algnatuna
Journal of Arabic Linguistics and Education
Vol. 9 No. 2, June 2024, 190-206
P-ISSN: 2477-5371; E-ISSN: 2503-2690

EVALUATING THE EFFECTIVENESS OF ARABIC LANGUAGE SUMMATIVE TESTS IN SENIOR HIGH SCHOOLS: A COMPREHENSIVE ANALYSIS

Sahrani
Institut Agama Islam Negeri Pontianak, Indonesia
sahrani176@gmail.com

Sri Herlina
Institut Agama Islam Negeri Pontianak, Indonesia
sriherlina779@gmail.com

Sumin
Institut Agama Islam Negeri Pontianak, Indonesia
amien.ptk@gmail.com

Hermansyah
Institut Agama Islam Negeri Pontianak, Indonesia
hermansyahii@yahoo.com

Abstract
The aims of this study is to describe the validity, reliability, difficulty level, differentiation power and effectiveness of the Examination of Arabic language Sumative Test Items for Class X Students of Al-Mumtaz IT Senior High School. This research uses a quantitative approach. The data obtained were analyzed using IBM SPSS statistics 23. The subjects of this study were class X students totaling 29 students. The results of the analysis of 30 multiple choice questions showed that from the validity aspect it was found that 67% were valid and 33% were invalid. From the reliability aspect, it was found that the reliability level of the items was high, namely the r coefficient of 0.864. From the aspect of difficulty level, it was found that about 63% of the items were easy, about 37% of the items were moderate. From the aspect of distinguishing power, it was found that about 60% of the items were good, about 27% of the items were not good, about 13% of the items were quite good. And from the aspect of eliminators it was found that about 56% were accepted, about 26% were rejected and about 17% were revised.

Keywords*: Test Item Analysis, Arabic language, Summative Test.*

Abstrak
Tujuan penelitian ini mendeskripsikan validitas, reliabilitas, tingkat kesukaran, daya beda dan efektifitas Pengecoh soal Soal Tes Sumatif Bahasa Arab Siswa Kelas X Sekolah Menengah Atas Islam Terpadu (SMAIT) Al-Mumtaz. Penelitian ini menggunakan pendekatan kuantitatif. Data yang diperoleh di analisis menggunakan IBM SPSS statistik 23. Subjek penelitian ini siswa kelas X berjumlah 29 siswa. Hasil analisis dari 30 butir soal pilihan ganda menunjukkan bahwa dari aspek validitas ditemukan 67% valid dan 33% tidak valid. Dari aspek reliabilitas ditemukan bahwa tingkat reliablitas butir soal termasuk tinggi yaitu *r* koefisien sebesar 0,864. Dari aspek

Tingkat kesukaran ditemukan bahwa sekitar 63% butir soal yang mudah, sekitar 37% butir soal yang sedang. Dari aspek daya beda ditemukan bahwa sekitar 60% butir soal yang baik, sekitar 27% butir soal yang kurang baik, sekitar 13% butir soal yang cukup baik. Dan dari aspek pengecoh ditemukan sekitar 56% diterima, sekitar 26% ditolak dan sekitar 17% direvisi.

Kata Kunci: *Analisis Butir Soal, Bahasa Arab, Tes Sumatif.*

INTRODUCTION

Education has an important role in creating individuals who have good quality and potential. The high quality of education reflects the successful implementation of formal education in a country. Through the education process, a person will gain the necessary knowledge, understanding and skills. Good quality education is closely related to the learning process carried out at school[1]. There are three main elements in the learning process, namely teaching objectives, learning experiences, and learning outcomes. In Law Number 14 of 2005 concerning Teachers and Lecturers, Chapter I Article 1 Paragraph 1 explains that a teacher is a professional educator who has the main responsibility in educating, teaching, guiding, directing, training, assessing, and evaluating students at various levels of education such as early childhood education, basic education, and secondary education. These roles are very important in achieving the expected educational goals[2]. So one of the roles of teachers in education is to evaluate student learning outcomes. Formally, in Permendikbud no. 66 of 2013 explained that Educational Assessment as a process of gathering and processing information to measure the achievement of student learning outcomes includes: authentic assessment, self-assessment, portfolio-based assessment, repeat, daily repetition, midterm replication, end semester test, competency level test, competency level quality test, national exam, and school / madrasah exam [3].

Evaluation is a very important part and stage that must be passed in the learning system by a teacher to evaluate the effectiveness of learning. According to Carl H. Witherington, Evaluation is a statement that something has or does not have value. A similar statement was also expressed by Wand and Brown, Evaluation refers to the action or process of determining the value of something. From the views of these two experts, it can be concluded that evaluation is the process of describing students and assessing them based on value and meaning. This definition emphasizes that

[1] Hery Susanto, Achi Rinaldi, and Novalia, "Analisis validitas reliabilitas tingkat kesukaran dan daya beda pada butir soal ujian akhir semester ganjil mata pelajaran Matematika kelas XII IPS di SMA Negeri 12 Bandar Lampung tahun ajaran 2014/2015," *Al-Jabar: Jurnal Pendidikan Matematika* 6, no. 2 (2015): 203–218.

