Journal of Education and Islamic Contemporary Issue
Vol. 1 No. 1 March 2025
https://jeici.kips.or.id/index.php/jeici
DOI:

Anates: Analysis of Objective Test Item Quality

Ummul Aini[1]*, Nurrahmi Lathifa[2], Mimi Jelita[3]
[123]Universitas Islam Negeri Sjech M. Djamil Djambek Bukittinggi, Indonesia
E-mail: [1]ummul221999@gmail.com, [2]nurrahmilathifa23@gmail.com, [3]mimijelita259@gmail.com

ABSTRACT

The teacher's task is not only to compile and create learning outcome tests, but teachers must also pay attention to the quality of the questions that will be given to students starting from the level of validity, reliability, level of difficulty, distinguishing power, and effectiveness of distractors. However, in reality, teachers often only make test questions, and teachers forget that the quality of the questions that will be given to students must be known so that improvements or refinements can be made to each question item. The purpose of the study was to describe the quality of multiple-choice questions for the final semester exam for Islamic Religious Education and Character Education for grade VI seen from the level of validity, reliability, level of difficulty, distinguishing power, and effectiveness of distractors. The research method used quantitative descriptive, the population used all answer sheets of grade VI students, using the total sampling technique, so the sample amounted to one set of multiple-choice questions for the final semester exam. The data collection technique used documentation, and was processed using the Anates application program version 4.09. The results of the study showed that out of 20 multiple-choice questions, 11 questions had very good quality, 1 question had good quality, 7 questions had moderate quality, and 1 question had poor quality.

Keywords: *validity, reliability, level of difficulty, discriminatory power*

Article history

Received:	*Revised:*	*Accepted:*	*Published:*
25 April 2024	*7 Juli 2024*	*21 Oktober 2024*	*05 Maret 2025*

INTRODUCTION

Education is a conscious, systematic, ongoing effort to develop the potential that humans bring, instill character and provide skills by the goals of education, namely to enlighten the life of the nation and develop whole people, people who believe in and are devoted to God Almighty and have noble character, have knowledge and skills, physical and spiritual health, a solid and independent personality and a sense of social and national responsibility.

One component of education is teachers. Teachers have several competencies that can support their duties. These competencies are pedagogical competence, social competence, professional competence, and personality competence. In pedagogical competence, teachers are required to conduct evaluations. Evaluation plays an important role in determining the achievement of goals.

"Based on the Law of the Republic of Indonesia Number 20 of 2003 concerning the National Education System, it states that educational evaluation is an activity of controlling, guaranteeing, and determining the quality of education for various components of education at every path, level, and type of education as a form of accountability for the implementation of education."

Evaluation is an action carried out to determine the level of success of an educational, teaching, or training program that has been implemented. For this reason, a teacher's task is not only as an educator who teaches and guides in class but must also evaluate his students. In carrying out evaluation activities, of course, good quality information or data is needed. Such data can be obtained by conducting measurements and assessments. To measure the expected abilities or skills of students, a test is needed. A test is a systematic instrument or tool consisting of a set of questions or tasks to measure student learning outcomes. There are two written tests that teachers usually use in schools to test student learning outcomes, namely essay tests and objective tests (multiple choice questions).

The test itself must be good so that the evaluation process functions properly and according to its objectives. This is often forgotten by teachers in the field, they only stop at reporting the evaluation results without feeling the need to know how good the test they have used is. Therefore, teachers need to conduct item analysis. The activity of analyzing item questions is an activity that must be carried out by teachers to determine the quality of each item that has been written. From the results of analyzing the item questions, it can be used to make improvements or refinements to each item question. This activity is the process of collecting, summarizing, and using information from student answers to make decisions about each assessment. The goal is to review each item question to obtain quality questions before the questions are used. Quality questions are questions that can provide information as precisely as possible according to their objectives, including being able to determine which students have or have not mastered the material taught by the teacher. Questions are said to have good quality if they have high validity, reliability, and discrimination power, as well as a moderate level of difficulty, and no less importantly, the questions can measure the competencies that are expected to be achieved. One of the tests used in assessing student learning outcomes is the final semester exam (UAS).

The quality of the final semester exam (UAS) greatly influences the information obtained by teachers about their students' ability to master the material during one semester, because good quality questions will provide more accurate information to teachers (Muluki, Bundu, & Sukmawati, 2020). Questions are one of the instruments for conducting assessments, especially the final semester exam (UAS) assessment, so the questions must be of good quality so that the assessment results are truly measurable.

