Test Item Quality Analysis of Final Mathematic Assessment in Islamic Elementary School

Arifatul Lutfina*
UIN Salatiga, Indonesia
Email: arifatullutfina32@gmail.com

Peni Susapti
UIN Salatiga, Indonesia
Email: peni.susapti@iainsalatiga.ac.id

Abdel Hakim Abdel Rahman
Suez Canal University, Egypt
Email: abdelhakim.abdelrahman_afro@suez.ed.eg

*Corresponding Author

Abstract
This study aimed to determine the test item quality of the final mathematic assessment for the 5th-grade students of MI Muhammadiyah Kalibanger. This quantitative descriptive study applied documentation, interviews, and observation to collect the data. The data were analyzed for validity, reliability, difficulty level, differentiating power, and function. The analysis result showed that the items were categorized as good. 60% of the analyzed questions were deemed valid, and 100% of the analyzed items were reliable, with a reliability value higher than 0.40. The difficulty level was also found to be in the good category, as 56% of the total questions have a moderate difficulty level. Most questions were found to have poor differentiating power, while the questions' function was found to be effective since 80% of the total multiple-choice questions can work properly.

Keywords: test item quality, mathematics, final exam

INTRODUCTION

Education serves as a nation's foundation and may significantly affect all life sectors (Berrueta-Clement, 1984; Karnani, 2007; Anheier & Hammack, 2010). It functions as one of the benchmarks to see a country's progress. Quality education is believed to prepare a generation that brings a bright future for a country. Therefore, a country needs to put a massive effort into improving its education quality. The improvement should encompass all aspects of education, including students, programs, facilities, infrastructures, and curriculum, among other pivotal aspects. Indonesian education has undergone many curricular changes to adjust the education process to the country's condition (Ilmi et.al., 2020; Nur & Madkur, 2014; Prihantoro, 2015).

The latest curriculum, the 2013 curriculum, primarily aims to prepare a generation with noble characters and the ability to develop their knowledge, attitudes, and skills (Mahfud, 2019; Ayuningsih, et.al., 2020; Komara, 2017). This goal is in line with the national education goals stated in Law 20/2003 on the National education system, stating that the Indonesian education system aims to develop students' religiosity, spirituality, attitudes, knowledge, and skills.

The law defines education as a conscious and planned effort to create a learning atmosphere and learning process so that students actively develop their potential to have religious, spiritual strength, self-control, personality, intelligence, noble character, and skills needed by themselves, society, nation, and state (Gaol & Sitepu, 2020; Ikhwan, et.al., 2020; Intania, & Sutama, 2020).

Indonesian education also aims to prepare students to contribute to society, the nation, and the state (Usman, et.al, 2013; Hashim & Langgulung, 2008; Fatimah & Aly, 2020). One of the subjects in Education is Mathematics, which mainly focuses on students' knowledge. Through mathematics, students are required to think critically, systematically, innovatively, and logically. Critical and logical thinking skills deal not only with the theoretical issue but also with real-life problems and solutions (Bailin, 1987; Ennis, 1993; Sternberg, 1986).

Teachers are also required to conduct evaluations to see their students' skill development. The evaluation result would be the basis for planning the next learning process and future improvement. Evaluation is an inseparable part of education (Liu & Chen, 2015; Smith, 1999; Haghani, et.al., 2008). It begins with an assessment. Assessment is carried out to collect data, obtain information about the student's learning achievement, and compare the information to the learning goal. It usually encompasses a number of learning aspects. In the 2013 curriculum, the assessment is conducted following Bloom's taxonomy. Bloom (Daryanto, 2012) states that the concept of education basically includes three domains namely the cognitive domain which is related to the learning progress of students includes three domains, namely the cognitive domain related to thinking or intellectual abilities, the affective domain related to the ability of attitudes, values, and feelings, and the psychomotor domain related to the skills aspect. The assessment focused more on students' cognitive domain.