[2] Evi Rizki Amelia, "Analisis Butir Soal Ulangan Akhir Semester Gasal Mata Pelajaran Administrasi Umum Kelas X Jurusan Otomatisasi Tata Kelola Perkantoran Smk Cut Nyak Dien Semarang Tahun Pelajaran 2019/2020," *Thesis.*, (Semarang: Unnes, 2020), 74, https://lib.unnes.ac.id/29614/1/7101413025.pdf.

[3] Raswan, N. Lalah Alawiyah, and Taufik Luthfi, "Model of Arabic Authentic Assessment Instruments: Speaking (Kalām) At Madrasah Aliyah," *Alsinatuna: Journal of Arabic Linguistics and Education* 8, no. 1 (2022): 51–64, https://doi.org/10.28918/alsinatuna.v8i1.1739.

evaluation is related to value and meaning[4]. Evaluation activities are activities that must exist in a learning activity. This is because evaluation is an assessment process used to determine the success of learning that has been carried out, as well as a reference in improving the next learning process [5].

In the era of increasingly widespread globalization, information is very important for all institutions, including educational institutions. Information technology acts as a tool or means of learning, while information systems act as the core in the learning process. Every education manager, including lecturers, is expected to keep up with technological developments in order to remain relevant to the times[6]. However, no matter how great the technology used as a tool for assessment, it is not meaningful if the assessment instrument developed does not meet the characteristics of a good assessment instrument (test), especially in terms of its validity. There are five characteristics that mark a good test, namely validity, reliability, differentiation, appropriate level of difficulty, and practicality[7].

In conducting evaluation activities, it is important to use a good tool or instrument that is suitable for the object being measured. According to Arikunto an evaluation tool is said to be good if it is able to evaluate something with results that are in accordance with the situation being evaluated[8]. Budi Susanto also stated the same thing that the question instrument being analyzed is considered good if it meets the criteria of validity, reliability, distinguishing power, difficulty level, and effectiveness of the checker. These criteria should be met in every instrument used in learning evaluation activities, including tests that are often used to measure learning outcomes[9].

Arabic learning evaluation is a method used to assess the extent of learning achievement that has been determined after students learn. The results of the evaluation aim to measure the level of learning success and mastery of Arabic on the material that has been taught. In addition, through this level of mastery of Arabic, information can also be obtained about the problems and difficulties experienced by students in the Arabic language learning process[10]. After carrying out the test, the teacher must analyze the items to evaluate the quality of the questions used as an assessment of the

[4] Ina Magdalena, *Dasar-Dasar Evaluasi Pembelajaran* (Sukabumi: CV Jejak, 2022), 65.

[5] Elok Rufaqoh et al., "Jacob's Analytical Assessment In The Evaluation Of Arabic Writing Skills At Madrasah Aliyah Negeri Program Keagamaan," *Alsinatuna Journal of Arabic Linguistics and Education* 9, no. 1 (2023): 29–40.

[6] Abdusima Nasution, Junaidi Arsyad, and Syamsul Kurniawan, "Menakar Kompetensi Dosen Pendidikan Agama Islam Di Era Society 5.0," *Darul 'Ilmi* 11, no. 1 (2024): 106–121.

[7] Muhammad Lukman Arifianto et al., *Evaluasi Dan Pengembangan Tes Interaktif Bahasa Arab* (Yogyakarta: Tinggak Media, 2021), 43.

[8] Suharsimi Arikunto, *Dasar-Dasar Evaluasi Pendidikan* (Jakarta: Bumi Aksara, 2018), 35.

[9] Budi Susanto, "Analisis Butir Soal Ulangan Tengah Semester Pada Mata Pelajaran Matematika Di SMP Negeri 2 Punggur," *Disertasi.*, (Lampung: IAIN Metro, 2021), 85.

[10] Moh. Matsna and Erta Mahyudin, *Pengembangan Evaluasi Dan Tes Bahasa Arab* (Jakarta:Serambi, 2012), 70.

development of student learning outcomes. This is because quality questions are questions that are able to provide accurate information about students' understanding of the material taught[11].

Item analysis is needed to test the quality of each item and collection of questions in various aspects. Item analysis can be done qualitatively or quantitatively. The main aims of item analysis is to obtain information about the characteristics of each item, either through item review or empirical analysis. The results of this analysis can be used to determine the quality of the questions and the quality of student learning from the analyzed test results[12]. Anastasil and Urbin stated that item analysis has many benefits, including: 1) assisting test users in evaluating the quality of the tests used; 2) relevant for the preparation of informal tests such as tests prepared by teachers for students in the classroom; 3) supporting effective item writing; 4) materially improving classroom tests; 5) increasing the validity and reliability of questions[13].

Generally, in educational institutions at all levels, whether elementary, junior high, high school, or college, both private and public, evaluations or tests must be given to students in the hope of knowing the extent of the students' success in absorbing the material taught. SMAIT Al-Mumtas also conducts an evaluation at the end of each lesson. However, based on the results of the analysis conducted by the researcher on the tests given by the teacher, it was found that there were several items that did not meet the evaluation standards, both from the aspects of validity, reliability, difficulty level, differentiation, and the effectiveness of the examiner. Therefore, researchers are interested in conducting research on "_Analysis the Arabic Language Summative Test Items For Class X Students of Al-Mumtaz IT Senior High School". Researchers hope that in the future teachers can make good items and can improve the quality of the items, so that the items can really measure students' abilities and can distinguish which students are eligible for promotion.

