LANGUAGE CIRCLE: Journal of Language and Literature 19 (1) October 2024 p-ISSN 1858-0165
Available online at https://journal.unnes.ac.id/journals/LC e-ISSN 2460-853X

Analysis of Mid-Semester Assessment Questions for Class XI Arabic at MAS Al-Islam Petala Bumi, Indragiri Hulu

Neni Naqiyah
UIN Maulana Malik Ibrahim Malang, Indonesia
neni.naqiyah18@gmail.com

Latifah Handayani
UIN Maulana Malik Ibrahim Malang, Indonesia
latifahhandayani577@gmail.com

Abdul Basid
UIN Maulana Malik Ibrahim Malang, Indonesia
abdulbasid@bsa.uin-malang.ac.id

Abstract

Question Item Analysis is still very much needed by teachers to create quality measuring instruments for learning outcomes and provide real measuring results. This research aims to analyze validity, reliability, level of difficulty, distinguishing power, and distracting power. The method used in this research is descriptive quantitative. The data source in this research is the student answer sheets in the mid-semester assessment of class XI Arabic in MAS Al-Islam, Petala Bumi, Indragiri Hulu 2022-2023 school year. Data was collected using documentation and analyzed using the ANATES program. The results of this research include 1) Validity: 36% of questions are valid and 64% are invalid. 2) Reliability: reliability coefficient value 0.20 (low reliability). 3) Difficulty level: 56% difficult questions, 44% medium questions, and 0% easy questions. 4) Differentiation power: 20% of questions are very good, 0% of questions are good, 36% of questions are fair, 28% of questions are bad, and 16% of questions are very bad. 5) Distracting power: 6% of questions are very good, 28% of questions are good, 44% of questions are fair, 20% of questions are bad, and 2% of questions are very bad. The mid-semester assessment questions for Arabic class

Keywords: differentiation power, difficulty level, distracting power, reliability, validity

INTRODUCTION

Question Item Analysis is still very much needed by teachers to create quality measuring instruments for learning outcomes (Mahmudi, 2020, p. 137). As stated in the National Education Standards, article 28:3 points (a) that the pedagogical competence that a teacher must have is not only competence in managing students' understanding, designing and implementing learning, and developing students' potential but pedagogical competence This also includes competencies in designing and implementing learning evaluations so that they can improve student learning outcomes (Setyawan & Fathoni, 2017, pp. 144–145). For this reason, after the evaluation tool has been prepared and the teacher must analyze it, it produces a quality learning outcome measuring tool or test that can measure student learning outcomes accurately (R. Y. Kurniawan, Prakoso, Hakim, Dewi, & Widayanti, 2017, p. 180).

Based on the facts that exist in the process of evaluating student learning outcomes, most teachers do not carry out item analysis (Nuraeni, Simarmata, Sukmaningthias, & Sari, 2021, p. 16). This is caused by the teacher's low willingness and ability to analyze the questions so that teachers do not know the strengths and weaknesses of the questions given to students (Muhson, Lestari, Supriyanto, & Baroroh, 2015, p. 200). Santosa and Badawi added that this could also result in low student learning outcomes because teachers do not

know whether the questions applied can truly measure their students' abilities (Santosa & Badawi, 2022, p. 1680).

This is the case in the questions used to measure students' Arabic language skills. Most Arabic teachers do not analyze the questions before they are applied to measure their students' Arabic language skills (Mahmudi, Yasin, & Ardiyanti, 2024, p. 2545). Mahmudi emphasized that item analysis is an absolute thing that must be carried out in preparing valid and reliable questions (Mahmudi, 2020). Brownlie, Burke, and Laan mentioned five indicators in summative assessment consisting of validity, reliability, fairness, authenticity, and flexibility (Brownlie, Burke, & Laan, 2024, p. 30). Different from Wulan and Rusdiana stated that analyzing the questions includes five minimum requirements so that the test can be categorized as good, namely, validity, reliability, distinguishing power, level of difficulty, and distracting power (Wulan & Rusdiana, 2014, p. 234). In line with Norman Grounlund who said that test items are said to be of high quality if they meet validity, reliability, distinguishing power, level of difficulty, and distracting power (Azmi, 2023, p. 2).

Mid-semester assessment (PTS) or what is called mid-term exam (UTS) before the 2013 curriculum is an activity carried out by teachers to measure students' achievement of competency levels after carrying out 8-9 weeks of learning activities (DIKDAS, 2021, p. 136). The scope of the mid-semester assessment includes all indicators that represent all basic competencies in that period. The questions used for the Arabic mid-semester assessment must also be of high quality. Because mid-semester assessments can provide feedback to students, build their motivation to learn Arabic and prepare students for end-of-semester assessments. Apart from that, the mid-semester assessment can also help teachers find out the achievements of Arabic language learning during that period and identify the extent of students' understanding of the material that has been taught during the half semester so that in the end the teacher can correct things that need to be improved or replaced in the half semester. semester after that (Direktorat Pembinaan Sekolah Menegah Atas, 2017, pp. 12–13).

Kumar and his colleagues defined item analysis as a relatively simple and valuable procedure that provides a method for analyzing observations, interpretations of students' attained knowledge, and information regarding the quality of test items (Kumar, Jaipurkar, Shekhar, Sikri, & Srinivas, 2021, p. S87). In line with this understanding, Zaenal Arifin stated that item analysis or test quality analysis is a stage that must be taken to determine the degree of quality of the questions both as a whole and the items that are part of the test (Arifin, 2011, p. 246). From the explanation above, it can be concluded that analysis of Arabic language mid-semester assessment items is an activity to determine the level of goodness of Arabic language items contained in the mid-semester assessment (PTS) starting from validity, reliability, distinguishing power, level of difficulty, and strength. Distractor so that we can use the resulting information to improve the questions.