Bloom's taxonomy divides cognitive assessment into lower-order thinking skills (LOTS) and higher-order thinking skills (HOTS) (Qasrawi &

BeniAbdelrahman, 2020; Jansen & Möller, 2022; Mispani, et.al., 2021). In this regard, teachers are expected to facilitate students' HOTS development through a supportive learning process. A previous study showed that a student-centered, HOTS-based assessment capable of triggering curiosity might stimulate students' HOTS (Susapti & Istiqlal, 2021). The concept of HOTS assessment is used to determine the extent to which students can foster critical thinking and think logically to solve problems faced in life. Students are expected to be able to analyze, evaluate and even create.

Assessment in education has been widely applied in formal schools ranging from elementary schools to secondary schools. The implementation of the expected assessment can be in the form of daily. Our initial observation in MI Muhammadiyah Kalibanger found that the final assessment was done merely by distributing the test to students without analyzing the quality of the test item, particularly in mathematics. Teachers appeared not to perform the item analysis because they prepared the items merely based on the available blueprint. The blueprint was considered sufficient to design test items with easy, moderate, or difficult levels, and item analysis was deemed unnecessary. The item quality analysis was deemed time-consuming because teachers should make the analysis manually.

Question analysis is a crucial step in determining an assessment instrument's quality. It is important for teachers to understand the quality of items used for the assessment to accurately measure students' critical and systematic thinking skills in Mathematics. The question quality analysis could also provide teachers valuable information to enhance their item quality. The analysis could be done by collecting data from students' answers.

The analysis result would allow teachers to understand the quality of their questions, whether they are in a good, moderate, or poor category.

Based on the earlier description, we found it necessary to see the quality of mathematic questions for the final assessment in MI Muhammadiyah Kalibanger. This study aimed to analyze the quality of final mathematics assessment questions for the 5th-grade students of MI Muhammadiyah Kalibanger for the 2021/2022 academic year. To this end, previous relevant works were also reviewed to determine the standpoint of this study. To our knowledge, studies with the same research object have not been reported.

METHODS

This field research was conducted in Muhammadiyah Islamic Elementary School, Kalibanger, Gemawang District, Temanggung Regency (Hereafter, MI Muhammadiyah Kalibanger). This quantitative descriptive study collected numerical data to produce new findings. The population of the study was students' answer sheets of the final mathematic assessment for classes 5A and 5B. The questions analyzed in this study were in multiple-choice and essay formats. In total, there were 15 multiple-choice and ten essay questions. A nonprobability sampling technique was applied. Thus not all members of the population have an equal chance to be selected as research samples. The research sample was the answer sheet of 5A and 5B students (n=41).

A research needs a method that can be used to collect the data needed. Data collection techniques that can be done include interviews,

questionnaires, observations, and combination of the three. The data collection techniques in this research consisted of documentation, interviews, and observation. The documentation method is used to obtain documents and archives in the form of question grids, questions, answer keys, student answer sheets. In addition, this method is also intended to obtain an overview such as geographical location, history of establishment, vision, mission, goals, and the number of educators and education personnel. The interview technique in this study was used to obtain information from the head of the madrasah and the Mathematics teacher related to final mathematic assessment for 5th-grade. The observation method is a reinforcement and complement to the data obtained from interviews and documentation methods.

The data were analyzed quantitatively following a number of aspects: validity, reliability, difficulty level, distinguishing power, and the distractor function. Each aspect was analyzed using SPSS and Microsoft Office Excel. The validity coefficient was used to determine the validity of the item. A question was deemed valid if the validity coefficient value was between 0.30 and 0.50 (Azwar, 2012). A higher coefficient indicates a higher validity.

The formulas used to determine reliability are K-R. 20, K-R.21, and Cronbach's Alpha Formula (r11). In this study, Cronbach's alpha was determined following a method explained by Arikunto (Zein & Darto, 2013, p. 83). A reliability coefficient of higher than 0.60 and less than 0.80 were categorized as high reliability. Meanwhile, the reliability coefficient of less than 0.40 and less than 0.40 were categorized as Low. The item difficulty was presented in the difficulty index. A good item typically has a moderate level

of difficulty, represented by the difficulty index between 0.30 and 0.70. An item is deemed easy if the difficulty index ranges from 0.70-1.00, while it is considered difficult if the difficulty index ranges from 0.00 to 0.30. The distractor index was used to see the function of the distractor. A good item typically has a distractor index of 51-75% or 126-150%. A distractor is considered good if it is chosen by more than 5% of the test takers.

