 questions from the FIKAR 2025 Arabic Grammar Olympiad, the first five questions used as the analysis sample fall into the LOTS (Lower Order Thinking Skills) and MOTS (Middle Order Thinking Skills) categories, and none are categorized as HOTS. Thus, a comprehensive analysis of cognitive level distribution is still needed. This indicates that olympiad questions still tend to measure thinking skills at the LOTS and MOTS levels and have not provided sufficient opportunities for participants to develop higher-order thinking skills (HOTS). This issue becomes the focus of the present study.

Various studies show that item analysis is generally conducted to assess the quality of test instruments, both quantitatively and in terms of the level of thinking skills measured. Logayah et al. (2024) and Rashwan et al. (2024) found that most multiple-choice questions were categorized as valid and reliable, had balanced difficulty levels, adequate discrimination power, and functioning distractors. However, improvements were still needed for several items.[3], [7] However, these studies focused mainly on quantitative quality and did not examine the thinking skills measured by the items. On the other hand, studies reviewing the HOTS aspect indicate that questions that measure higher-order thinking skills remain scarce. Musliha et al. (2021), Kadir et al. (2024), and Aulia & Baroroh (2024) revealed that most questions still focus on low to middle-order thinking skills, while questions at the HOTS level—especially evaluation (C5) and creation (C6)—are still very minimal or even not found.[6], [12], [13] Although these studies analyzed cognitive levels, the analyses generally ended at the classification of cognitive levels, without linking them to learning objectives within the context of the Merdeka Curriculum.

Based on these two tendencies, previous studies often separate quantitative quality analysis from cognitive-level analysis. In addition, these studies do not relate the test instrument to the eight dimensions of the deep learning graduate profile. This study expands the analysis of test items by not only evaluating the quantitative quality of questions and cognitive levels according to Bloom's Taxonomy, but also examining the extent to which Arabic Grammar Olympiad questions include elements of deep learning that support the achievement of graduate profile outcomes in Indonesia. Unlike previous studies that generally focused on validity, reliability, difficulty level, discrimination power, distractor effectiveness, or cognitive-level classification, this study introduces a new perspective by linking test instruments to the eight dimensions of the graduate profile within the deep learning framework of the Merdeka Curriculum.

Therefore, this research is important because it will provide implications for the development of evaluation instruments, particularly Arabic grammar questions. This study is expected to enrich studies on item analysis by integrating aspects of instrument quality, cognitive levels of the revised Bloom's Taxonomy, and the concept of deep learning oriented toward graduate profiles. On the other hand, the results of this study can serve as evaluation material and recommendations for question designers of the Arabic Grammar Olympiad to design questions that are not only valid, reliable, and proportionally distributed across cognitive levels, but also support the achievement of the eight dimensions of the deep learning graduate profile. Thus, olympiad questions will not merely function as a competitive selection tool but also as a means to improve the quality of Arabic language learning in line with the direction of national education policy.

LITERATURE REVIEW

Quality of Test Items

Rust (2002) explains that an assessment can be considered truly reliable if it consistently produces the same results regardless of who administers it.[14] Meanwhile, Matazu and Julius (2021) state that, through item analysis, researchers can examine the characteristics of test items and improve test quality.[15] The main function of item analysis procedures is to measure the usefulness of each test item.[16], [17] The primary objective of item analysis is to obtain information about each item's characteristics, and the results of this analysis can be used to determine the quality of the questions.[18], [19] Quantitative analysis is conducted after the questions have been administered and real data are available for analysis.[7] This analysis includes five main indicators: validity, reliability, difficulty level, discrimination power, and distractor effectiveness.[20]

Imaduddin et al. (2022) state that a question that cannot differentiate the abilities of two students may not meet the criteria of a good test item.[8] Validity refers to the level of accuracy and precision of a testing instrument in performing its measurement function. A test can be considered highly valid if the measurement instrument performs its measurement function properly and produces accurate results.[21] Reliability refers to a series of measurements or measuring instruments that show consistency when the measurement is repeated, or the extent to which the measurement results can be trusted.[22] The difficulty level is determined by the proportion of participants who answer a question correctly [23], which then indicates whether the question is too easy or too difficult for students.[15] Discrimination power can be observed by comparing individuals who answer a particular item correctly with the total test score [7], and it is used to identify differences between high-achieving and low-achieving students. This reflects the difference between the percentages of high-achieving and low-achieving students who answer correctly.[15] Distractor effectiveness refers to the quality of distractors in answer choices [7], indicating whether they successfully divert participants' attention from the correct answer. Ideally, all distractors should be sufficiently similar to the correct answer.[24]

Cognitive Level

The questions are then analyzed according to the revised Bloom's Taxonomy by Anderson and Krathwohl, based on their cognitive levels. They state that this cognitive process dimension represents a continuum of increasingly complex thinking, starting from the most basic ability, remembering, to the highest ability, creating.[9] There are six cognitive levels: remembering (C1), understanding (C2), applying (C3), analyzing (C4), evaluating (C5), and creating (C6). These levels are further categorized into LOTS, MOTS, and HOTS. C1 (Remembering) and C2 (Understanding) fall into the LOTS category, C3 (Applying) falls into the MOTS category, while C4 (Analyzing), C5 (Evaluating), and C6 (Creating) belong to the HOTS category.[13], [25], [26]