In the 2023/2024 school year, Ira Yoshita Cahyaningrum, Anies Fuady, and Sunismi have conducted similar research. The research was entitled "Analysis of Odd Semester Final Summative Problem Items for Grade VII Mathematics Subjects with the Help of Anates Software Application". The results showed that the reliability of the question had a number of 0.51 which showed an adequate level. There is a division of a superior group consisting of 8 students with high scores, and an asor group consisting of 8 students with low scores. The differentiating power in working on summative questions at the end of the odd semester is classified as good, but the overall quality of question exemptions is less effective. Most of the question exemptions have not functioned

[11] Mila Ningrum Masitoh, "Analisis Butir Soal Penilaian Tengah Semester Genap Mata Pelajaran Pendidikan Agama Islam Kelas V Di SDN 1 Bumi Harjo Tahun Ajaran 2021/2022," *Disertasi.*, (Kebumen: IAINU, 2022), 23.

[12] Elviana, "Analisis Butir Soal Evaluasi Pembelajaran Pendidikan Agama Islam Menggunakan Program Anates," *Jurnal MUDARRISUNA: Media Kajian Pendidikan Agama Islam* 10, no. 2 (2020): 209-224.

[13] Sridadi, Riky Dwihandaka, And Ariyo Bagiastomo, "The Analysis Of Test Item On Learning Outcomes Of Physiical Education Subject Of Class Viii Students," *Jurnal Pendidikan Jasmani Indonesia* 16, no. 1 (2020): 28–40.

properly, and there is 1 question that is not legible[14]. This study was conducted with the aim of identifying the quality of summative questions for Analysis the Arabic Language Summative Test Items for Class X Students of al-Mumtaz IT Senior High School. This research will evaluate important aspects such as validity, reliability, level of difficulty, differentiation, and effectiveness of the questionnaire.

METHOD

In this study, a quantitative approach was used as the method used. The data source used is the text of analyze the Arabic Language Summative test items for class X students of al-Mumtaz IT senior high school. Data collection was carried out through documentation consisting of question grids, question sheets, answer keys, and student answer sheets. The data that has been collected is then analyzed using IBM SPPS statistics 23. The analysis aims to describe the level of validity, reliability, difficulty level, distinguishing power, and effectiveness of the checkers from the data that has been collected[15].

Validity is a measure of the extent to which a test can measure what it actually wants to measure. Data or information is said to be valid if it is in accordance with the actual situation. Criterion validity focuses more on comparing the test results of the test being tested with the test results of standardized tests that have been considered valid.

Table 1. Validation Value Criteria

Interval	Interpretation
0,800-1,00	Very high
0,600-0,799	High
0,400-0,599	Moderately high
0,200-0,399	Low
0,000-0,199	Very low

Reliability is a measure that describes the extent to which a measuring instrument is reliable. Test reliability refers to the level of accuracy or constancy of the instrument in evaluating the object being assessed[16]. A reliable test is a test that is consistent, not hesitant or indecisive. To obtain a reliable test, the test must be tested more than once on the same subject in different time periods, so that the results obtained remain consistent and reliable.

[14] Ira Yoshita Cahyaningrum, Anies Fuady, dan Sunismi, "Analisis Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran Matematika Kelas VII Dengan Berbantuan Aplikasi Software Anates," *Mathema: Jurnal Pendidikan Matematika* 5, no. 2 (2023): 67-81.

[15] Sitti Mania et al., "Analisis Butir Soal Ujian Akhir Sekolah," *Al Asma : Journal of Islamic Education* 2, no. 2 (2020): 274- 278, https://doi.org/10.24252/asma.v2i2.16569.

[16] Susanto "Analisis Validitas Reabilitas Tingkat Kesukaran ," 215.

Table 2. Criteria for Reliability Value

Interval	Interpretation
0,90 - 1,00	Very high
0,70 - 0,89	High
0,50 - 0,69	Medium
0,30 - 0,49	Low
Less than 0,30	Veri Low

Difficulty analysis refers to the evaluation of test questions based on their level of difficulty, with the aim of identifying questions that fall into the easy, medium, and difficult categories. An ideal question is one that is neither too easy nor too difficult. The level of difficulty of the question is determined based on the student's ability to answer it, not based on the teacher's point of view as the question maker[17]. Thus, there is an opportunity for students to be able to answer correctly more on each question, whether it has a medium, difficult, or very difficult level of difficulty.

Table 3. Criteria for Level of Difficulty

Interval	Interpretation
0,00 – 0,30	Difficult
0,31 – 0,70	Medium
0,71 – 1,00	Easy

Differentiability analysis refers to the examination of test questions to evaluate their ability to differentiate between low and high performing students. Item discriminating power refers to the ability of a test question to distinguish between high and low ability students. One of the important requirements of a good test instrument is to have a good level of discrimination.

Table 4. Criteria for Discriminating Power

Distinguishing Power	Interval	Interpretation
0,00 - 0,20	0 -25,9%	Not good
0,21 – 0,40	26 – 50,99%	Fairly good
0,41 – 0,70	51 – 75,99%	Good
0,71 - 1,00	76 -100%	Veri good

Distractor effectiveness is the extent to which the wrong choice is able to confuse test takers who do not know the correct answer. The more test takers who choose the distractor, then the distractor can function properly. The criteria for measuring the effectiveness of distractors on each question can be seen in the following table 5.