In the realm of evaluation, analyzing the quality of each test is very minimal, as is the case at MAS Al-Islam Petala Bumi, Indragiri Hulu, especially in Arabic language material. Based on the results of interviews with Arabic language teachers at the school, it was stated that the Arabic UTS questions were made by themselves and they had carried out an analysis of the quality of the questions made, but only on content validity (content analysis) while empirical validity, reliability, level of difficulty, power differentiating, and distracting power have not been implemented. Moreover, the Arabic

language teacher is not a graduate of Arabic language education, but is a graduate of economics and only has the Arabic language skills he had when he was still at school. Based on the results of the researchers's observations of the Arabic language learning scores of class XI students, the scores were still low. This shows that the questions or tests are classified as difficult for students, so they are unable to differentiate between students who have high and low Arabic language skills.

Research on item analysis has been carried out previously, including analysis of multiple-choice items in terms of level of difficulty, differentiation, distractibility, and reliability (Kumar et al., 2021), analysis of the level of difficulty and differentiating power of English summative test items (Alareifi, 2023), item analysis of multiple-choice questions (MCQ-Type A) for the summative assessment of Professional Examination (Htoon & Aung, 2024), the impact of differentiating power on the level of difficulty and differentiating power of multiple-choice questions (Rezigalla et al., 2024), an indicator of the quality of summative assessments that teachers make effectively (Brownlie et al., 2024), the influence of distracting power on the level of difficulty and differentiating power of multiple-choice questions (Chauhan, Chauhan, Vaza, & Chauhan, 2023), analysis of Arabic language midterm exam questions (Mahmudi, Nurwardah, Rochma, & Nurcholis, 2023), and analysis of end-of-year assessment questions for grade VII Arabic (Tanjung, Fahmi, Rahmanita, Filzafati, & Qomari, 2024).

What this research has in common with the first research is that they both analyze the level of difficulty, differentiation, distractibility, and reliability of the test, while the difference is that this research also analyzes the validity of the test, and the object and place of research are different (Kumar et al., 2021). The second study has similarities in analyzing the level of difficulty and differentiating power of summative test items, while the differences in this research will also analyze validity, reliability, and distracting power, as well as different research objects and places (Alareifi, 2023). The third study has similarities in analyzing the level of difficulty and differentiating power of multiple-choice questions, while the difference is this research will also analyze validity, reliability, and distracting power, as well as different research objects and places (Htoon & Aung, 2024). The fourth study has similarities in analyzing distracting power, level of difficulty, and differential power in multiple-choice questions, while the difference is this research will also analyze validity and reliability, using descriptive quantitative methods as well as different research objects and places (Rezigalla et al., 2024). The fifth study, the seventh study has similarities in the indicators of effective summative assessment, namely validity and reliability, while the difference is this research will analyze five indicators in the summative assessment items (validity, reliability, level of difficulty, distinguishing power and distracting power, as well as different research objects and places (Brownlie et al., 2024). The sixth study has similarities in analyzing the difficulty index, differentiation power, and distracting power of multiple-choice questions using Epi-Info 7TM software. While the difference is this research will also analyze the validity and reliability analyzed with ANATES software as well as different research objects and places (Chauhan et al., 2023). The seventh study has similarities in analyzing Arabic language questions in terms of validity, reliability, level of difficulty, and differentiability. While the difference is this research will also analyze the power of distraction, as well as different research

objects and places (Mahmudi et al., 2023). The eighth study has similarities in analyzing Arabic language questions in terms of validity, reliability, level of difficulty, distinguishing power, and distracting power. While the difference is the object of this research is the 8th-grade mid-semester assessment questions and different research locations (Tanjung et al., 2024).

Based on the similarities and differences above, the position of this research is to find findings, namely five specifications or indicators of good test quality, namely the level of validity, reliability, different power, level of difficulty, and distracting power of mid-semester assessment (PTS) items in the form of multiple-choice questions in class XI Arabic subjects at MAS Al-Islam Petala Bumi, Indragiri Hulu.

Based on the problems above, this research aims to analyze the Arabic language mid-semester assessment (PTS) questions for class XI It is hoped that the results of this analysis will determine the quality of the mid-semester assessment questions (PTS) for class XI Arabic at MAS Al-Islam Petala Bumi, Indragiri Hulu includes of validity, reliability, difficulty level, differentiation power, and distracting power. Apart from that, it can also provide valuable input for Arabic teachers in designing questions that truly measure Arabic students' language skills and improve the quality of Arabic language learning in schools.

METHODOLOGY

The type of research used in this research is descriptive quantitative research. Researchers use a quantitative approach because it is a method used to display data and results related to calculations, numbers, and statistical analysis (Sugiyono, 2010, p. 13), and researchers want to analyze the mid-semester assessment questions in depth. Descriptive research is research conducted to determine the value of independent variables, either one or more variables (independent) without making comparisons or connecting them with other variables (Siregar, 2018, p. 15). Descriptive research has the main goal of presenting enough data through describing the present to clarify understand and guide the future and develop conclusions through what the data shows.

The data source in this research is the student answer sheets in the mid-semester assessment of class XI Arabic subjects in MAS Al-Islam, Petala Bumi, Indragiri Hulu 2022-2023 school year. In this research, data was obtained through open interviews and documentation. Interviews were conducted with Arabic language teachers at MAS Al-Islam, Petala Bumi Indragiri Hulu to obtain data used in the preliminary study. Apart from that, researchers also use documentation to obtain data about question sheets, student answer sheets, and the grid for preparing questions for the mid-semester assessment of Class XI Arabic subjects which were tested in MAS Al-Islam, Petala Bumi, Indragiri Hulu 2022-2023 school year. After the data was collected, the author carried out data analysis using the calculation method via a computer program, namely the ANATES program version 4.0.9. The items analyzed consisted of 50 multiple-choice questions with four options (A, B, C, D) analyzed for the level of validity, level of reliability, level of difficulty, distinguishing power, and distracting power using the specified item analysis formula

RESULT AND DISCUSSION

Result

Validity

Writers are calculating the validity of PTS multiple-choice questions in Arabic language class XI at MAS Al-Islam Petala Bumi, Indragiri Hulu carried out using the ANATES V4 program. The calculation results are then categorized into predetermined correlation significance limits. If r table ≤ r count, then the question item is

valid. But if r table ≥ r count, then the question item is invalid(Siregar, 2018, p. 77). In this study, there were 50 questions with a significance level of 5% so the r table was 0.273.