Graduate Profile in Deep Learning

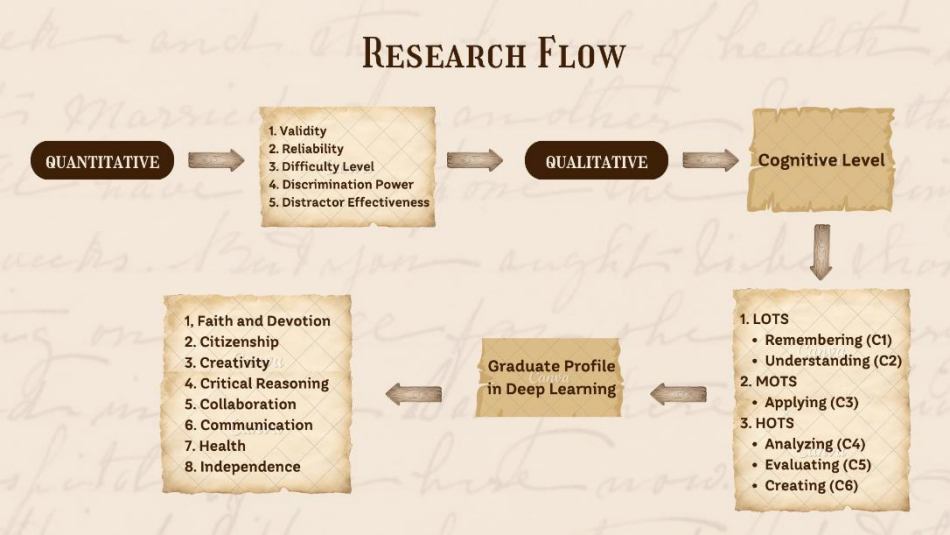
Furthermore, the questions are examined for the presence of deep learning elements that support the achievement of the graduate profile. This analysis refers to the Constructive Alignment theory proposed by Biggs (1996). Biggs emphasizes that all components within the learning system, from the curriculum to the assessment instruments, must be designed in alignment and directed toward learning activities that support the achievement of the intended learning outcomes.[11] Thus, assessment cannot be treated as a standalone component; rather, it must align with the curriculum and the graduate competencies being developed. In the context of the Merdeka Curriculum, this alignment is manifested through the concept of deep learning, which emphasizes the development of eight dimensions of the graduate profile: faith and devotion to God Almighty; citizenship; creativity; critical reasoning; collaboration; communication; health; and independence.

Ideally, these dimensions should be reflected in every learning activity, including assessment.[10]

METHOD

Design

This study employs both quantitative and qualitative methods (Mixed-Method) to obtain a comprehensive understanding of test quality, cognitive levels based on Bloom's Taxonomy, and the presence of deep learning concepts in the test items. First, the quantitative approach emphasizes numerical data, is objective, and analyzes data using statistical techniques.[27] Second, based on the quantitative results obtained, a qualitative approach is conducted to explore the phenomenon in depth through non-numerical data such as words, texts, or symbols, which are analyzed descriptively and interpretatively with emphasis on meaning, context, and the experiences of subjects or the content of a document.[28] The research stages are illustrated in Picture 1.



Picture 1. Research Flow

In the first stage, this research uses a quantitative, ex post facto research design, which observes phenomena based on events that have already occurred, without the researcher's involvement in influencing or manipulating the research object.[29] In this case, the researcher was not involved in preparing the test items or administering the test, but only observed and analyzed existing test results to assess the quality of the items using five indicators: validity, reliability, difficulty level, discrimination power, and distractor effectiveness. After the quantitative stage was completed, the research continued with a qualitative approach using a content analysis design. Content analysis is a research technique used to identify, analyze, and interpret both implicit and explicit meanings in a text or document to obtain a systematic and objective understanding of the message contained in the data.[27]

Data Collection Technique

The data collection technique used in this study was documentation. This technique involved collecting relevant documents, namely the question sheets and participants' answer sheets for question code 2A from the second preliminary round of the Arabic Grammar Olympiad, consisting of 30 multiple-choice items. 21 participants had answered these questions during the olympiad held on Tuesday, November 18, 2025. Since 21 participants took part in the olympiad, the sampling technique used was saturated sampling. This technique was chosen because the population size was relatively small (fewer than 30 individuals), so the entire population was used as the sample.[30]

The documents were obtained directly from the organizing committee of the FIKAR 2025 event held at Universitas Pendidikan Indonesia.

Data Analysis Technique

Processing for the quantitative analysis used Microsoft Excel to determine validity and reliability. Meanwhile, the Anates application was used to determine the difficulty level, discrimination power, and distractor effectiveness. After completing the quantitative analysis stage, the research continued with a qualitative analysis using Krippendorff's (2013) content analysis model. The process began with Unitizing, which involved determining the FIKAR 2025 Arabic Grammar Olympiad questions as the unit of analysis. Next was Sampling, which involved selecting the second preliminary round questions with code 2A as the sample to be analyzed. This was followed by Recording/Coding, in which each item was systematically coded according to Bloom's Taxonomy and the presence of deep learning concepts. The coding results were then processed in the reducing stage by filling in tables on the research instrument sheet, thereby simplifying and systematically presenting the data. Based on these tables, the researcher used inference to interpret the dominance of cognitive levels and deep learning concepts in the items. Finally, the process ended with Narrating, which involved presenting convincing results in the research report.