As a teacher, a teacher does not only create tests, but more than that, a teacher must also improve the quality of the tests he or she compiles. Teachers who have a lot of experience teaching and compiling test questions, also sometimes find it difficult to realize that their tests are still not perfect. If in a test that is conducted, almost all students get bad scores, it means that the test that was compiled was probably too difficult. Conversely, if all students get good scores, it can be interpreted that the test is too easy. Therefore, one of the best ways that a teacher must do is to analyze each item of questions used so that good, less good, and bad questions can be identified so that improvements can be made.

In general, teachers in schools often only make test questions as an assessment activity. However, teachers forget that the quality of the questions given to students must be known so that improvements or refinements can be made to each question item. Consideration of the criteria for good question items is expected to be able to provide accountable information. In other words, teachers are required to be able to prepare and carry out assessments well so that the learning objectives that have been set can be achieved optimally.

This problem is often found in schools, one of which is at SD Surya Kids Bukittinggi. Teachers at SD Surya Kids Bukittinggi rarely analyze exam questions because they do not understand how to analyze questions, so teachers do not know the quality of the questions being tested on students. Teachers' knowledge and skills in analyzing test items are still relatively low. If the quality of each test item is not known for certain, it will affect the tendency for errors in interpreting test results. This certainly has an impact on the bias of information obtained from assessment tools regarding students' actual abilities. Therefore, test items made by teachers need to be analyzed to determine their quality, so that it can be followed up whether the test items are suitable for reuse, need to be improved or should be replaced with new questions. Based on this problem, research on the analysis of test item quality, especially in the subjects of Islamic Religious Education and Character Education.

Based on these problems, this study aims to describe the quality of multiple-choice questions in the final exam of the odd semester for Islamic Religious Education and Character Education for Grade VI at SD Surya Kids Bukittinggi in terms of validity, reliability, level of difficulty, discriminating power, and effectiveness of distractors.

METHOD

This study uses a quantitative descriptive research type. The definition of quantitative descriptive is that the research is conducted quantitatively and only describes the actual conditions according to the object being studied (Arikunto, 2015). The population used is all answer sheets of students in the Odd Semester Final Exam for Islamic Religious Education and Character Education for grade VI at SD Surya Kids Bukittinggi in the 2023/2024 academic year with a total of 18 people.

The sample used in this study used the total sampling technique, namely all members of the population were used as samples for the Odd Semester Final Exam for the subject of Islamic Religious Education and Character Education for class VI at SD Surya Kids Bukittinggi for the 2023/2024 academic year, the sample of this study was one set of Odd Semester Final Exam questions consisting of 20 Multiple Choice questions.

The data collection technique used is documentation, which is related to research data such as student names, questions, and answers for the Final Semester Exam. The data analysis technique used is descriptive statistical analysis. And the data processing process using the Anates 4.09 application program is then concluded. Quantitative analysis includes analysis of validity, reliability, discrimination power, level of difficulty, and effectiveness of distractors.

The basis for making decisions regarding validity testing is:

1. If the calculated r value > r table, then the question item or statement in the questionnaire is significantly correlated with the total score (the questionnaire item is declared valid)
2. If the calculated r value < r table, then the question items or statements in the questionnaire do not correlate significantly with the total score (the questionnaire items are declared invalid) (Widiyanto, 2010)

Furthermore, the basis for decision making in reliability testing is as follows:

1. If the Cronbach's Alpha value > 0.60 then the instrument is declared reliable or consistent.
2. If the Cronbach's Alpha value is < 0.60 then the instrument is declared unreliable or inconsistent (Wiratna, 2014:193)

The basis for making decisions on differential power is:

Item Distinguishing Power	Information
0 - 0.20	The test items have weak discriminatory power.
0.21 – 0.40	The test items have moderate discriminating power.
0.41 – 0.70	The test items have good discriminating power.
0.71 – 1.00	The test items have very strong discriminatory power.
Negative marked	The test items have very poor discriminatory power.

Source: Arikunto, 2003: 213,218

The basis for decision making of the difficulty index is often classified as follows:

Difficulty Index	Information
0.00 to 0.30	Difficult
0.31 to 0.70	Currently
0.71 to 1.00	Easy

The basis for making decisions on distractor effectiveness:

a. If three distractor answers work, then the question is said to have very good distractor effectiveness.

b. If two distractor answers work, then the question is said to have good distractor effectiveness.

c. If there is only one distractor answer that works, then the question is said to have poor distractor effectiveness.

d. If all the distractor answers do not work, then the question is said to have bad distractors.