After the analysis results were categorized, it was discovered that the questions analyzed had 18 questions categorized as valid, and 32 questions categorized as invalid. The complete validity analysis data is in Figure 1. Next, the analysis results are made into percentages based on category. The percentage results of the validity analysis of the items in the Arabic PTS multiple-choice questions are presented in Table 1 as follows:

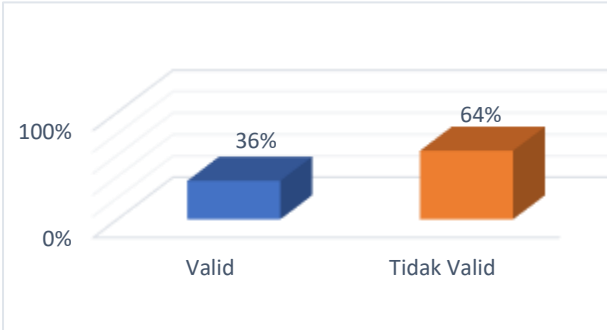
No Butir Baru	No Butir Asli	Korelasi	Signifikansi	No Butir Baru	No Butir Asli	Korelasi	Signifikansi
1	1	0,081	-	26	26	0,121	-
2	2	-0,495	-	27	27	0,481	Sangat Signifikan
3	3	0,509	Sangat Signifikan	28	28	0,098	-
4	4	0,203	-	29	29	-0,069	-
5	5	0,122	-	30	30	0,043	-
6	6	0,122	-	31	31	0,031	-
7	7	0,530	Sangat Signifikan	32	32	0,280	Signifikan
8	8	-0,313	-	33	33	0,031	-
9	9	-0,069	-	34	34	-0,273	-
10	10	0,043	-	35	35	0,181	-
11	11	0,531	Sangat Signifikan	36	36	0,121	-
12	12	0,081	-	37	37	0,131	-
13	13	0,031	-	38	38	0,331	Signifikan
14	14	-0,244	-	39	39	0,306	Signifikan
15	15	0,420	Sangat Signifikan	40	40	0,281	Signifikan
16	16	0,328	Signifikan	41	41	-0,229	-
17	17	0,221	-	42	42	0,081	-
18	18	0,187	-	43	43	0,333	Signifikan
19	19	0,464	Sangat Signifikan	44	44	0,117	-
20	20	0,144	-	45	45	0,372	Sangat Signifikan
21	21	0,680	Sangat Signifikan	46	46	0,280	Signifikan
22	22	-0,112	-	47	47	0,238	-
23	23	0,509	Sangat Signifikan	48	48	0,517	Sangat Signifikan
24	24	0,122	-	49	49	0,509	Sangat Signifikan
25	25	-0,069	-	50	50	0,131	-

Figure 1. Validity Analysis of Arabic PTS Questions

Table 1. Percentage of Validity Analysis of Arabic PTS Question Items

No	Validity Criteria	Question Items	Amount	Percent
1	> 0.273 (Valid)	3, 7, 11, 15, 16, 19, 21, 23, 27, 32, 38, 39, 40, 43, 45, 46, 48, 49	18	36 %
2	< 0.273 (Invalid)	1, 2, 4, 5, 6, 8, 9, 10, 12, 13, 14, 17, 18, 20, 22, 24, 25, 26, 28, 29, 30, 31, 33, 34, 35, 36, 37, 41, 42, 44, 47, 50	32	64 %

Based on the table above, the researchers present the question validity chart as follows:



Reliability

Reliability analysis of PTS multiple-choice questions Arabic language class XI at MAS Al-Islam Petala Bumi, Indragiri Hulu carried out using the ANATES V4 program. The results of the reliability analysis are then categorized based on reliability criteria. The criteria used to interpret reliability are shown in Table 2 as follows(Arikunto, 2006):

Table 2. Classification of Reliability Criteria

No	Reliability Criteria	Interpretation
1	0.801 – 1,000	Very high
2	0.601 – 0.800	Tall
3	0.401 – 0.600	Currently
4	0.201 – 0.400	Low
5	0.000 – 0.200	Very low

```
Rata2= 13,46
Simpang Baku= 3,80
KorelasiXY= 0,11
Reliabilitas Tes= 0,20
```

Figure 2. Reliability Analysis of Arabic PTS Questions

Based on the results of the ANATES calculations above, the reliability of the Arabic PTS questions is 0.20. This can be interpreted as the Mid-Semester Assessment (PTS) questions for class XI Arabic at MAS Al-Islam Petala Bumi, Indragiri Hulu it could be said to be less reliable in the very low category. These results indicate that the Arabic Mid-Semester (PTS) assessment questions cannot be reused (reliable) because they tend to give inconsistent results.