RESULTS AND DISCUSSION

Analysis of Test Items in The Arabic Grammar Olympiad

The analysis of the test items includes validity testing, reliability testing, difficulty level testing, discrimination power testing, distractor effectiveness testing, and cognitive level analysis. The validity test was conducted in Microsoft Excel and analyzed using dichotomous data, with correct answers scored 1 and incorrect answers scored 0. Item validity was assessed by measuring the relationship between each item's score and the total test score.[21] Therefore, the item validity analysis used a correlation technique, and the results are presented in Table 1.

Table 1. Validity Test Results

Item No.	Correlation	Information
1	NAN	-
2	NAN	-
3	0,364	Valid
4	NAN	-
5	0,214	Invalid
6	NAN	-
7	0,475	Valid
8	-0,011	Invalid
9	0,420	Valid
10	0,262	Invalid
11	-0,011	Invalid
12	0,872	Valid
13	NAN	-
14	0,611	Valid
15	NAN	-
16	-0,339	Invalid
17	0,174	Invalid
18	-0,086	Invalid
19	0,465	Valid
20	0,257	Invalid

21	0,873	Valid
22	NAN	-
23	0,711	Valid
24	0,311	Invalid
25	0,148	Invalid
26	0,474	Valid
27	0,668	Valid
28	0,514	Valid
29	0,370	Valid
30	0,437	Valid

Based on the validity test results, out of 30 items in the FIKAR 2025 Arabic Grammar Olympiad, only 13 items (43%) were declared valid, while 10 items (33%) were categorized as invalid, and 7 items (24%) could not have their validity values calculated (NAN) because all participants answered these questions correctly. These findings indicate that most of the questions do not fully meet the validity criteria as instruments for accurately measuring participants' abilities. Therefore, improvements in item design are still needed to distinguish participants' abilities better. Next, the reliability test was conducted using Microsoft Excel, and the results are presented in Table 2.

Table 2. Reliability Test Results

Formula	Reliability Coefficient	Degree of Reliability
Kr-20	0,708	High
Kr-21	0,639	High
CA	0,692	High

Based on reliability test results using three formulas, KR-20, KR-21, and Cronbach's Alpha, the reliability coefficients ranged from 0.61 to 0.80. KR-20 was used as the main formula because the tested items were multiple-choice questions.[21] Meanwhile, KR-21 and Cronbach's Alpha were used to strengthen and confirm the reliability results. According to the reliability criteria proposed by Guilford and Fruchter [21], this range falls into the high reliability category. These findings indicate that the Arabic Grammar Olympiad test instrument shows good consistency, meaning that the measurement results are trustworthy and relatively stable when used under similar conditions. Next, the difficulty level test was conducted using the Anates application, and the results are presented in Table 3.

Table 3. Difficulty Level Test Results

Item No.	Correct Amount	Difficulty Index	Information
1	21	100	Easy
2	21	100	Easy
3	20	95,24	Easy
4	21	100	Easy
5	20	95,24	Easy
6	21	100	Easy
7	12	57,14	Medium
8	20	95,24	Easy
9	19	90,48	Easy
10	17	90,95	Easy
11	20	95,24	Easy

12	17	80,95	Easy
13	21	100	Easy
14	16	76,19	Easy
15	21	100	Easy
16	14	66,67	Medium
17	10	47,62	Medium
18	20	95,24	Easy
19	15	71,43	Easy
20	19	90,48	Easy
21	16	76,19	Easy
22	21	100	Easy
23	18	85,71	Easy
24	19	90,48	Easy
25	19	90,48	Easy
26	19	90,48	Easy
27	17	80,95	Easy
28	20	95,24	Easy
29	11	52,38	Medium
30	18	85,71	Easy

Based on the difficulty level analysis, most Arabic Grammar Olympiad test items fall into the easy category (87%), with the remaining (13%) in the medium category, and none are classified as difficult. The dominance of easy questions indicates that the test instrument does not yet provide a proportional level of challenge for olympiad participants, leaving their abilities unmeasured. This condition limits variation in participants' responses and may reduce the test's effectiveness as an academic selection tool, limiting its ability to distinguish abilities more sharply. Next, the discrimination power test was conducted using the Anates application, and the results are presented in Table 4.

Table 4. Discrimination Power Test Results

Item No.	Top Group	Bottom Group	Different	DP Index (%)	Information
1	6	6	0	0	Poor
2	6	6	0	0	Poor
3	6	5	1	16,67	Poor
4	6	6	0	0	Poor
5	6	5	1	16,67	Poor
6	6	6	0	0	Poor
7	4	1	3	50	Very good
8	6	6	0	0	Poor
9	6	4	2	33,33	Good
10	6	4	2	33,33	Good
11	6	6	0	0	Poor
12	6	2	4	66,67	Very good
13	6	6	0	0	Poor
14	6	2	4	66,67	Very good