RESULTS AND DISCUSSION

1. Validity Test

Validity comes from the word validity, which means the extent to which a measuring instrument is accurate and precise in performing its function (Sudaryono, 2016). Validity is the ability of a measuring instrument to accurately measure the conditions to be measured. An instrument is said to have high validity if the instrument can provide results that are in line with what is to be measured. Testing the validity of multiple-choice questions for the final semester exam for Islamic Religious Education and Character Education for class VI of SD Surya Kids Bukittinggi in the 2023/2024 Academic Year using the biserial correlation coefficient formula with the help of the Anates version 4.09 application program. The results of the calculation of the validity of the questions will then be consulted with the r table at a significance level of 5%. As the number of students who took the final semester exam for the Islamic Religious Education and Character Education for class VI of SD Surya Kids Bukittinggi in the 2023/2024 Academic Year was 18 students. So with a level of 5% and a sample of 18 students, the r table shows a value of 0.455, with the condition that the validity test decision is if r table < r count then the question item is declared valid. Conversely, if r table > r count then the question item is declared invalid.

Based on the results of the analysis of multiple choice questions using the Anates version 4 application program, the overall validity of the questions or XY

Correlation is 0.82 when compared to rtable, then the multiple choice questions as a whole can be said to be valid. Meanwhile, if the validity of the questions is analyzed, there are 12 questions or 60% declared valid on numbers 3, 4, 5, 6, 13, 14, 15, 16, 17, 18, 19 and 20 with the criteria rcount> 0.455, while 8 questions with a percentage of 40% are declared invalid on numbers 1, 2, 7, 8, 9, 10, 11 and 12 with the criteria rcount <0.455.

Table 1. Validity Analysis of Multiple Choice Questions

No	Validity	Question Item Number	Amount	Percentage
1.	> 0.455 (Valid)	3, 4, 5, 6, 13, 14, 15, 16, 17, 18, 19 and 20	12	60%
2.	< o.455 (Invalid)	1, 2, 7, 8, 9, 10, 11, and 12	8	40%
Amount		**20**	**20**	**100%**

Based on these results, the multiple-choice questions in the final exam of the odd semester for the subject of Islamic Religious Education and Character Education for grade VI of SD Surya Kids Bukittinggi in the 2023/2024 Academic Year, there were 12 questions or 60% declared valid and 8 questions or 40% declared invalid, the results of the study were mostly able to measure what was to be measured.

The results of this study are in line with the theory put forward by Ngalim Purwanto which states that a test is said to have a very good level of item validity if the test can provide results by what is to be measured (Purwanto, 2013). Meanwhile, for 8 questions with a percentage of 40% that are invalid, it can be caused by various factors. This is by what Ground stated in Arifin's book, that three factors affect the validity of test results, namely the instrument factor used for the test, the administration and scoring factors, and the factor of student answers (Arifin, 2014).

Based on the results of the analysis of the validity of the final exam questions for this odd semester, it can be influenced by the instrument or questions used, because the questions have never been analyzed before, so teachers should conduct a validity analysis of the instruments that will be used to identify questions that are included in the invalid category and replaced with new questions by paying more attention to the rules for compiling good questions, especially for multiple choice questions so that they are worthy of being tested to be able to measure the level of student ability or to measure what is to be measured.

2. Reliability

Reliability comes from the translation of the word reliability, which has the origin of the words rely and ability. When combined, the two words will lead to an understanding of the ability of measuring instruments to be trusted and to be a basis for decision making (Purwanto, 2012). Reliability is also a value that shows the consistency of a measuring instrument in measuring the same symptoms. An instrument is said to be reliable if the instrument can produce consistent research data, because with consistency, data can be trusted to be true.

Reliability testing on multiple-choice questions and essay questions was carried out using the Anates application program version 4.09 from a series of comprehensive assessment instruments. The results obtained will then be interpreted with the reliability test criteria, namely if the test results are greater than 0.60, then the question can be said to have high reliability. Based on the results of the analysis of multiple-choice questions, a reliability of 0.90 was obtained, which means that the question has a high level of reliability.

Table 2. Analysis of Question Item Reliability

No	Type of Questions	Number of Question Items	Reliability	Interpretation
1.	Multiple choice	20	0.90	High Reliability

A test instrument that has good validity on each item will also have a high level of reliability. Similarly, Arikunto's opinion states that a test consisting of many items will be more valid than a test consisting of only a few items. The high or low level of validity can indicate the high or low level of reliability coefficient (Arikunto, 2015).