Difficulty Level

Analysis of the level of difficulty in PTS Arabic multiple-choice questions for class XI at MAS Al-Islam Petala Bumi, Indragiri Hulu carried out using the ANATES V4 program. The results of the difficulty level analysis are then categorized based on the difficulty level criteria. The criteria used to interpret the level of difficulty of the questions shown in Table 3 are as follows (Bhat & Prasad, 2021):

Table 3. Difficulty Level Criteria

No	Difficulty Level Criteria	Interpretation
1	< 0.30	Hard
2	0.30 – 0.70	Currently
3	> 0.70	Easy

After the results of the analysis were categorized, it was discovered that the questions analyzed had 28 questions in the difficult category, 22 questions in the medium category, and 0 questions in the easy category. Complete difficulty level analysis data is in Figure 3. The percentage of difficulty level analysis of multiple-choice questions is shown in Table 4 as follows:

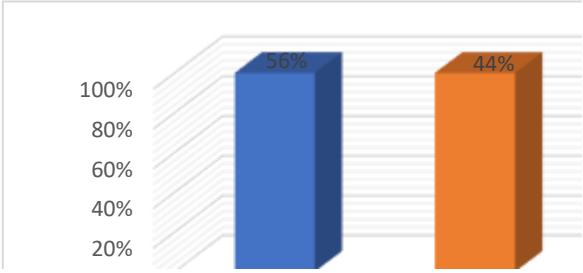
No Butir Baru	No Butir Asli	Jml Betul	Tkt. Kesukaran %	No Butir Baru	No Butir Asli	Jml Betul	Tkt. Kesukaran %
1	1	3	23,08	26	26	2	15,38
2	2	4	30,77	27	27	3	23,08
3	3	4	30,77	28	28	4	30,77
4	4	5	38,46	29	29	3	23,08
5	5	1	7,69	30	30	1	7,69
6	6	1	7,69	31	31	3	23,08
7	7	2	15,38	32	32	1	7,69
8	8	4	30,77	33	33	3	23,08
9	9	3	23,08	34	34	5	38,46
10	10	1	7,69	35	35	3	23,08
11	11	3	23,08	36	36	2	15,38
12	12	3	23,08	37	37	3	23,08
13	13	3	23,08	38	38	3	23,08
14	14	6	46,15	39	39	6	46,15
15	15	5	38,46	40	40	4	30,77
16	16	7	53,85	41	41	2	15,38
17	17	6	46,15	42	42	3	23,08
18	18	8	61,54	43	43	5	38,46
19	19	4	30,77	44	44	7	53,85
20	20	4	30,77	45	45	4	30,77
21	21	5	38,46	46	46	1	7,69
22	22	2	15,38	47	47	2	15,38
23	23	4	30,77	48	48	6	46,15
24	24	1	7,69	49	49	4	30,77
25	25	3	23,08	50	50	3	23,08

Figure 3 Analysis of Difficulty Level of Arabic PTS Questions

Table 4. Percentage Analysis of Difficulty Levels for Arabic PTS Questions

No	Difficulty Level Criteria	Question Items	Amount	Percent
1	< 0.30 (Difficult)	1, 5, 6, 7, 9, 10, 11, 12, 13, 22, 24, 25, 26, 27, 29, 30, 31, 32, 33, 35, 36, 37, 38, 41, 42, 46, 47, 50	28	56 %
2	0.30 – 0.70 (Medium)	2, 3, 4, 8, 14, 15, 16, 17, 18, 19, 20, 21, 23, 28, 34, 39, 40, 43, 44, 45, 48, 49	22	44 %
3	> 0.70 (Easy)	-	0	0 %

Based on the table above, the researchers present a chart of the difficulty levels of the questions as follows:



Different power

Calculation of the differentiating power of multiple-choice questions in PTS Arabic for class XI at MAS Al-Islam Petala Bumi, Indragiri Hulu carried out using the ANATES V4 program. The calculation of the differentiating power analysis is then categorized based on the differentiating power criteria. The criteria used to interpret differential power in this study are shown in Table 5 as follows (Htoon & Aung, 2024, p. 1453):

Table 5. Differentiating Power Criteria

No	Differential Power Criteria	Interpretation
1	> 0.40	Very good, accepted
2	0.30 – 0.39	OK, the problem is accepted but needs to be fixed
3	0.20 – 0.29	Enough, revised matter
4	0.00 – 0.19	Ugly, about not being used/thrown away
5	Negative	Very Ugly, about not being used/thrown away

After the analysis results were categorized, it was discovered that the multiple-choice questions analyzed for their differentiating power contained 10 questions in the very good category, 0 questions in the good category, 18 questions in the fair category, 14 questions in the poor category, and 8 questions in the very bad category. The complete differentiating power analysis data is in Figure 4. Next, the analysis results are made into percentages based on the differentiating power category. The percentage of results of the analysis of the differentiating power of Arabic PTS multiple-choice questions analyzed using ANATES V4 can be seen in Table 6 as follows:

No Butir Baru	No Butir Asli	Kel. Atas	Kel. Bawah	Beda	Indeks DP %	No Butir Baru	No Butir Asli	Kel. Atas	Kel. Bawah	Beda	Indeks DP %
1	1	0	0	0	0,00	26	26	1	0	1	25,00
2	2	1	2	-1	-25,00	27	27	2	0	2	50,00
3	3	2	0	2	50,00	28	28	1	1	0	0,00
4	4	1	1	0	0,00	29	29	1	2	-1	-25,00
5	5	0	0	0	0,00	30	30	0	0	0	0,00
6	6	0	0	0	0,00	31	31	2	1	1	25,00
7	7	2	0	2	50,00	32	32	1	0	1	25,00
8	8	1	2	-1	-25,00	33	33	1	1	0	0,00
9	9	1	1	0	0,00	34	34	1	2	-1	-25,00
10	10	0	0	0	0,00	35	35	1	0	1	25,00
11	11	2	0	2	50,00	36	36	1	0	1	25,00
12	12	1	1	0	0,00	37	37	1	0	1	25,00
13	13	1	1	0	0,00	38	38	1	0	1	25,00
14	14	1	2	-1	-25,00	39	39	3	2	1	25,00
15	15	2	1	1	25,00	40	40	2	1	1	25,00
16	16	2	1	1	25,00	41	41	0	1	-1	-25,00
17	17	2	1	1	25,00	42	42	1	1	0	0,00
18	18	3	2	1	25,00	43	43	3	1	2	50,00
19	19	3	1	2	50,00	44	44	2	3	-1	-25,00
20	20	1	1	0	0,00	45	45	1	0	1	25,00
21	21	4	0	4	100,00	46	46	1	0	1	25,00
22	22	1	1	0	0,00	47	47	1	0	1	25,00
23	23	2	0	2	50,00	48	48	3	1	2	50,00
24	24	1	0	1	25,00	49	49	2	0	2	50,00
25	25	0	1	-1	-25,00	50	50	1	0	1	25,00