15	6	6	0	0	Poor
16	4	6	-2	-33,33	No Discrimination
17	4	3	1	16,67	Poor
18	6	6	0	0	Poor
19	5	2	3	50	Very good
20	5	5	0	0	Poor
21	6	1	5	83,33	Very good
22	6	6	0	0	Poor
23	6	3	3	50	Very good
24	6	5	1	16,67	Poor
25	6	5	1	16,67	Poor
26	6	4	2	33,33	Good
27	6	3	3	50	Very good
28	6	5	1	16,67	Poor
29	5	2	3	50	Very good
30	6	4	2	33,33	Good

Based on the discrimination power analysis, most Arabic Grammar Olympiad test items fall into the poor category, totaling 17 items (57%). In addition, 8 items (27%) fall into the very good category, 4 items (13%) into the good category, and 1 item is classified as no discrimination or very poor. These findings indicate that the majority of the questions do not yet optimally distinguish between high- and low-ability participants. This condition occurs because many questions were answered correctly by all participants or were answered incorrectly by only one participant. However, a small number of items have demonstrated very good discrimination power and have the potential to serve as models for improving the test instrument's quality. Next, the distractor effectiveness test was conducted using the Anates application, and the results are presented in Table 5.

Table 5. Distractor Effectiveness Test Results

Item No.	A	B	C	D	Answer key	Ineffective Answer
1	0	21	0	0	B	ACD
2	0	21	0	0	B	ACD
3	20	0	0	1	A	BC
4	0	21	0	0	B	ACD
5	20	0	0	1	A	BC
6	0	0	21	0	C	ABD
7	0	0	12	9	C	AB
8	20	0	0	1	A	BC
9	2	0	0	19	D	BC
10	3	0	17	1	C	B
11	1	20	0	0	B	CD
12	3	1	17	0	C	D
13	0	21	0	0	B	ACD
14	1	2	2	16	D	-
15	0	21	0	0	B	ACD

16	5	14	0	2	B	C
17	11	10	0	0	B	CD
18	1	0	0	20	D	BC
19	1	15	5	0	B	D
20	0	19	2	0	B	AD
21	0	3	2	16	D	A
22	21	0	0	0	A	BCD
23	18	1	0	2	A	C
24	2	19	0	0	B	CD
25	1	0	19	1	C	B
26	19	1	1	0	A	D
27	4	0	17	0	C	BD
28	20	1	0	0	A	CD
29	8	2	11	0	C	D
30	1	1	18	1	C	-

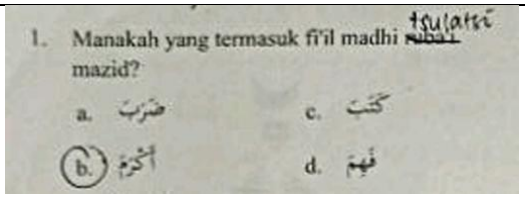
Based on the distractor effectiveness analysis, of the 90 distractor options (excluding the answer keys), only 36 (40%) were considered effective because test participants selected them. Given that 21 participants took the test, a distractor is considered effective if it is selected by at least one participant, in accordance with the 5% threshold criterion.[21] Meanwhile, 54 (60%) distractors were categorized as ineffective because none of the participants selected them. These findings indicate that most distractors do not yet have sufficient similarity to the correct answer, making them less capable of diverting participants' attention and thereby reducing the quality of the test items.

Next, the questions were categorized by cognitive level according to the revised Bloom's Taxonomy by Anderson and Krathwohl. After conducting the analysis, the results are presented in Table 6.

Table 6. Cognitive Level Analysis Results

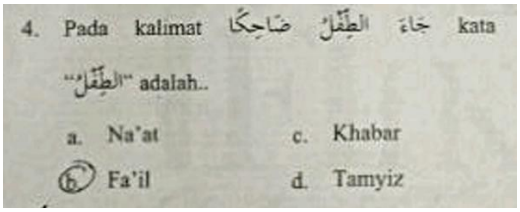
Category	Question Number	Amount	Percentage
C1 – Remembering (LOTS)	1, 8, 13, 14, dan 28	5	17%
C2 – Understanding (LOTS)	2, 4, 7, 10, 11, 16, 17, 18, 19, 20, 21, 22, 23, 24, dan 27	15	50%
C3 – Applying (MOTS)	3, 5, 15, dan 29	4	13%
C4 – Analyzing (HOTS)	6, 9, 12, 25, 26, dan 30	6	20%
C5 – Evaluating (HOTS)	-	0	0%
C6 – Creating (HOTS)	-	0	0%

Based on the analysis results, 20 questions fall into the LOTS category, comprising 5 at the C1 (Remembering) level and 15 at the C2 (Understanding) level. One example of a question at the C1 (Remembering) level is Question 1.



Picture 2. Question Number One

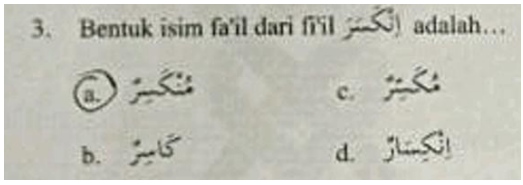
This question requires participants to identify the form of *fi'l māḍī tsulātsī mazīd* among the available options. Cognitively, this question falls at the C1 level because participants only need to recall the pattern (*wazan*) of *fi'l tsulātsī mazīd* to determine the correct answer. The cognitive level of this question is relatively low because the difference between *tsulātsī mazīd* and *tsulātsī mujarrad* can be recognized directly through the additional letters in the verb form. In addition, the distractors used are relatively simple, since all options are *fi'l māḍī* with the *ḍamīr huwa*. Therefore, participants only need to recognize the pattern without conducting a deeper analysis. In the context of an olympiad, questions like this only measure basic mastery of morphological concepts and are less challenging for participants, who generally already possess a good foundational competence in Arabic. Furthermore, one example of a question at the C2 (Understanding) level is Question 4.