RESULTS AND DISCUSSION

MI Muhammadiyah Kalibanger was a Compulsory Education Madrasah at the beginning of its establishment. Then in 1951, under the care of K.H. Bahrun and associates, the Compulsory Education Madrasah changed its name to Madrasah Muhammadiyah Kalibanger. MI Muhammadiyah Kalibanger is located in Kalibanger village, Gemawang District, Temanggung Regency

MI Muhammadiyah Kalibanger at first quantitatively consisted of only a few students, but gradually this madrasah was able to develop both quantitatively and qualitatively. In 2016, MI Muhammadiyah Kalibanger received an accreditation score of B. The accreditation obtained caused this madrasah to grow and be recognized by the community until in 2019, the total number of students was 199 children.

The vision of MI Muhammadiyah Kalibanger is graceful in character and excellent in achievement. As for the missions is a) fostering an appreciation of the teachings of Islam so that it becomes a guide in behavior, b) fostering a spirit of creativity, competition, and achievement to all madrasa residents, c) enforcing discipline, safety, cleanliness, beauty,

harmony, and shade, d) organizing education with effective and quality learning in achieving academic and non-academic achievements, e) organizing extracurricular coaching and life skills training to explore and develop the interests, talents of high potential students so that they can develop optimally, f) prepare quality and actual learning services by involving educators and education personnel in training, workshops, education in accordance with their fields (linear), g) organizing Islamic nuanced education by creating a religious environment in madrasah, h) fostering a culture of akhlakul karimah in all madrasah residents, and i) organizing education that provides skills, cadre, and leadership.

The objectives that MI Muhammadiyah Kalibanger wants to achieve are two, namely general and specific. Two objectives above cannot be separated from the main objectives of education as stated in the law.

The general objective of MI Muhammadiyah Kalibanger are a) able to actively carry out yaumiyah worship correctly and orderly, b) improve student achievement in academic and non-academic fields, c) be of good character, d) participants memorize juz 30 (Juz Amma) and selected surahs in the Qur'an, e) able to foster a culture of reading and writing for madrasah residents, f) making the school a fun and child-friendly learning park so that the school community is able to manage knowledge, g) can compete and not lose with students from other madrasah in the field of science, and h) personality, healthy lifestyle, and care for the environment.

The specific objective of MI Muhammadiyah Kalibanger are a) pursuing the fulfillment of vital facilities in supporting the creation of a literacy madrasah-oriented education system, b) realizing a learning climate

that integrates the use of learning resources and facilities in madrasah and outside madrasah, c) develop the curriculum in accordance with the demands of society, the environment, and reading culture, d) implement a competency-based education system, e) making extracurricular activities a means of making students more trained and accustomed to dealing with problems, both technical and organizational, and f) provide the widest possible opportunity for learners to develop their talents and interests.

The Validity of Final Mathematic test items

The analysis result showed that fifteen question (60%) were deemed valid, while 10 questions (40%) were deemed invalid. The question items checked for validity consisted of 10 multiple choice questions and 5 essay questions. Table 1 presents the validity analysis of each question.

The validity test was performed using IBM SPSS 25. The validity coefficient was then compared to the rtable with a significance level of 0.05. Since this study employed forty-one subjects and a 5% significance level, the rtable value was 0.316. An item is deemed valid if the rcount is higher than the rtable. The analysis result showed that fifteen question (60%) were deemed valid, while 10 questions (40%) were deemed invalid. Valid questions indicated that they were functional and were able to measure what was supposed to be measured, whereas invalid questions mean that they were not able to measure aspects that should be measured, which may be accounted for by various factors. To conclude, fifteen out of twenty-five questions (60%) were valid.