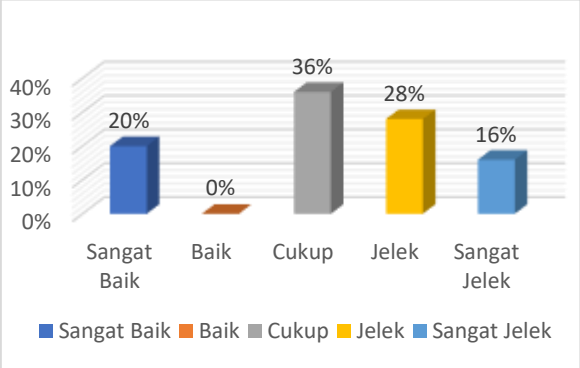
Figure 4 Analysis of the Differentiating Power of Arabic PTS Questions

Table 6. Percentage Analysis of Differentiating Power of Arabic PTS Question Items

No	Differentiating Power Criteria	Question Items	Amount	Percent
1	> 0.40 (Very Good)	3, 7, 11, 19, 21, 23, 27, 43, 48, 49	10	20 %
2	0.30 – 0.39 (Good)		0	0 %
3	0.20 – 0.29 (Enough)	15, 16, 17, 18, 24, 26, 31, 32, 35, 36, 37, 38, 39, 40,	18	36 %

		45, 46, 47, 50		
4	0.00 – 0.19 (Bad)	1, 4, 5, 6, 9, 10, 12, 13, 20, 22, 28, 30, 33, 42	14	28 %
5	Negative (Very Bad)	2, 8, 14, 25, 29, 34, 41, 44	8	16 %

Based on the table above, the researchers present a chart of the different power questions as follows:



Distracting power

Analysis of the effectiveness of distractors in PTS Arabic multiple-choice questions for class XI in MAS Al-Islam Petala Bumi, Indragiri Hulu carried out using the ANATES V4 computer program. The results of the analysis are then categorized using distractor effectiveness criteria. A distractor can be said to work well if it is chosen by at least 5% of test participants. So it can be concluded that the questions analyzed will be effective if 1 or more students are selected because 1 student constitutes 5% of the total sample in this study.

After the analysis results were categorized, it was discovered that 3 questions had very good distractor effectiveness, 14 questions had good distractor effectiveness, 22 questions had fairly good distractor effectiveness, 10 questions had poor distractor effectiveness, and 1 item had very poor distractor effectiveness. Complete data on the effectiveness of distractors can be seen in Figure 5. Next, the results of the analysis are made into percentages based on the category of distractor effectiveness. The percentage of effectiveness of multiple-choice question distractors analyzed using ANATES V4 can be seen in Table 8 as follows:

Table 7. Criteria for Distracting Power

No	Criteria for Distracting Power	Interpretation
1	**	Answer key
2	++	Very good
3	+	Good
4	-	Enough
5	--	Bad
6	---	Very bad

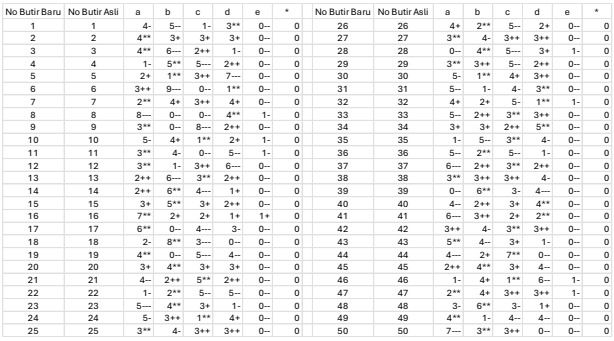


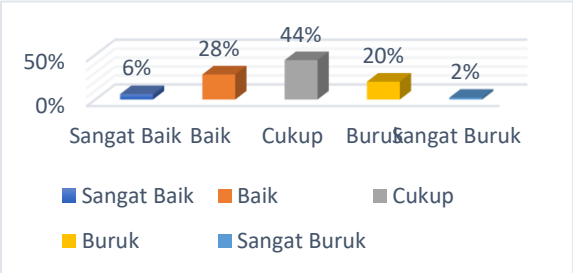
Figure 5 Analysis of the Distracting Power of Arabic PTS Questions

Table 8. Percentage Analysis of the Distracting Power of Arabic PTS Question Items

No	Criteria for Distracting Power	Question Items	Amount	Percent
1	Very good	10, 32, 47	3	6 %
2	Good	2, 7, 15,	14	28 %

		16, 20, 24, 25, 27, 30, 34, 38, 42, 46, 48		
3	Enough	1, 3, 4, 5, 11, 12, 13, 14, 21, 23, 26, 28, 29, 31, 33, 35, 37, 39, 40, 41, 43, 45	22	44 %
4	Bad	6, 8, 9, 17, 18, 22, 36, 44, 49, 50	10	20 %
5	Very bad	19	1	2 %

Based on the table above, the researchers present a chart of the distracting power of the questions as follows:



Discussion

Item analysis must be carried out regularly to create good questions so that Arabic PTS questions can be used to evaluate students' cognitive skills effectively. The analyzed questions will provide feedback to teachers about their educational and teaching actions. Designing multiple-choice questions is a complex and time-consuming process that involves multidisciplinary and integrated curricula (Bhat & Prasad, 2021, p. 344).