[17] Santoso, “Analisis Butir Soal Ulangan Tengah Semester," 70.

Table 5. Exception Effectiveness Criteria

Interval	Interprtion
>200%	Not good
0% - 25%	Less good
26% - 50%	Fairly goog
51% -75%	Good
76% -125%	Very good

Furthermore, based on these considerations, the quality of the items that have been analyzed can be determined, including questions with good quality, good enough, less good and not good.

RESULT AND DISCUSSION

The data obtained by researchers are the results of answers that have been analyzed for the number of correct answers from all students of Arabic Language class X al-Mumtaz IT as follows.

Table 6. Results of Student Answer Data

No	Name	Answers obtained																												Total	Score
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30		
1	Aisyah	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	1	0	14	46,67
2	Alessandro	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	26	86,67
3	Annisa	0	1	1	1	0	0	0	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	0	1	20	66,67
4	Assyifa	1	1	1	1	1	1	0	0	0	1	0	1	0	0	1	0	0	0	1	1	1	1	1	1	1	0	1	1	1	0	19	63,33
5	Aulia	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1	1	1	1	19	63,33
6	Aziyati	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	26	86,67
7	Damar	1	1	1	1	0	0	1	0	1	0	0	1	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	14	46,67
8	Dihyan	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	25	83,33
9	Fadhil	0	1	1	1	0	0	0	1	0	0	1	1	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	19	63,33
10	Faizah	1	0	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	0	1	23	76,67
11	Fathia	0	1	1	0	0	1	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	9	30
12	Fathina	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	29	96,67
13	Hafidz	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	25	83,33
14	Melinda	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	28	93,33
15	M. Abiyyu	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	0	20	66,67
16	M. Afgan	1	0	1	1	1	1	0	1	0	1	1	1	0	1	1	0	0	1	1	0	1	1	1	1	1	1	1	0	1	1	22	73,33
17	M. Faris	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	28	93,33
18	Nadira	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30	100
19	Najla	1	1	1	0	0	1	0	1	1	0	0	1	0	1	0	1	1	0	1	0	1	1	0	0	0	1	0	0	0	1	15	50
20	Nur Syafiqah	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	26	86,67
21	Puan	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	0	1	0	0	1	0	0	0	20	66,67
22	Rafi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30	100
23	Rafifa	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	24	80
24	Rahmat	0	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	0	1	1	0	0	1	1	1	1	1	21	70
25	Rasya	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	28	93,33
26	Salwa	0	1	1	1	1	0	1	1	1	0	0	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1	0	1	1	1	21	70
27	Shin Shia	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	0	1	0	1	1	0	0	0	0	1	1	0	1	20	66,67
28	Syafiqa	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	26	86,67
29	Togi	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	26	86,67
	Jumlah	23	25	29	27	23	20	16	26	18	20	19	26	18	23	22	19	24	11	23	20	23	26	25	23	24	22	23	19	15	21		

The following research results are then presented and described through the results of the calculation of aspects of validity, reliability, difficulty level, differentiation and effectiveness of the Arabic language summative test items for class X students of Al-Mumtaz IT senior high school.

Validity Test

A test is said to be valid or valid, if the test really measures what it wants to measure, it must be able to measure the level of learning outcomes achieved in the implementation of a desired goal[18]. With 29 students and 30 items, it is known that the r table is 0.316 at the 0.05 significance level. From the results of the calculation, 10 invalid items were obtained, while 30 valid items. The following is a recap of valid and invalid question items on table 7 below.

Table 7. Validity of Question Items

No	Pearson Correlation		Sig. (2-tailed)	
1	0,348	Low	0,064	Not valid
2	-0,018	Very low	0,926	Not valid
3	.a	High	0	Valid
4	.555**	Fair	0,002	Valid
5	.612**	High	0	Valid
6	0,197	Very low	0,305	Not valid
7	.440*	Fair	0,017	Valid
8	.561**	Fair	0,002	Valid
9	0,313	Low	0,099	Not valid
10	.573**	Fair	0,001	Valid
11	.593**	Fair	0,001	Valid
12	0,188	Low	0,329	Not valid
13	.506**	Fair	0,005	Valid
14	0,2	Low	0,299	Not valid
15	.525**	Fair	0,003	Valid
16	.537**	Fair	0,003	Valid
17	0,223	Low	0,245	Not valid
18	.556**	Fair	0,002	Valid
19	.480**	Fair	0,008	Valid
20	.472**	Fair	0,01	Valid
21	.480**	Fair	0,008	Valid
22	.539**	Fair	0,003	Valid
23	.622**	High	0	Valid
24	.447*	Fair	0,015	Valid
25	.400*	Fair	0,032	Valid
26	0,307	High	0,106	Not valid
27	0,282	Fair	0,138	Not valid
28	.664**	High	0	Valid
29	0,298	Low	0,117	Not valid
30	.466*	Fair	0,011	Valid

[18] Eva Trifiani Damayanti, "Analisis Butir Soal Evaluasi Pembelajaran Pendidikan Agama Islam Kelas XI IPA & IPS DI SMAN 3 Probolinggo," *Disertasi.*, (Malang: UIN Maulana Malik Ibrahim, 2012), 90.