Table 1. Test Item validity

No.		Significance Level	r count	r table	Validity
Multiple choice	Essay				
1.		0.435	0.125	0.316	Invalid
2.		0.000	0.764	0.316	Valid
3.		0.002	0.472	0.316	Valid
4.		0.002	0.475	0.316	Valid
5.		0.008	0.408	0.316	Valid
6.		0.000	0.521	0.316	Valid
7.		0.021	0.359	0.316	Valid
8.		0.678	-0.067	0.316	Invalid
9.		0.000	0.536	0.316	Valid
10.		0.000	0.560	0.316	Valid
11.		0.385	0.139	0.316	Invalid
12.		0.011	0.395	0.316	Valid
13.		0.017	-0.372	0.316	Invalid
14.		0.044	0.317	0.316	Valid
15.		0.012	0.391	0.316	Valid
	1.	0.044	0.316	0.316	Valid
	2.	0.102	0.259	0.316	Invalid
	3.	0.165	0.221	0.316	Invalid
	4.	0.000	-0.693	0.316	Invalid
	5.	0.000	0.634	0.316	Valid
	6.	0.009	0.405	0.316	Valid
	7.		1	0.316	Valid
	8.	0.109	-0.254	0.316	Invalid
	9.	0.604	-0.083	0.316	Invalid
	10.	0.889	-0.22	0.316	Invalid

Table 2. Item Validity Distribution

No.	Coefficient Of Correlation	No. (Multiple choice)	No. (*Essay*)	Total	Percentage
1.	>0.316 (Valid)	2, 3, 4, 5, 6, 7, 9, 10, 12, 14, 15	1, 5, 6, 7	15	60%
2.	<0.316 (Invalid)	1, 8, 11, 13	2, 3, 4 ,8, 9. 10	10	40%

The Reliability of Final Mathematic assessment questions

The analysis result showed that all questions were reliable, with multiple choice and essay questions having a correlation coefficient of 0.496 and 0.443, respectively. Table 3 presents the reliability test result.

The reliability test was performed using Cronbach's alpha. The result was then compared to the rtable with a significance level of 0.05. A reliability coefficient of lower than 0.4 was considered low, that of 0.4- 0.60 was considered moderately reliable, and a reliability coefficient of higher than 0.70 was considered highly reliable (Surapranata, 2006).

The analysis result showed that all questions were reliable, with multiple choice 100% and essay questions having a correlation coefficient of 0.496 and 0.443, respectively.

Difficulty level of each question in final mathematic assessment

The analysis result showed that seven questions (28%) were categorized as difficult, fourteen questions (56) were moderately difficult, and four questions (16%) were considered easy. Table 4 presents the level of difficulty of each question.

Table 3. Reliability Test Result

No.		r count	r table	Category
Multiple choice	Essay			
1.		0.492	0.316	Sufficient
2.		0.492	0.316	Sufficient
3.		0.492	0.316	Sufficient
4.		0.492	0.316	Sufficient
5.		0.492	0.316	Sufficient
6.		0.492	0.316	Sufficient
7.		0.492	0.316	Sufficient
8.		0.492	0.316	Sufficient
9.		0.492	0.316	Sufficient
10.		0.492	0.316	Sufficient
11.		0.492	0.316	Sufficient
12.		0.492	0.316	Sufficient
13.		0.492	0.316	Sufficient
14.		0.492	0.316	Sufficient
15.		0.492	0.316	Sufficient
	1.	0.443	0.316	Sufficient
	2.	0.443	0.316	Sufficient
	3.	0.443	0.316	Sufficient
	4.	0.443	0.316	Sufficient
	5.	0.443	0.316	Sufficient
	6.	0.443	0.316	Sufficient
	7.	0.443	0.316	Sufficient
	8.	0.443	0.316	Sufficient
	9.	0.443	0.316	Sufficient
	10.	0.443	0.316	Sufficient

Table 4. Item Difficulty level

No.		Difficulty Level	Category
Multiple choice	Essay		
1.		0.41	Moderate
2.		0.54	Moderate
3.		0.88	Easy
4.		0.54	Moderate
5.		0.76	Easy
6.		0.66	Moderate
7.		0.61	Moderate
8.		0.02	Difficult
9.		0.51	Moderate
10.		0.44	Moderate
11.		0.73	Easy
12.		0.27	Difficult
13.		0.27	Difficult
14.		0.71	Easy
15.		0.66	Moderate
	1.	0.67	Moderate
	2.	0.46	Moderate
	3.	0.33	Moderate
	4.	0.28	Difficult
	5.	0.66	Moderate
	6.	0.50	Moderate
	7.	0.37	Moderate
	8.	0.28	Difficult
	9.	0.23	Difficult
	10.	0.23	Difficult

The items' difficulty level distribution is presented in Table 5.