Picture 3. Question Number Four

This question asks participants to determine the grammatical function of the word الطِّفْلُ in the sentence جَاءَ الطِّفْلُ ضَاحِكًا. To answer it, participants need to understand the relationship between the elements within the structure of a *jumlah fi'liyah*. The word الطِّفْلُ is an *ism marfū'* that appears after the *fi'l ma'lūm* جَاءَ, and therefore functions as the *fā'il*. Thus, this question belongs to the C2 cognitive level because it requires understanding to infer the grammatical function of a word within a sentence's structure. The cognitive demand of this item is relatively low, since the grammatical clues are quite clear from the word position and the *i'rāb* marker. However, the distractor *tamyīz* is structurally less relevant to the sentence context and can therefore be eliminated relatively easily. This option would be more effective if it were replaced with *nā'ib al-fā'il*, which is grammatically closer to the function of *fā'il*. In the context of an olympiad, a question like this still falls within the basic level of understanding and does not fully challenge the analytical abilities of participants who already possess strong knowledge of *naḥwu*.

Next, there are four Olympiad questions categorized as MOTS, which correspond to the C3 (Applying) cognitive level. One example of a question at the C3 (Applying) level is Question 3.

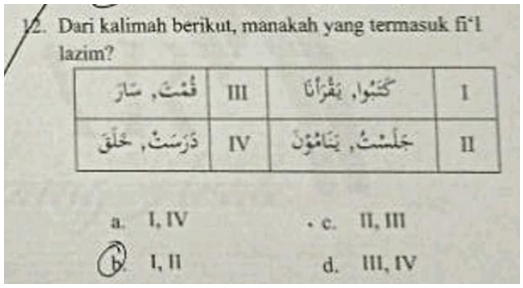


Picture 4. Question Number Three

This question requires participants to determine the form of *ism fā'il* from the verb إِنْكَسَرَ by applying the rules of *ṣarf*. The verb follows the pattern إِنْفَعَلَ, so its *ism fā'il* form is مُنْكَسِرٌ. Therefore, participants must apply their knowledge of morphological patterns to produce the correct word

form, which places this item at the C3 cognitive level. The cognitive demand of this question is considered moderate, since participants need to understand the patterns of word transformation in *ṣarf*. However, in the context of an olympiad, the cognitive requirement of this question remains relatively limited because the solution process is procedural, namely, following a fixed, established pattern of word formation. The question does not yet require more complex morphological analysis, such as forms involving *ḥurūf al-'illah*.

Furthermore, there are six Olympiad questions categorized as HOTS, specifically those at the C4 (Analyzing) cognitive level. However, no questions were found at the C5 (Evaluating) or C6 (Creating) levels. One example of a question at the C4 (Analyzing) level is Question 12.


12. Dari kalimah berikut, manakah yang termasuk fi'l lazim?

قُمْتُ, سَارَ	III	كَتَبُوا, يَقْرَأَانِ	I
دَرَسْتُ, خَلَقَ	IV	جَلَسْتُ, يَنَامُوْنَ	II

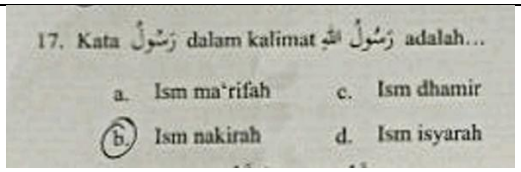
a. I, IV
b. I, II
c. II, III
d. III, IV

Picture 5. Question Number Twelve

This question requires participants to analyze the characteristics of several *fi'l* forms presented in four groups of data. Participants must distinguish between *fi'l lāẓim* and *fi'l muta'addī* for each verb, and then determine which group contains only *fi'l lāẓim*. This process requires participants to organize information and analyze each verb within the groups before determining the correct answer. Therefore, the item belongs to the C4 cognitive level (analyzing). The cognitive demand of this question is relatively higher than the previous questions because participants must evaluate several verbs simultaneously before selecting the correct group. The verbs presented are also more varied in terms of their types and the *ḍamīr* used. However, the analysis required is still limited to the classification of verb types and does not yet involve applying the concept within a broader syntactic context. In the context of an olympiad, this question has begun to measure analytical ability, but it could be further developed to assess higher-order thinking skills better. For example, the task could involve analyzing sentence structures that contain particular verbs, rather than merely identifying word forms in isolation.