Based on the results of data analysis obtained using IBM SPSS statistics 23, the validity coefficient (r table) is 0.316 or a question item if the significance is <0.05, it can be said that the question item is valid. So after observing the table above, it is known that of the 30 items analyzed, 33% invalid questions were obtained and 67% valid items.

Reliability Test

Reliability test is very important to determine whether the question is reliable or not. This test should be done at least 2 times with the same question and different times. In this first reliability test, it was carried out only for items that were said to be valid. There are 20 items that are eligible for the reliability test with a significance level of 5%, the degree of freedom of the r table value = 0.561 and the question is said to be reliable if the Cronbach alpha value is> 0.6. The first reliability test obtained the following data.

Table 8. Reliability of Problem Items

Cronbach's Alpha	N of Items
.837	21

Based on table 8 above explains that the results of the reliability test on 20 question items obtained the Cronbach alpha result of 0,872. So it can be concluded that the value of 0.872 is greater than r table 0.561 and greater than 0.6, then the question items are reliable. Then it can also be interpreted that 0.872 is included in the range of 0.70-0.89 in the correlation index, it can be concluded that the level of reliability of the items is high.

Furthermore, the second validity test the researcher uses the split-half technique or tajzi'atu al-ikhtibar (split-half method). This technique is done by separating the scores into two groups, namely odd groups and even groups. What is meant by odd and even groups is the answer in the form of scores from odd number items and even number items. And the items are all tested. The following results are obtained from the variable test.

Table 9. Reliability 1

Cronbach's Alpha	Part 1	Value	.687
		N of Items	15[a]
	Part 2	Value	.768
		N of Items	15[b]
	Total N of Items		30
Correlation Between Forms			.650
Spearman-Brown Coefficient	Equal Length		.788
	Unequal Length		.788
Guttman Split-Half Coefficient			.781

The second reliability test table 9 above shows that the correlation coefficient from the calculation is 0.781. This only shows the level of test reliability for half of the questions. Meanwhile, what is expected is the level of reliability for all questions. To obtain the correlation coefficient of the reliability level of all questions, the Spearman-Brown formula is used as follows[19].

Reliability of the whole tes	=	2 x reliability of half the test 1 + half test reliability
$r = \frac{2 \times 0{,}781}{1 + 0{,}781}$	=	$\frac{1{,}522}{1{,}781}$
r	=	0,864

Based on the results of the analysis obtained r count of 0,864> r table (0.561), it means that the level of reliability of the test with the two-split technique is reliable. Similarly, when confirmed with the correlation index, it can be seen that the level of reliability of these items is high, because it is between 0,70 and 0,89.

Level of Item Difficulty

Analysis of the level of difficulty is very important to examine the items in terms of their difficulty, because the results can be known that there are items that are suitable to be given to students. This analysis aims to find out easy items, medium items and difficult items. The level of difficulty of the question can be seen from how many students can answer it, not seen from teachers who are experts in making questions[20]. The results of the analysis of the level of difficulty explained on the following table 10 below:

Table 10. Level of Problem Item Difficulty

Soal	N		Mean	Difficulty Level
	Valid	Missing		
S1	29	0	0,793	Easy
S2	29	0	0,862	Easy
S3	29	0	1	Easy
S4	29	0	0,931	Easy
S5	29	0	0,793	Easy

[19] Ina Magdalena et al., "Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas III SDN Karet 1 Sepatan," *BINTANG : Jurnal Pendidikan Dan Sains* 3, no. 2 (2021): 198–214, https://ejournal.stitpn.ac.id/index.php/bintang.

[20] Magdalena et al.

Soal	N		Mean	Difficulty Level
	Valid	Missing		
S6	29	0	0,69	Medium
S7	29	0	0,552	Medium
S8	29	0	0,897	Easy
S9	29	0	0,621	Medium
S10	29	0	0,69	Medium
S11	29	0	0,655	Medium
S12	29	0	0,897	Easy
S13	29	0	0,621	Medium
S14	29	0	0,793	Easy
S15	29	0	0,759	Easy
S16	29	0	0,655	Medium
S17	29	0	0,828	Easy
S18	29	0	0,379	Medium
S19	29	0	0,793	Medium
S20	29	0	0,69	Medium
S21	29	0	0,793	Easy
S22	29	0	0,897	Easy
S23	29	0	0,862	Easy
S24	29	0	0,793	Easy
S25	29	0	0,828	Easy
S26	29	0	0,759	Easy
S27	29	0	0,793	Easy
S28	29	0	0,655	Medium
S29	29	0	0,517	Medium
S30	29	0	0,724	Easy

The table shows that there are 63% of items that are categorized as easy and there are 37% of items that are categorized as moderate. From this information, it can be followed up that it is better for easy items to be re-examined so that it can be known why the question can be easily done by students, and easily students can find which answer key and which is just an exception.