Table 5. Item Difficulty Level Distribution

No.	Difficulty Index	Item number (Multiple Choice)	Item number (*Essay*)	Total	Percentage (%)
1.	0.00-0.30 (Difficult)	8, 12, 13	4, 8, 9, 10	7	28%
2.	0.30-0.70 (Moderate)	1, 2, 4, 6, 7, 9, 10, 15	1, 2, 3, 5, 6. 7	14	56%
3.	0.70-1.00 (Easy)	3, 5, 11, 14	-	4	16%

The multiple-choice item difficulty level analysis was made using SPSS 25, while the essay item difficulty level was analyzed manually using Microsoft Office Excel. Items that are not too difficult and too easy are considered good items. Items with a difficulty index between 0.00-0.30 were considered easy, while those between 0.30-0.70 was deemed moderate, and those between 0.00-0.30 was deemed difficult (Suharsimi, 2013). Items with moderate difficulty levels could be used again in the future test, while those falling into difficult or easy categories should be revised and improved before being used in the future test.

To conclude, seven test items (28%) were categorized as difficult, fourteen items (56%) were moderate, and four (16%) items were considered easy. In other words, the test items used in the final mathematic assessment have a good level of difficulty. The reliability test was performed using Cronbach's alpha. The result was then compared to the rtable with a significance level of 0.05. A reliability coefficient of lower than 0.4 was

considered low, that of 0.4- 0.60 was considered moderately reliable, and a reliability coefficient of higher than 0.70 was considered highly reliable (Surapranata, 2006).

The analysis result showed that all questions were reliable, with multiple choice 100% and essay questions having a correlation coefficient of 0.496 and 0.443, respectively.

Discriminating power of test items in final mathematic assessment

The analysis result also showed that, as displayed in Table 6, four questions (16%) had a very poor discriminating power, nine questions (36%) had a poor discriminating power, five questions (20%) had a fair discriminating power, six questions (24%) had a good discriminating power, and one question (4%) had very good discriminating power.

The discriminating power of multiple-choice items was performed using SPSS 25, while that of the essay was analyzed manually using Microsoft Office Excel. Items with a discriminating power index between 0.40-0.70 were considered to have good discriminating power, while those between 0.70 - 1.00 were considered to have very good discriminating power.

Items with good discriminating power could be retained for future tests, while those with poor discriminating power should be removed. The analysis result shows that four items (16%) have very poor discriminating power, nine items (36%) have a poor discriminating power, five items (20%) have fair discriminating power, six items (24%) have a good discriminating power, and one item (4%) has a very good discriminating power. To conclude, the test items of the final mathematic assessment have poor ability

to distinguish students with high and low ability. Table 7 presents the item distribution in terms of its discriminating power.

Table 6. Test item's discriminating power

No.		Discriminating Power index	Category
Multiple choice	Essay		
1.		0.125	Poor
2.		0.764	Very good
3.		0.472	Good
4.		0.475	Good
5.		0.408	Good
6.		0.521	Good
7.		0.359	Fair
8.		-0.067	Very poor
9.		0.536	Good
10.		0.560	Good
11.		0.139	Poor
12.		0.395	Fair
13.		-0.372	Very poor
14.		0.317	Fair
15.		0.391	Fair
	1.	0.07	Poor
	2.	0.25	Fair
	3.	0.02	Poor

Item's Distractor efficiency in final mathematic assessment

Multiple choice items in this study provided four choices, and only one of them was the correct answer, while the other served as the distractors. In other words, there were forty-five distractors used in the assessment.

The analysis result indicated that 32 distractors (71.1%) worked properly, whereas thirteen distractors did not perform as expected. The result also demonstrates that only five items (33.3%) had efficient distractors, while the rest, 10 items (66.6%), had poorly performing distractors. Table 8 presents the distractor efficiency analysis.