Overall, the Arabic Grammar Olympiad questions are still dominated by LOTS-level questions (67%), while MOTS questions account for only 13% and HOTS questions account for 20%. This composition does not align with the recommended proportional distribution of cognitive levels, which suggests 30% HOTS, 40% MOTS, and 30% LOTS.[31] The low proportion of HOTS questions is due to the relatively similar design of many items, such as identifying syntactic positions, types of *jumlah*, types of *ism* or *fi'l*, and the functions of *ḥarf*. This condition indicates that the olympiad's role as a platform for measuring higher-level academic ability has not yet been fully reflected in the design of the questions. Given that olympiad participants are generally top-performing students who have already passed a school selection process, the questions should require greater analytical skills and more complex grammatical reasoning. Therefore, the future development of Olympiad questions should consider a balanced distribution of cognitive levels by increasing the number of HOTS questions, so that the competition does not merely measure basic mastery of Arabic concepts but also critical and analytical thinking skills at a higher level.

After conducting the analysis, the researcher also identified several ambiguities in the Arabic Grammar Olympiad questions, one of which appears in Question 17.

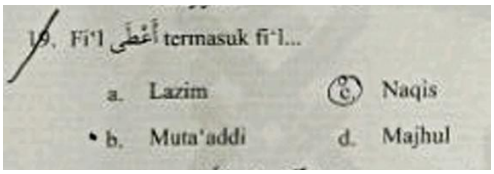

17. Kata رَسُولُ dalam kalimat رَسُولُ اللهِ adalah...

a. Ism ma'rifah c. Ism dhamir

b. Ism nakirah d. Ism isyarah

Picture 6. Question Number Seventeen

In Question 17, participants are required to classify "رَسُولُ" as one of the following types of nouns: *ism ma'rifah*, *nakirah*, *ḍamīr*, or *ism isyārah*. The word "رَسُولُ" does not contain *alif-lām*, is not a pronoun, and is not a demonstrative noun. Therefore, if viewed independently, the word "رَسُولُ" can be categorized as *ism nakirah*, which corresponds to the answer key provided by the committee. However, according to *nahwu* rules, every noun that becomes *muḍāf* to a *ma'rifah* noun will also take the status of *ma'rifah*. In the phrase "رَسُولُ اللهِ", the word "اللهِ" is an *ism ma'rifah*, so the word "رَسُولُ" as *muḍāf* becomes *ism ma'rifah* through the *iḍāfah* construction. The wording of the question does not indicate that the word "رَسُولُ" should be interpreted separately or independently; instead, it is explicitly placed within an *iḍāfah* structure. Based on this analysis, the correct answer according to *nahwu* rules should be *ism ma'rifah*, making the answer key *ism nakirah* less appropriate given the wording of the question. This is also reflected in the participants' responses, with the majority choosing option A (*ism ma'rifah*), as shown in Table 5. Another question identified as having an issue is Question 19.


19. Fi'l أَعْطَى termasuk fi'l...

a. Lazim c. Naqis

b. Muta'addi d. Majhul

Picture 7. Question Number Nineteen

In Question 19, participants are asked to determine whether the verb "أَعْطَى" belongs to *fi'l lāzim*, *fi'l muta'addī*, *fi'l nāqiṣ*, or *fi'l majhūl*. The verb "أَعْطَى" means "to give," an action that semantically requires an object for the sentence to be complete. Therefore, based on its syntactic function, this verb is classified as *fi'l muta'addī*, making option B the answer according to the provided key. However, the verb "أَعْطَى" can also be categorized as *fi'l nāqiṣ* because, in terms of its morphological structure, it contains a *ḥarf 'illah* in the *lām al-fi'l*, namely the letter "ى". As a result, the question potentially has two correct answers. This ambiguity is reflected in the participants' responses, as five participants selected option C (*nāqiṣ*).

The Relationship Between the Test Instrument and the Eight Dimensions of the Graduate Profile

When aligned with the eight dimensions of the deep learning graduate profile, the test instrument serves as a representation of the intended direction of the learning design. The analysis indicates that the Arabic Grammar Olympiad test instrument in FIKAR 2025 covers only some of the eight dimensions of the deep learning graduate profile, as shown in Table 7.

Table 7. Results of the Deep Learning Graduate Profile Analysis

Graduate Profile	Question Number	Amount	Percentage
Faith and devotion to God Almighty	9, 10, 11, 15, 16, 17, 18, 23, 25, dan 27	10	33%
Citizenship	-	0	0%
Creativity	-	0	0%

Critical reasoning	6, 9, 12, 25, 26, dan 30	6	20%
Collaboration	-	0	0%
Communication	-	0	0%
Health	-	0	0%
Independence	All Questions	30	100%

The independence profile appears across all test items because the test is individual, requiring participants to rely on their own abilities to understand the questions and determine the correct answers. In addition to independence, the profile of faith and devotion to God Almighty appears in 10 items (33%). This profile is reflected through the use of Qur'anic verses or Islamic contexts as linguistic objects to be analyzed. These questions not only test participants' grammatical abilities but also connect Arabic language learning with religious values. Furthermore, the critical reasoning profile comprises 6 items (20%), all at the C4 (Analyzing) cognitive level and belonging to the HOTS category. This finding indicates alignment between higher-order cognitive demands in assessment and the critical reasoning profile emphasized in deep learning. However, the limited number of HOTS questions means that this profile has not yet been optimally distributed across the entire test instrument.