Differentiating power

Daryanto said that the differentiating power of a question is the ability of a question to distinguish between high-ability students and low-ability students[21]. Good test items should be able to distinguish between capable test takers and incapable test takers, in other words, good test items

[21] Anida Rahmaini dan Aitya Nur Taufiq, "Analisis Butir Soal Pendidikan Agama Islam Di SMK N 1 Sedayu Tahun Ajaran 2017/2018 (Analisis Tingkat Kesukaran, Daya Pembeda Dan Fungsi Distraktor Pada Soal Pilihan Ganda Kelas XI)," *Jurnal Mudarrisuna* 8, no. 1 (2018): 1–24, https://www.jurnal.ar-raniry.ac.id/index.php/mudarrisuna/article/view/2787.

should be answered correctly by capable students and answered incorrectly by incapable test takers[22]. The table 11 is an analysis of the differentiation of test items that have been tested.

Table 11. Differentiated Power of Problem Items

Items	Scale Mean If Item Deleted	Scale Variance If Item Deleted	Corrected Item-Total Correlation	Distinguishing Power Index	Cronbach's Alpha If Item Deleted
S1	21,72	26,207	0,277	Fairly good	0,834
S2	21,66	27,734	-0,085	Not good	0,843
S3	21,52	27,544	0	Not good	0,837
S4	21,59	26,108	0,52	Good	0,829
S5	21,72	25,064	0,56	Good	0,824
S6	21,83	26,791	0,109	Not good	0,84
S7	21,97	25,463	0,357	Fairly good	0,831
S8	21,62	25,815	0,519	Good	0,828
S9	21,9	26,167	0,224	Fairly good	0,836
S10	21,83	24,933	0,508	Good	0,825
S11	21,86	24,766	0,528	Good	0,824
S12	21,62	27,03	0,13	Not good	0,837
S13	21,9	25,167	0,431	Good	0,828
S14	21,72	26,85	0,123	Not good	0,838
S15	21,76	25,333	0,461	Good	0,827
S16	21,86	25,052	0,466	Good	0,827
S17	21,69	26,793	0,152	Not good	0,837
S18	22,14	24,909	0,485	Good	0,826
S19	21,72	25,635	0,417	Good	0,829
S20	21,83	25,433	0,398	Good	0,829
S21	21,72	25,635	0,417	Good	0,829
S22	21,62	25,887	0,495	Good	0,828
S23	21,66	25,377	0,578	Good	0,825
S24	21,72	25,778	0,381	Good	0,83
S25	21,69	26,079	0,336	Good	0,832
S26	21,76	26,333	0,229	Good	0,835
S27	21,72	26,493	0,208	Not good	0,836
S28	21,86	24,409	0,607	Good	0,821
S29	22	26,214	0,206	Not good	0,837
S30	21,79	25,527	0,394	Fairly good	0,83

From the table 11, can be seen that there are 60% of the items that are categorized as good, there are 27% of the items that are categorized as less good, there are 13% of the items that are categorized as quite good. So it can be concluded that most of the The Arabic Language Summative

[22] Rusmayani, "Analisis Butir Soal Penilaian Akhir Semester Genap Mata Pelajaran Pendidikan Agama Islam Di SMP Bintang Persada Tabanan-Bali," *Widya Balina* 5, no. 1 (2020): 41–49, https://doi.org/10.53958/wb.v5i1.50.

Test Items For Class X Students Of Al-Mumtaz It Senior High School have good item quality because 60% are in the range between 51-75.99%.

Excerpts Effectiveness

Checkers on multiple choice possible answers are divided into two, namely the answer key and distractors. Of the many alternative answers only one is correct, namely the answer key and the possibility of an incorrect answer is called an exception. The aims of analyzing distractors is to find out how many students answer correctly according to the answer key and how many choose distractors or exceptions.

The following is a test of distractors or exceptions using IBM SPSS statistics 23 then manually clarified using excel and the results also show the same percentage, so for more clarity we can see from the following formulas and tables.

Formula:

$D = A / N \times 100\%$

D: Distractor/Excellent Rate

A: Number of students who chose the answer

N: Total number of students

Table 12. Percentage of Excerpts Effectiveness

Items	Item Statistics			Distractor index	answer key
	Mean	Std. Deviation	N		
S1	0,79	0,412	29	Very good	E
S2	0,86	0,351	29	Very good	C
S3	1	0	29	Very good	A
S4	0,93	0,258	29	Very good	C
S5	0,79	0,412	29	Very good	C
S6	0,69	0,471	29	Good	E
S7	0,55	0,506	29	Good	D
S8	0,9	0,31	29	Very good	A
S9	0,62	0,494	29	Good	A
S10	0,69	0,471	29	Good	C
S11	0,66	0,484	29	Good	A
S12	0,9	0,31	29	Very good	B
S13	0,62	0,494	29	Good	C
S14	0,79	0,412	29	Very good	D
S15	0,76	0,435	29	Very good	B
S16	0,66	0,484	29	Good	D

Items	Item Statistics			Distractor index	answer key
	Mean	Std. Deviation	N		
S17	0,83	0,384	29	Very good	C
S18	0,38	0,494	29	Fairly good	B
S19	0,79	0,412	29	Very good	E
S20	0,69	0,471	29	Good	B
S21	0,79	0,412	29	Very good	E
S22	0,9	0,31	29	Very good	B
S23	0,86	0,351	29	Very good	B
S24	0,79	0,412	29	Very good	C
S25	0,83	0,384	29	Very good	D
S26	0,76	0,435	29	Very good	B
S27	0,79	0,412	29	Very good	A
S28	0,66	0,484	29	Good	E
S29	0,52	0,509	29	Good	A
S30	0,72	0,455	29	Good	D

Based on the table 12 above, it can be explained that of the 30 multiple choice questions with the effectiveness of the triggers, 60% are in the good category, 37% are in the very good category and 3% are in the good enough category.