Table 7. Discriminating power distribution

Item number	Discriminating power index	Item number (Multiple Choice)	Item number (Essay)	Total	Percent. (%)
1.	Negative (Very poor)	8. 13	9. 10	4	16%
2.	0.00-0.20 Poor	1. 11	1, 3, 4, 5, 6, 7, 8	9	36%
3.	0.20-0.40 Fair	7, 12, 14, 15	2	5	20%
4.	0.40-0.70 (Good)	3, 4, 5, 6, 9. 10	-	6	24%
5.	0.70-1.00 Very good	2	-	1	4%

Table 8. Multiple choice Items' Distractor Efficiency

Item number	Answer key	Distractor efficiency		Category
		Good	Poor	
1.	B	A, C, D	-	Very Good
2.	A	B, C, D	-	Very Good
3.	B	C	A, D	Poor
4.	D	C	A, B	Poor
5.	C	A	B, D	Poor
6.	B	A, C	D	Good
7.	D	A, B, C	-	Very good
8.	D	B, C	A	Good
9.	A	B, C	D	Good
10.	C	A, B, D	-	Very good
11.	B	A, C	D	Good
12.	D	A, B, C	-	Very good
13.	C	B, D	A	Good
14.	B	A, D	C	Good
15.	B	C, D	A	Good

The following table 9 presents the distribution of distractor's efficiency of each item.

Table 9. Item distractor efficiency distribution

Item number	Distractor Efficiency Criteria	Item Number (Multiple Choice)	Total	Percentage (%)
1.	3 items Very good	1, 2, 7, 10, 12	5	33.3%
2.	2 items (Good)	6, 8, 9, 11, 13, 14. 15	7	46.6%
3.	1 item (Less Good)	3, 4, 5	3	20%

The distractor efficiency was analyzed manually using Microsoft Office Excel. An item is considered to have a very good distractor if it has three functioning distractors. Items with two functioning distractors are considered to have good distractor, while those with only one functioning distractor is considered poor. Lastly, items with no functioning distractors are considered very poor.

A distractor is considered good if it is chosen by at least 5% of test takers. In this study, there were 45 distractors, and thirty-two of them were considered to function properly, while the rest, 13 distractors (28.8%), were considered not to function properly. Out of fifteen items, five items (33.3%) had very good distractors, while seven questions (46.6%) had good distractors, and the rest, three questions (20%), had poor distractors. In other words, 80% of the multiple-choice items were categorized as effective because the distractor functions properly.

CONCLUSION

This study concluded that the items of the final mathematic assessment for the fifth-grade students were good, with 60% of them deemed to be valid, 100% of them considered reliable with a reliability coefficient of 0.40, and 56% of items with moderate difficulty. However, while the distractors (80%) in this assessment could work properly, this test exhibited poor discriminating power. Good items could be retained for a future test, while those considered difficult and easy with poor discriminating power should be revised before being used for future tests.

REFERENCES

Agustina, Nelly. (2019). *Analisis Soal Ujian Akhir Semester Ganjil Mata Pelajaran Pendidikan Agama Islam Ditinjau dari Tingkat kesukaran dan Daya Pembeda pada Kelas VIII di SMP Islam Terpadu Bina Insani Metro*. Skripsi: Fakultas Tarbiyah dan Ilmu Keguruan Institut Agama Islam Negeri Metro.

Anastasi, Anne and Urbina, Susana (1997). *Psichological Testing*. (Seventh Edition). New Jersey: Prentice Halltric.

Anheier, H. K., & Hammack, D. C. (Eds.). (2010). *American foundations: Roles and contributions*. Brookings Institution Press.

Arikunto, Suharsimi. (2013). *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.

Ayuningsih, W., Syafaruddin, S., & Amiruddin, M. S. (2020). *Implementation of Islamic Education Curriculum Development in Al-Ulum Islamic School Medan*. Budapest International Research and Critics in Linguistics and Education (BirLE) Journal, 3(2), 1033-1044.

Azwar, Saifuddin. (2012). *Reliabilitas dan Validitas*. Yogyakarta: Pustaka Pelajar.