Validity

In this study, of the 50 Arabic PTS questions, 18 questions (36%) were valid and 32 questions (64%) were invalid. Puspitaningsih et al stated that a good test is a test that obtains a high level of validity so that it can truly measure students' abilities (Puspitaningsih, Febrianto, & Putro, 2019, p. 118). A test with a good level of validity is a test that gets 80% of the questions declared valid (Mahmudi et al., 2023, p. 565). The Arabic PTS question items for MAS Al-Islam Petala Bumi, Indragiri Hulu have not yet reached the percentage of 80%, that is, they have only reached 36% or are not yet valid. This is because teachers do not understand the procedures for preparing learning outcomes instruments properly, especially in the use of effective sentences, alternative forms of answers, level of difficulty, distinguishing power, and so on.

Similar results were seen in Tanjung et al.'s research, where the majority of questions were categorized as invalid with a percentage of 70% while only 30% were valid (Tanjung et al., 2024). Likewise, in Al

Azmi's research, the questions analyzed had low validity with a percentage of 54% invalid and only 46% valid (Azmi, 2023). The majority of questions that are invalid in this study are because the calculated r-value produced is smaller than the r table of 0.273, such as those in numbers 4 and 5 which have calculated r values of 0.203 and 0.122. The questions that are categorized as valid in this research are those that have a calculated r value greater than the r table (0.273) as found in numbers 3 and 7 which have a calculated r value of 0.509 and 0.530.

For this reason, teachers must save valid questions in the question bank so they can be used again in the next exam. Meanwhile, invalid questions must be corrected before they are used so that the results of the exam provide actual results. In this case, teachers can use good tools or programs or ask professionals to correct the validity of the questions so that the questions used to evaluate student learning outcomes have high validity.

Reliability

In this study, the reliability of PTS Arabic language questions for class XI MAS Al-Islam Petala Bumi, Indragiri Hulu was 0.20 which was classified in the low reliability group. Zainal Arifin stated that a good test is a test that can be used repeatedly and provides stable and consistent results (Arifin, 2012, p. 326). The tests used in decision making must have at least a minimum reliability coefficient of 0.85 (A. Kurniawan et al., 2022, p. 160). This is because the majority of PTS Arabic questions for class XI are categorized as difficult with a percentage of 56%, while there are no easy ones. So with the time provided and questions that are too difficult, it is possible that students will answer carelessly and will produce a low reliability coefficient value.

The results of this research are different from previous research conducted Tanjung et al showed a high level of reliability of the test items with r calculated 0.625 > r table 0.396 (Tanjung et al., 2024, pp. 356–357). This is because the questions that were prepared in this research were applied or tested on 25 students, which means more than in this study which was only tested on 13 students. Apart from that, the length of the questions in this study was 40 questions, which shows that the number of questions was less than in this study which amounted to 50 questions.

Based on the paragraph above, the researchers can conclude that when preparing tests teachers must pay attention to the level of difficulty of the questions with an ideal proportion of easy and difficult questions. Because questions that have an ideal level of difficulty are in the form of a normal curve to produce an optimal score distribution and can increase test reliability (Arifin, 2012, p. 327). Apart from that, teachers must also pay attention to the length of the test that will be used to measure students' Arabic language skills. The longer the test tends to increase the reliability of the test, because the more samples measured, the more correct answers, and the lower the guess factor, so that the measurement results become more stable and consistent (Arifin, 2012, p. 327). Meanwhile, if the questions are long but the sample measured is small, it will result in low test reliability. For this reason, if a small number of PTS Arabic questions are given to students, then the length of the questions prepared will not be too long.

Difficulty Level

In this study, of the 50 PTS Arabic language questions for class Mahmudi stated that a good test is not the easiest or the most difficult, because if the test is too easy or too difficult, the respondent's abilities cannot be known. The reason is that exams that are too easy do not attract students' interest, and exams that are too difficult make students less active in assessing (Mahmudi, 2020, p.

140). The proportion of difficulty levels of questions should be distributed normally to achieve good learning achievement. As for proportions a good level of difficulty for questions is 25% easy questions: 5% medium questions: 25% difficult questions or 20%: 60%: 20% or 15%:70%: 15% (Arifin, 2012, p. 347). The PTS Arabic language questions for MAS Al-Islam Petala Bumi, Indragiri Hulu are not yet said to be a good test, because in preparing them they did not pay attention to the proportion of the good questions.

The results of this research are in line with previous research conducted by Tanjung et al showed that the level of difficulty of the questions was not good, where only one question was included in the difficult category and the rest were included in the easy category (Tanjung et al., 2024, pp. 356–357). Apart from that, it is also in line with research by Mahmudi, et al which stated that the question items were not good, with a percentage of 80% easy questions, 20% medium questions, and 0% difficult questions (Mahmudi et al., 2023, p. 568). This is because it does not match the list of good difficulty levels for multiple-choice questions which are in the ratio 25%:5%:25%.

Based on the paragraph above, the researchers can conclude that when preparing tests teachers must pay attention to the proportion of difficulty levels of the items from easy questions, medium questions, and difficult questions. Difficult questions in this research need to be revised and retried if the questions do not meet the criteria for discriminating power and distracting power such as item 2. However, if the difficult questions meet the criteria for discriminating power and distracting power such as item 7, then the question can be selected and accepted as an alternative to be stored in the question bank. So that Arabic PTS questions can show students' true learning achievements.