Meanwhile, the other five graduate profiles, citizenship, creativity, collaboration, communication, and health, are not yet accommodated in the analyzed test instrument. The absence of the citizenship profile indicates that the questions do not relate Arabic grammar to social contexts, national values, or cultural diversity. The creativity profile also does not appear because all questions are still oriented toward single and closed answers, leaving no room for participants to generate new ideas or construct language independently. The absence of collaboration and communication profiles can be explained by the test format, which is an individual multiple-choice test. However, this also highlights the limitations of objective test instruments in measuring profiles that involve social and interactive dimensions. Meanwhile, the health profile is not represented because the questions focus entirely on cognitive linguistic aspects without linking them to health awareness. From the perspective of constructive alignment, this condition indicates an imbalance between the curriculum's holistic goals and the assessment used to evaluate them.

CONCLUSION AND IMPLICATIONS

This study concludes that the test instrument of the Arabic Grammar Olympiad FIKAR 2025 has not yet fully reflected the characteristics of high-quality assessment and has not optimally supported deep learning. The instrument still needs improvement to accurately and comprehensively measure participants' abilities, particularly in presenting questions aligned with the deep learning graduate profile and that challenge participants at appropriate cognitive levels. This indicates that assessment has not yet been fully positioned as an important component in achieving the deep learning graduate profile. Assessment in competitive events such as olympiads should be designed to fairly differentiate participants' abilities, stimulate higher-order reasoning, and support the achievement of the graduate profile in accordance with the direction of educational policy.

The findings of this study provide both theoretical and practical implications for the development of Arabic language assessment instruments. Theoretically, the results emphasize that test instruments should not be understood merely as tools for measuring cognitive achievement and the technical quality of items, but should also be analyzed in terms of their alignment with the deep learning graduate profile as the curriculum objective. In this way, assessment is positioned as an essential component in the development of students' holistic competencies. Practically, this study can serve as a reference for test developers, teachers, and organizers of Arabic language

olympiads in designing test instruments that are more proportional, contextual, and oriented toward deep learning. Through such improvements, assessment will not only function as a competitive selection tool, but also as a means to develop higher-order thinking skills and strengthen the graduate profile in line with the direction of national education policy.

BIBLIOGRAPHY

[1] A. Harahap, *Evaluasi Pembelajaran Berbasis Hots Dalam Kurikulum Merdeka*. Indramayu: CV. Adanu Abimata, 2024.

[2] M. Laili, "Ketepatan Kontruksi Butir Pilihan Ganda Bahasa Arab," *ALSUNIYAT J. Penelit. Bahasa, Sastra, dan Budaya Arab*, vol. 3, no. 2, pp. 111–124, 2020, doi: https://doi.org/10.17509/alsuniyat.v3i2.25272

[3] D. S. Logayah, M. Ruhimat, R. Arrasyid, and M. R. F. Islamy, "Item analysis of National Geography Olympiad multiple-choice questions (MCQs) in Indonesia," *Cogent Soc. Sci.*, vol. 10, no. 1, p., 2024, doi: https://doi.org/10.1080/23311886.2024.2354971

[4] M. Taufik and R. D. Susanti, "Pelatihan dan pendampingan olimpiade matematika berbasis strategi heuristik," vol. 7, no. 1, pp. 1–2, 2023. https://doi.org/10.31764/jmm.v7i1.12226

[5] J. Setiawan, A. Sudrajat, Aman, and D. Kumalasari, "Development of higher order thinking skill assessment instruments in learning Indonesian history," vol. 10, no. 2, 2021, doi: https://doi.org/10.11591/ijere.v10i2.20796

[6] S. Kadir, S. Sarif, and A. H. N. Fuadi, "Item Analysis of Arabic Thematic Questions to Determine Thinking Level Ability," *ELOQUENCE J. Foreign Lang.*, vol. 3, no. 1, pp. 26–39, 2024, doi: https://doi.org/10.58194/eloquence.v3i1.1498

[7] N. I. Rashwan, S. R. Aref, O. A. Nayel, and M. H. Rizk, "Postexamination item analysis of undergraduate pediatric multiple-choice questions exam: implications for developing a validated question Bank," *BMC Med. Educ.*, vol. 24, no. 1, pp. 1–9, 2024, doi: https://doi.org/10.1186/s12909-024-05153-3

[8] M. F. Imaduddin, H. Maulani, and I. H. Taufik, "Test the Validity and Reliability of Arabic Learning Questions," *Arab. J. Arab. Stud.*, vol. 7, no. 2, pp. 198–207, 2022, doi: https://doi.org/10.24865/ajas.v7i2.523

[9] L. Anderson and D. Krathwohl, *Kerangka Landasan Untuk Pembelajaran, Pengajaran, Dan Asesmen Revisi Taksonomi Pendidikan Bloom*. Yogyakarta: Pustaka Pelajar, 2015.

[10] Kemendikdasmen RI, "Pembelajaran Mendalam Menuju Pendidikan Bermutu untuk Semua," 2025.

[11] S. Lestari and F. N. Yusuf, "Aligning Assessment Practices with Learning Objectives : A Case of EFL Classes in Indonesia," vol. 10, no. May, pp. 145–163, 2025. https://doi.org/10.21093/ijeltal.v10i1.1973

[12] S. Musliha, D. Sudana, and Y. Wirza, "The Analysis of Higher Order Thinking Skills (HOTs) in the Test Questions Constructed by English Teachers," vol. 595, no. Icollite, pp. 610–617, 2021. https://doi.org/10.2991/assehr.k.211119.095

[13] N. Aulia and R. U. Baroroh, "Innovation of Qawaid Nahwiyah Assessment in Arabic Textbooks Class XII Based on HOTS Assessment," vol. 02, no. 2021, 2024, doi: https://doi.org/10.32332/ijalt.v6i01.8678

[14] Rust, "Learning and Teaching Briefing Papers Series:Theories of learning," p. 3, 2002.