Bailin, S. (1987). Critical and creative thinking. Informal logic, 9(1).

Berrueta-Clement, J. R. (1984). *Changed Lives: The Effects of the Perry Preschool Program on Youths through Age 19. Monographs of the High/Scope Educational Research Foundation, Number Eight*. Monograph Series,

High/Scope Foundation, 600 North River Street, Ypsilanti, MI 48197.

Depdikbud. (2014). *Permendikbud nomor 58 tahun 2014 tentang Kurikulum 2013 Sekolah Menengah Pertama/Madrasah Tsanawiyah*. (Online). Tersedia:http://staff.unila.ac.id/ngadimunhd/files/2012/03/Permen-58-ttg-Kurikulum- SMP.doc.

Ennis, R. H. (1993). Critical thinking assessment. Theory into practice, 32(3), 179-186.

Fatimah, M., & Aly, A. (2020). *Religious culture development in community school: a case study of boyolali middle school, central Java, Indonesia*. Humanities & Social Sciences Reviews, 8(2), 381-388.

Gaol, R. L., & Sitepu, A. (2020). The Influence of Used Good-Based Learning Media on the Value of Chracter Education and Student's Motivation to Study. Budapest International Research and Critics in Linguistics and Education (BirLE) Journal, 3(4), 1696-1703.

Hashim, C. N., & Langgulung, H. (2008). *Islamic religious curriculum in Muslim countries: The experiences of Indonesia and Malaysia*. Bulletin of Education & Research, 30(1), 1-19.

Ikhwan, A., Farid, M., Rohmad, A., & Syam, A. R. (2020, May). Revitalization of Islamic Education Teachers in the Development of Student Personality. In 1st Borobudur International Symposium on Humanities, Economics and Social Sciences (BIS-HESS 2019) (pp. 162-165). Atlantis Press.

Ilmi, Z., Darma, D. C., & Azis, M. (2020). *Independence in learning, education management, and industry 4.0: habitat Indonesia during COVID-19*. Journal of Anthropology of Sport and Physical Education, 4(4), 63-66.

Intania, E. V., & Sutama, S. (2020). The role of character education in learning during the COVID-19 pandemic. Jurnal Penelitian Ilmu Pendidikan, 13(2), 129-136.

Jansen, T., & Möller, J. (2022). Teacher judgments in school exams: Influences of students' lower-order-thinking skills on the assessment of students' higher-order-thinking skills. Teaching and Teacher Education, 111, 103616.

Karnani, A. (2007). *The mirage of marketing to the bottom of the pyramid: How the private sector can help alleviate poverty*. California management review, 49(4), 90-111.

Kurniawan, Tutut. (2015). *Analisis Butir Soal Ulangan Akhir Semester Gasal*

Mata Pelajaran IPS Sekolah Dasar. Journal of Elementary Education, (Online), (4) (1), (http://journal.unnes.ac.id/sju/index.php/jee).

Komara, E. (2017). *Curriculum and civic education teaching in Indonesia*. EDUCARE, 10(1).

Liu, S., & Chen, P. (2015). Research on fuzzy comprehensive evaluation in practice teaching assessment of computer majors. International Journal of Modern Education and Computer Science, 7(11), 12.

Mahfud, C. (2019). *Evaluation of islamic education curriculum policy in Indonesia. Premiere Educandum*: Jurnal Pendidikan Dasar Dan Pembelajaran, 9(1), 34-43.

Mu'arifah, Isnani. (2017). *Analisis Kualitas Butir Soal Pilihan Ganda Mata Pelajaran Biologi Kelas X dan XI pada Ulangan Akhir Semester Tahun Ajaran 2016/2017 di MAN Kota Palangka Raya*. Skripsi: Fakultas Tarbiyah dan Ilmu Keguruan Institut Agama Islam Palangka Raya.

Nur, M. R., & Madkur, A. (2014). *Teachers'voices On The 2013 Curriculum For English Instructional Activities*. IJEE (Indonesian Journal of English Education), 1(2), 119-134.

Nurul Agustin, Vivin. (2013). *Peningkatan Aktivitas dan Hasil Belajar Siswa melalui Model Problem Based Learning (PBL)*. Journal of Elementary Education, (Online), (2) (1), (http://journal.unnes.ac.id/sju/index.php/jee).