The follow-up after analyzing the effectiveness of the triggers, according to Arikunto, is to analyze the quality of the multiple choice. The method is to see the following criteria: a. Accepted because it is good, meaning that all the distractors on the question have been chosen by 5% of the students. b. Rejected because it is not good, meaning that the distractors are not chosen by students at all (0%). c. Rewritten (revised) because it is not good, meaning that the distractors have not performed their functions well (distractors are chosen by less than 5%)[23]. The following is an explanation of whether the effectiveness of an exception can be accepted, revised or rejected.

Table 13. Effectiveness of Eligibility Checkers

Items	A	B	C	D	E
1	retrieved	revised	rejected	retrieved	Retrieved
2	retrieved	retrieved	retrieved	rejected	Rejected
3	retrieved	rejected	rejected	rejected	Rejected
4	rejected	rejected	retrieved	rejected	Retrieved
5	diterima	revised	retrieved	revised	Retrieved
6	rejected	rejected	revised	rejected	Retrieved
7	retrieved	retrieved	retrieved	retrieved	retrieved
8	retrieved	rejected	revised	retrieved	rejected
9	retrieved	retrieved	retrieved	rejected	retrieved

[23] Rahmaini dan Taufiq, “Analisis Butir Soal Pendidikan Agama Islam," 20.

Items	A	B	C	D	E
10	direvisi	retrieved	retrieved	rejected	rejected
11	retrieved	retrieved	rejected	rejected	retrieved
12	retrieved	retrieved	rejected	rejected	rejected
13	retrieved	retrieved	retrieved	retrieved	revised
14	retrieved	retrieved	revised	retrieved	rejected
15	retrieved	retrieved	retrieved	retrieved	rejected
16	rejected	revised	retrieved	retrieved	revised
17	retrieved	rejected	retrieved	retrieved	rejected
18	retrieved	retrieved	retrieved	retrieved	rejected
19	retrieved	direvisi	rejected	rejected	retrieved
20	revised	retrieved	retrieved	retrieved	ditolak
21	revised	revised	revised	retrieved	retrieved
22	revised	diterima	rejected	rejected	retrieved
23	revised	diterima	revised	revised	revised
24	rejected	revised	retrieved	retrieved	retrieved
25	revised	revised	retrieved	retrieved	rejected
26	retrieved	retrieved	retrieved	retrieved	rejected
27	retrieved	retrieved	retrieved	rejected	rejected
28	retrieved	revised	retrieved	revised	retrieved
29	retrieved	retrieved	retrieved	rejected	retrieved
30	revised	retrieved	rejected	retrieved	retrieved

The table 13 shows that out of 30 items there are 150 choices. So it can be concluded that: 1) there are 56% of choices in the accepted category, meaning that all distractors have been >5% chosen by students. 2) 27% of the choices are in the rejected category, meaning that the distractors are not selected by students at all (0%). 3) 17% of the choices are in the revised category, means that the distractors have not performed their functions well and need to be revised so that students choose below 5%.

An exception is said to function well if it is selected by 5% of students. If an exemplar is chosen evenly, then the exemplar is very good. Making a good Examiner on a multiple choices test is difficult, because poor Examiners will result in low differentiating power, if one or two Examiners do not function low.

CONCLUSION

Arabic learning evaluation results can be used as a measuring tool to determine how far the learning that has been determined can be achieved after students learn. From the results of the Arabic language evaluation, it is intended to measure the success of learning, and the level of mastery of Arabic language that is sufficient for the material that has been taught. According to

Budi Susanto that the question instrument analyzed is considered good if it meets the criteria of validity, reliability, differentiating power, difficulty level, and effectiveness of the examiner.

The results of the research and discussion of the analyze the arabic language summative test items for class x students of al-mumtaz it senior high school, it can be concluded that the quality of the items is good. This is measured from the aspects of validity, reliability, level of difficulty, differentiation and effectiveness of checkers. The analyzed question items totaled 30 out of 29 students. The results of the analysis showed that from the validity aspect it was found that 67% were valid and 33% were invalid. From the reliability aspect, it was found that the reliability level of the question items was high, namely the r coefficient of 0.864. From the aspect of difficulty level, it was found that about 63% of the items were easy, about 37% of the items were moderate. From the aspect of distinguishing power it was found that about 60% of the items were good, about 27% of the items were not good, about 13% of the items were quite good. And from the aspect of eliminators it was found that about 56% were accepted, about 26% were rejected and about 17% were revised.

REFERENCES

Amelia, Evi Rizki. "Analisis Butir Soal Ulangan Akhir Semester Gasal Mata Pelajaran Administrasi Umum Kelas X Jurusan Otomatisasi Tata Kelola Perkantoran Smk Cut Nyak Dien Semarang Tahun Pelajaran 2019/2020." *Thesis.*, (Semarang: Unnes, 2020), 74. https://lib.unnes.ac.id/29614/1/7101413025.pdf.