[15] S. Matazu and E. Julius, "Item Analysis: A Veritable Tool for Effective Assessment in

Teaching and Learning," *J. Educ. Pract.*, vol. 12, no. 21, pp. 22–28, 2021, doi: https://doi.org/10.7176/JEP/12-21-04

[16] S. Bhattacherjee, A. Mukherjee, K. Bhandari, and A. Rout, "Physical Violence Against Doctors: A Content Analysis from Online Indian Newspapers," *Indian J. Community Med.*, vol. 42, no. 1, pp. 147–50, 2022, doi: 10.4103/ijcm.IJCM.

[17] I. R. N. Fauziah, Syihabudin, and A. Sopian, "Analisis Kualitas Tes Bahasa Arab Berbasis Higher Order Thingking Skill (HOTS)," vol. 10, no. 1, pp. 45–54, 2020. https://doi.org/10.22373/ls.v10i1.7805

[18] M. Nojomi and M. Mahmoudi, "Assessment of multiple-choice questions by item analysis for medical students' examinations," *Res. Dev. Med. Educ.*, vol. 11, no. 1, p. 24, 2022, doi: https://doi.org/10.34172/rdme.2022.024

[19] Sahrani, S. Herlina, Sumin, and Hermansyah, "Evaluating The Effectiveness Of Arabic Language Summative Tests," vol. 9, no. 2, pp. 190–206, 2024. https://doi.org/10.28918/alsinatuna.v9i2.6858

[20] V. N. I. Sari, A. P. Y. Utomo, and Sumarwati, "Kualitas Soal Bahasa Indonesia di SMP Muhammadiyah 1 Pontianak: Analisis Butir Soal," *J. Pendidik. Bhs. dan Sastra Indones.*, vol. 11, no. 2, pp. 112–119, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/jpbsi/article/view/24018

[21] R. Istikomah, M. Al Farisi, and H. Maulani, "Classpoint AI : Kualitas Empiris Pembuat Soal Otomatis dalam Assessmen Bahasa Arab Classpoint AI : Empirical Quality of Automatic Item Test Generator in Arabic Foreign Language Assessment," vol. 8, no. 1, pp. 44–64, 2025, doi: https://doi.org/10.26555/insyirah.v8i1.13163

[22] B. S. Hariati *et al.*, "Analisis Kualitas Butir Soal Pilihan Ganda Elemen Dokumen Berbasis Digital Menggunakan ANATES," vol. 6, no. 01, pp. 46–58, 2026. https://doi.org/10.57008/jjp.v6i01.1956

[23] I. I. Umiyati and F. Fakhriyah, "Analisis Kualitas Butir Soal Konsep Sistem Gerak Manusia Kelas VI Sekolah Dasar Berdasarkan Validitas , Reliabilitas , Tingkat Kesukaran dan Daya Beda pada siswa . Miskonsepsi sering kali tidak terdeteksi jika butir soal yang digunakan tidak kualitas bank soal di sekolah . Penggunaan teknologi informasi , seperti Microsoft Excel , dan akurat . Syaifudin (2023) menyebutkan bahwa analisis butir soal secara mandiri oleh guru," no. November 2025, 2026.

[24] A. A. Rezigalla *et al.*, "Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items," *BMC Med. Educ.*, vol. 24, no. 1, pp. 1–7, 2024, doi: https://doi.org/10.1186/s12909-024-05433-y

[25] K. Masrifah, M. Ahsanuddin, and M. Ainin, "The Effectiveness of HOTS-Based Practice Questions in Enhancing Students ' Critical Thinking in Arabic Learning," *ALSUNIYAT J. Penelit. Bahasa, Sastra, dan Budaya Arab*, vol. 8, no. 2, 2025, doi: https://doi.org/10.17509/alsuniyat.v8i2.87847 ALSUNIYAT.

[26] A. Fudhaili, N. Nurjannah, and A. Arifin, "Integrating Kahoot to Improve Reading Skills through HOTS-Based and Religious Moderation-Oriented Materials," vol. 8, no. 2, 2025.

[27] A. Fauzy *et al.*, *Metodologi Penelitian*. 2022.

[28] R. Abubakar, *Pengantar metodologi penelitian*. Yogyakarta: SUKA-Press UIN Sunan Kalijaga, 2021.

[29] B. I. Sappaile, "Konsep Penelitian Ex-Post Facto," vol. 1, pp. 105–113, 2010.

[30] A. Veronica, M. Abas, and N. Hidayah, *Metodologi penelitian kuantitatif*. 2022.

[31] A. Susanto and P. Rahmah, "Cognitive Level Analysis of Problems in The Worksheets of Students (WS) Mathematics of Junior High School Analisis Tingkat Kognitif Soal Pada Lembar Kerja Peserta Didik (LKPD) Matematika MTs," vol. 3, no. 1, pp. 75–85, 2021. https://doi.org/10.24252/ajme.v3i1.20941