Differentiating Power

In this research, the differentiating power of PTS Arabic language questions for class, 8 questions (16%) were very bad, and no questions had good differential power. Fatimatuz Zahroh stated that the different power of the questions is used to find out which students have understood the subject and which students have not understood the subject and also to find out which students are smart and which students are weak (Zahrah, 2022, p. 207). If the discriminating power is more than 0.30, it means the instrument can be classified as good. The higher the item discrimination index, the more able the questions are to differentiate intelligent students from less intelligent students (A. Kurniawan et al., 2022, p. 49; Munip, 2017, p. 276). It can be concluded that PTS questions still cannot differentiate between smart and weak students because there are still many questions that have poor discrimination so they have to be discarded.

The results of this research are in line with previous research conducted by Tanjung et al. The majority of the items analyzed had poor discriminating power (Tanjung et al., 2024, p. 358). However, this is different from the results of Alareifi's research which had items with a good discriminating power of 45% and a very good 40% (Alareifi, 2023, p. 132). This is due to the large number of questions with non-functional distracting power. The discriminating power of questions is affected by any dysfunctional distractors which makes it very important to differentiate between high achieving and low achieving examinees.

Based on the results of the discriminating power analysis, the researcher suggests that items that already have good discriminating power should be included in the question bank and can be removed again in the next test because their quality is

sufficient. As for the items whose discriminating power is still low, they can be corrected and submitted again in the learning outcomes test and then analyzed again to see whether they have improved or not, or the question items are discarded or dropped and not issued again for the learning outcomes test. And especially for questions whose item discrimination index number is marked negative, it is best not to issue them again in the learning outcomes test, because such items are of very poor quality (smart tests answer more wrongly than tests that are not clever, in fact only a few answer wrong). Therefore PTS questions can differentiate students' true Arabic language abilities.

Distracting power

In this study, the distracting power of PTS Arabic language questions for class XI, 10 questions (20%) were poor, and 1 question (2%) was very poor. Distractor function analysis is only carried out on multiple-choice tests which are intended to determine participants' understanding, knowledge, and accuracy of the material being tested (Zahrah, 2022, p. 107). A distractor that is not chosen at all by the testee means that the distractor is bad, while a good distractor is a distractor that has a great appeal to be chosen by testees who have little or no mastery of the test material. A distractor can be said to be working well if it is chosen by at least 5% of test takers and is chosen more by the lower group (Munip, 2017, p. 281). From the results of the distractor analysis, it can be concluded that the multiple-choice questions in Arabic PTS already have fairly good distractors.

The results of this research are different from previous research conducted by Rezigalla et al., where in this study the difference between the very good and good categories was 37.3%, the fair category was 22%, and the poor category was 3.4% (Rezigalla et al., 2024, p. 3). Likewise, it is different from the results of Chauhan's research which had items with a good distracting power category of 61.50%, a fair category of 29.50%, and a poor category of 9% (Chauhan et al., 2023, p. 4). This is because of the problem. In both studies, the item difficulty index and item discrimination power were appropriate, resulting in a distractor that functioned well. Meanwhile, in this study, there were still many item difficulty indices that were not appropriate, and the differentiating power of the items so that the distracting power of the Arabic PTS questions was still categorized as sufficient.

Based on the discussion above to improve and maintain the quality of PTS Arabic multiple-choice questions, teachers must identify distracting forces that are not functioning. Once identified, the teacher needs to remove or replace with functioning distractors to maintain the number of functioning distractors between three and four.

CONCLUSION

Based on the findings and discussion above, researchers can conclude as follows: 1) Validity: 36% of questions are valid and 64% are invalid. 2) Reliability: reliability coefficient value 0.20 (low reliability). 3) Difficulty level: 56% difficult questions, 44% medium questions, and 0% easy questions. 4) Differentiation power: 20% of questions are very good, 0% of questions are good, 36% of questions are fair, 28% of questions are bad, and 16% of questions are very bad. 5) Distracting power: 6% of questions are very good, 28% of questions are good, 44% of questions are fair, 20% of questions are bad, and 2% of questions are very bad.

This discussion concludes that the mid-semester assessment questions (PTS) for Arabic language class for this reason, researchers recommend that Arabic teachers revise or replace questions that need to be corrected so that the mid-semester

assessment (PTS) questions used to measure Arabic learning outcomes can reflect students' true Arabic language abilities.

The researchers hope that this article can help readers analyze the questions in general and the questions to measure Arabic learning outcomes in particular. However, of course, the researchers also admit that this article is not free from all shortcomings, for this reason, more in-depth research is needed, and with all humility, the author accepts criticism, suggestions, comments, and opinions that will make this article a general study.

REFERENCES

Alareifi, R. M. (2023). Analysis of MCQs in Summative Exam in English: Difficulty Index, Discrimination Index and Relationship between them. *Journal of Educational and Human Sciences*, *20*, 124–135. https://doi.org/https://doi.org/10.33193/JEAHS.20.2023.325

Arifin, Z. (2011). *Evaluasi Pembelajaran Prinsip, Teknik, dan Prosedur*. Bandung: PT Remaja Rosdakarya.

Arifin, Z. (2012). *Evaluasi Pembelajaran*. Jakarta: Direktorat Jenderal Pendidikan Islam Kementerian Agama RI.

Arikunto, S. (2006). *Prosedur Penelitian Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta.

Azmi, F. Al. (2023). *Analisis Soal Sumatif Keterampilan Membaca Berdasarkan Teori ―Norman E Gronlund‖ di SMA An-Nuqayyah Madura*. Universitas Maulana Malik Ibrahim Malang.