Arifianto, Muhammad Lukman, et al. *Evaluasi Dan Pengembangan Tes Interaktif Bahasa Arab*. Yogyakarta: Tinggak Media, 2021.

Arikunto, Suharsimi. *Dasar-Dasar Evaluasi Pendidikan.* Jakarta: Bumi Aksara, 2018.

Cahyaningrum, Ira Yoshita, Anies Fuady, and Sunismi. "Analisis Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran Matematika Kelas VII Dengan Berbantuan Aplikasi Software Anates." *Matema: Jurnal Pendidikan Matematika.* 5, no. 67–81 (2023): 2023.

Damayanti, Eva Trifiani. "Analisis Butir Soal Evaluasi Pembelajaran Pendidikan Agama Islam Kelas XI IPA & IPS DI SMAN 3 Probolinggo." *Disertasi.*, Malang: UIN Maulana Malik Ibrahim, 2012.

Elviana. "Analisis Butir Soal Evaluasi Pembelajaran Pendidikan Agama Islam Menggunakan Program Anates." *Jurnal MUDARRISUNA: Media Kajian Pendidikan Agama Islam* 10, no. 2 (2020): 209-224.

Magdalena, Ina. *Dasar-Dasar Evaluasi Pembelajaran*. Sukabumi: CV. Jejak Publisher, 2022.

Magdalena, Ina, Septy Nurul Fauziah, Siti Nur Faziah, and Fika Sulaehatun Nupus. "Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas III SDN Karet 1 Sepatan." *BINTANG : Jurnal Pendidikan Dan Sains* 3, no. 2

(2021): 198–214. https://ejournal.stitpn.ac.id/index.php/bintang.

Mania, Sitti, Fitriani Fitriani, Ahmad Farham Majid, Nidya Nina Ichiana, and Andi Ika Prasasti Abrar. "Analisis Butir Soal Ujian Akhir Sekolah." *Al Asma : Journal of Islamic Education* 2, no. 2 (2020): 274. https://doi.org/10.24252/asma.v2i2.16569.

Masitoh, Mila Ningrum. "Analisis Butir Soal Penilaian Tengah Semester Genap Mata Pelajaran Pendidikan Agama Islam Kelas V Di SDN 1 Bumi Harjo Tahun Ajaran 2021/2022." *Disertasi.,* Kebumen: IAINU, 2022.

Matsna, Moh., and Erta Mahyudin. *Pengembangan Evaluasi Dan Tes Bahasa Arab*. Jakarta:Serambi, 2012.

Nasution, Abdusima, Junaidi Arsyad, and Syamsul Kurniawan. "Menakar Kompetensi Dosen Pendidikan Agama Islam Di Era Society 5.0." *Darul 'Ilmi* 11, no. 01 (2023): 106–121.

Rahmaini, Anida, and Aitya Nur Taufiq. "Analisis Butir Soal Pendidikan Agama Islam Di SMK N 1 Sedayu Tahun Ajaran 2017/2018 (Analisis Tingkat Kesukaran, Daya Pembeda Dan Fungsi Distraktor Pada Soal Pilihan Ganda Kelas XI)." *Jurnal Mudarrisuna* 8, no. 1 (2018): 1–24. https://www.jurnal.ar-raniry.ac.id/index.php/mudarrisuna/article/view/2787.

Raswan, N. Lalah Alawiyah, and Taufik Luthfi. "Model of Arabic Authentic Assessment Instruments: Speaking (Kalām) At Madrasah Aliyah." *Alsinatuna Journal of Arabic Linguistics and Education* 8, no. 1 (2022): 51–64. https://doi.org/10.28918/alsinatuna.v8i1.1739.

Rufaqoh, Elok, Sutaman Sutaman, Zakiyah Arifa, Nahla Ibrahim Eljack Ibrahim, and Hasyim Asy'ari. "Jacob's Analytical Assessment In The Evaluation Of Arabic Writing Skills At Madrasah Aliyah Negeri Program Keagamaan." *Alsinatuna Journal of Arabic Linguistics and Education* 9, no. 1 (2023): 29–4.

Rusmayani. "Analisis Butir Soal Penilaian Akhir Semester Genap Mata Pelajaran Pendidikan Agama Islam Di SMP Bintang Persada Tabanan-Bali." *Widya Balina* 5, no. 1 (2020): 41–49. https://doi.org/10.53958/wb.v5i1.50.

Sridadi, Riky Dwihandaka, and Ariyo Bagiastomo. "The Analysis of Test Item on Learning Outcomes of Physiical Education Subject of Class VIII Students." *Jurnal Pendidikan Jasmani Indonesia* 16, no. 1 (2020): 28–40.

Susanto, Budi. "Analisis Butir Soal Ulangan Tengah Semester Pada Mata Pelajaran Matematika Di SMP Negeri 2 Punggur." *Disertasi.,* Lampung: IAIN Metro, 2021.

Susanto, Hery, Achi Rinaldi, and Novalia. "Analisis validitas reliabilitas tingkat kesukaran dan daya beda pada butir soal ujian akhir semester ganjil mata pelajaran Matematika kelas XII IPS di SMA Negeri 12 Bandar Lampung tahun ajaran 2014/2015." *Al-Jabar: Jurnal Pendidikan Matematika* 6, no. 2 (2015): 203–218.

.