Nuryadi dan Nanang Khuzaini. (2016). *Evaluasi Hasil dan Proses Pembelajaran Matematika*. Yogyakarta: PT Leutika Nouvalitera.

Permendikbud nomor 36 tahun 2018 tentang Perubahan Atas Peraturan Menteri Pendidikan dan Kebudayaan Nomor 59 Tahun 2014 tentang Kurikulum2013 Sekolah Menengah Atas/Madrasah Aliyah.

Prihantoro, C. R. (2015). *The perspective of curriculum in Indonesia on environmental education*. International Journal of research studies in Education, 4(1), 77-83.

Qasrawi, R., & BeniAbdelrahman, A. (2020). The Higher and Lower-Order Thinking Skills (HOTS and LOTS) in Unlock English Textbooks (1st and 2nd Editions) Based on Bloom's Taxonomy: An Analysis Study. International Online Journal of Education and Teaching, 7(3), 744-758.

Safari. (2008). *Analisis Butir Soal: dengan manual, kalkulator, dan computer*. Jakarta: APSI Pusat.

Salmina, Mik dan Fadlillah Adyansyah. (2017). *Analisis Kualitas Soal Ujian Matematika Semester Genap Kelas XI SMA Inshafuddin Kota Banda Aceh*. Journal STKIP Bina Bangsa Getsempena, (Online), (4) (1).

Setiawan, Budiana dkk. (Eds). (2019). *Pengembangan Model Pembelajaran Matematika Berbasis Higher Order Thinking Skills (HOTS): Faktor- Faktor yang Mempengaruhi Keberhasilan Pembelajaran Matematika Berbasis HOTS*. Jakarta: Puslitjakdikbud.

Smith, S. R. (1999). AMEE guide no. 14: outcome-based education: part 2-planning, implementing and evaluating a competency-based curriculum. Medical teacher, 21(1), 15-22.

Sternberg, R. J. (1986). Critical Thinking: Its Nature, Measurement, and Improvement.

Sudijono, Anas. (2017). *Pengantar Evaluasi Pendidikan*. Jakarta: Raja Grafindo Persada.

Sugiyono. (2012). *Metode Penelitian Pendidikan (Pendekatan Kuantitatif, Kualitatif, dan R&D)*, Bandung: Alfabeta.

Surapranata, Sumarna. (2006). *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes Implementasi Kurikulum 2004*. Bandung: PT Remaja Rosdakarya.

Supranoto, Kusaeri. (2012). *Pengukuran dan Penilaian Pendidikan*. Yogyakarta: Graha Ilmu.

Suryani, Yulinda Erma. (2017). *Pemetaan Kualitas Empirik Soal Ujian Akhir Semester pada Mata Pelajaran Bahasa Indonesia SMA di Kabupaten Klaten*. Jurnal Penelitian dan Evaluasi Pendidikan, (Online), (21) (2), (http://journal.uny.ac.id/index.php/jpep)

Susapti, Peni dan Muhammad Istiqlal. (2021). *Pengembangan Higher Order Thinking Skills (HOTS) untuk Madrasah Ibtidaiyah*. Jurnal Pendidikan Guru MI: Al-Ibtida, (Online), Vol (8) (2), (http://dx.doi.org/10.24235/al.ibtida.snj.v8i2.88219).

Umar, J. (1991). *Pengantar Penilaian Pendidikan (makalah)*. Jakarta: Pusisjian, Baltibang, Depdikbud.

Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional. (2004). Jakarta: PT Armas Duta Jaya.

Usman, A. H., Shaharuddin, S. A., & Abidin, S. Z. (2017). *Humanism In Islamic Education: Indonesian References*. International Journal of Asia-Pacific Studies, 13(1).

Widiyanto, Joko (Ed). (2018). *Evaluasi Pembelajaran (Konsep, Prinsip dan Prosedur)*. Madiun: Unipma Press.

Zein, Mas'ud dan Darto. (2012). *Evaluasi Pembelajaran Matematika*. Daulat Riau.