Bhat, S. K., & Prasad, K. H. (2021). Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study. *Indian Journal of Ophthalmology*, *69*(2), 344. https://doi.org/10.4103/ijo.IJO

Brownlie, N., Burke, K., & Laan, L. van der. (2024). Quality indicators of effective teacher-created summative assessment. *Quality Assurance in Education*, *32*(1), 30–45. https://doi.org/10.1108/QAE-04-2023-0062

Chauhan, G. R., Chauhan, B. R., Vaza, J. V, & Chauhan, P. R. (2023). Relations of the Number of Functioning Distractors With the Item Difficulty Index and the Item Discrimination Power in the Multiple Choice Questions. *Cureus*, *15*(7). https://doi.org/10.7759/cureus.42492

DIKDAS, T. G. (2021). *Modul Belajar Mandiri Calon Guru Pegawai Pemerintah dengan Perjanjian Kerja (PPPK): Pedagogi*. Jakarta: Direktorat Jenderal Guru dan Tenaga Kependidikan Kementerian Pendidikan dan Kebudayaan.

Direktorat Pembinaan Sekolah Menegah Atas. (2017). *Panduan penilaian oleh pendidik dan satuan pendidikan Sekolah Menengah Atas*. Jakarta: Direktorat Jenderal Pendidikan Dasar dan Pendidikan Menengah.

Htoon, K. Z., & Aung, Y. P. (2024). Item analysis of multiple-choice questions in summative assessment for professional examination I of an outcome-based integrated MBBS curriculum. *International Journal of Research in Medical Sciences*, *12*(5), 1451–1456. https://doi.org/10.18203/2320-6012.ijrms20241226

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item Analysis of Multiple Choice Questions: A Quality Assurance Test for An Assessment Tool. *Medical Journal Armed Forces India*, *77*(1), S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.

007

Kurniawan, A., Febrianti, A. N., Hardianti, T., Ichsan, Desy, & Risan, R. (2022). Evaluasi Pembelajaran. In *Social Science Academic* (Vol. 1). Padang: PT. GLOBAL EKSEKUTIF TEKNOLOGI. https://doi.org/10.37680/ssa.v1i2.3582

Kurniawan, R. Y., Prakoso, A. F., Hakim, L., Dewi, R. M., & Widayanti, I. (2017). Pemberian Pelatihan Analisis Butir Soal Bagi Guru di Kabupaten Jombang : Efektif? *Jurnal Pemberdayaan Masyarakat Madani (JPMM)*, *1*(2), 179–193. https://doi.org/10.21009/jpmm.001.2.03

Mahmudi, I. (2020). *Evaluasi Pendidikan*. Sleman: Lintang Books.

Mahmudi, I., Nurwardah, A., Rochma, S. N., & Nurcholis, A. (2023). Item Analysis Of Arabic Language Examination. *IJAZ ARABI: Journal of Arabic Learning*, *6*(3), 563–573. https://doi.org/https://doi.org/10.18860/ijazarabi.v6i3.19821

Mahmudi, I., Yasin, A., & Ardiyanti. (2024). Pelatihan Penyusunan Tes Dan Rubrik Penilaian Bagi Guru Bahasa Arab Chongraksat Wittaya School Thailand. *Communnity Development Journal*, *5*(1), 2544–2548.

Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2015). Kelayakan Anbuso Sebagai Software Analisis Butir Soal Bagi Guru. *Jurnal Kependidikian*, *45*(2), 198–210.

Munip, A. (2017). *Penilaian Pembelajaran Bahasa Arab*. Yogyakarta: Fakultas Ilmu Tarbiyah dan Keguruan UIN Sunan Kalijaga Yogyakarta.

Nuraeni, Z., Simarmata, R. H., Sukmaningthias, N., & Sari, N. (2021). Pelatihan Software SPSS untuk Menghitung Validitas, Reliabilitas, dan Analisis Butir Soal bagi Mahasiswa Calon Guru di Palembang. *Jurnal Anugerah*, *3*(1), 15–23. https://doi.org/10.31629/anugerah.v3i1.3383

Puspitaningsih, F., Febrianto, R., & Putro, B. N. (2019). *Evaluasi Pembelajaran*. Trenggalek: Sembilan Mutiara Publishing.

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., … Adam, M. I. E. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, *24*(1), 1–7. https://doi.org/10.1186/s12909-024-05433-y

Santosa, S., & Badawi, J. A. (2022). Analisis Butir Soal Pilihan Ganda Tema Pertumbuhan dan Perkembangan Makhluk Hidup Kelas III Madrasah Ibtidaiyah. *Jurnal Basicedu*, *6*(2), 601–614. https://doi.org/https://doi.org/10.31004/basicedu.v6i2.2206

Setyawan, C. E., & Fathoni, M. (2017). Kompetensi Pedagogik Guru Bahasa Arab Dalam Merancang dan Melaksanakan Ealuasi Pembelajaran di Madrasah Aliyah Negeri (MAN) Yogyakarta. *At-Ta'dib*, *12*(1), 143. https://doi.org/10.21111/at-tadib.v12i1.865

Siregar, S. (2018). *Statistik Parametrik untuk Penelitian Kuantitatif*. Jakarta: Bumi Aksara.

Sugiyono. (2010). *Metode Penelitian Pendidikan: Pendekatan Kuantitatif, Kualitatif dan R&D*. Bandung: Alfabeta.

Tanjung, M. A. H. R., Fahmi, A. A., Rahmanita, F., Filzafati, I. H., &

Qomari, N. (2024). Analisis Butir Soal Penilaian Akhir Tahun Pelajaran Bahasa Arab Pelajaran Bahasa Arab Kelas VII MTs Al-Ma'arif Rakit Banjarnegara Jawa Tengah. *Mantiqu Tayr: Journal of Arabic Language*, *4*(1). https://doi.org/https://doi.org/10.25217/mantiqutayr.v4i1.4038

Wulan, E. R., & Rusdiana, A. (2014). *Evaluasi Pembelajaran Dengan Pedekatan Kurikulum 2013*. Bandung: Pustaka Setia.

Zahrah, F. (2022). *Evaluasi Pembelajaran SD/MI*. Kediri: Kreator Cerdas Indonesia.