3. Difficulty Level

The level of difficulty is the ratio between the number of students who answer the questions correctly and the number of students who take the test. The more students who answer correctly, the lower the level of difficulty of the questions. Good questions have a moderate level of difficulty, meaning they are not too easy and not too difficult. The results will be interpreted into three criteria, namely: questions with a difficulty index (P) of 0.00 to 0.30 are classified as difficult questions, questions with a difficulty index (P) of 0.31 to 0.70 are classified as moderate questions, and questions with a difficulty index (P) of 0.71 to 1.00 are classified as easy questions. Based on the analysis of the level of difficulty of multiple-choice questions for the final semester exam for Islamic Religious

Education and Character Education for grade VI of SD Surya Kids Bukittinggi using the Anates version 4.09 program, the following results were obtained:

Table 3. Analysis of the Level of Difficulty of Multiple Choice Questions

No	Difficulty Index	Question Items	Amount	Percentage
1.	0.00 - 0.30 (Difficult)	-	-	-
2.	0.31 - 0.70 (Moderate)	16	1	5 %
3.	0.71 - 1.00 (Easy)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19 and 20	19	95%

Based on the results of the calculation of the level of difficulty of the multiple-choice questions for the final exam of the odd semester for the subject of Islamic Religious Education and Character Education for class VI of SD Surya Kids Bukittinggi in the 2023/2024 Academic Year, 19 questions or 95% were classified as easy questions, 1 question or 5% were classified as medium questions, but there were no or no questions that were classified as difficult. This is different from what was stated by Sudijono, that questions can be said to have a good level of difficulty if the questions are not too difficult and not too easy, in other words, moderate or sufficient. Meanwhile, the results of this study showed that there were more easy questions than moderate questions.

For example, multiple choice question number 1 has an easy level of difficulty (88.89) because out of 18 students, only 2 students answered the question incorrectly, this shows that question number 1 in multiple choice is very easy to answer by students. While in multiple choice question number 16 has a medium level of difficulty (66.67) because out of 18 students, 6 students answered the question incorrectly, this shows that question number 16 in multiple choice is moderate or difficult to answer by students.

Based on this, teachers should re-examine the questions that have an easy difficulty index, namely questions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19 and 20 to find out the cause of these questions being easy for students to work on. Easy questions come from material that is very easy for students to understand, and have distractors that do not function, and do not stimulate students to think in solving a problem. Therefore, teachers should formulate questions, especially multiple choice questions, clearly in the formulation of questions and answer choices, the main questions do not provide clues to the correct choice and the answer choices are homogeneous. Teachers also need to use questions with operational verbs at cognitive levels C2 (understand) and C3 (apply), by Arikunto's statement that the cognitive domains that are suitable for application at the elementary school/MI level are knowledge, understanding, and application. Easy questions are discarded and not issued again in the next exam. Meanwhile, questions that have a good level of difficulty (categorized as moderate) are stored in a question bank so that they can be issued again in the next exam.

4. Distinguishing Power

According to Solichin in the journal of Muhamad Rishan and Sulaiman, discriminatory power is the ability of a test item to distinguish between students with high and low abilities. The discriminatory power test was conducted using the Anates application program version 4.09. The number of subjects in this study was 18 students, so it was included in the small group. Calculating the discriminatory power for small groups was first divided into upper and lower groups. The Anates application program version 4.09 only took 27% of the superior group (upper group) and 27% of the asor group (lower group). So out of 18 students, there were 5 students in the upper group and 5 students in the lower group. The results of the calculation of discriminatory power are interpreted into five criteria, namely: 0.00 to 0.20 = poor, 0.21 to 0.40 = sufficient (statistical), 0.41 to 0.70 = good, 0.71 to 1.00 = excellent and negative = all are not good. The following are the results of the analysis carried out with the Anates application program version 4.09 on the discriminatory power aspect of multiple-choice questions for the final exam of the odd semester for the subject of Islamic Religious Education and Character Education for class VI SD Surya Kids Bukittinggi:

Table 4. Analysis of Differentiating Power of Multiple Choice Questions

No	Distinguishing Power	No. Question Item	Amount	Percentage
1.	0.00 – 0.20 Bad (poor)	5, 7, 8, 9, 10, 11, 13, 14 and 20	9	45%
2.	0.21 – 0.40 Enough (statistical)	1, 2, 12, 15 and 17	5	25%
3.	0.41 – 0.70 Good	3, 4, 6 and 19	4	20%
4.	0.71 – 1.00 Very good (excellent)	16 and 18	2	10%

5.	Negative, everything is not good.	-	-	-
	Amount	**20**	**20**	**100%**

Based on the calculation results, 9 questions or 45% have poor discriminatory power, 5 questions or 25% have sufficient discriminatory power (statistical), 4 questions or 20% have good discriminatory power, 2 questions or 10% have excellent discriminatory power, and no questions have negative discriminatory power.

Based on the description above, it is known that the questions that have poor and sufficient discriminating power are more than the questions that have good discriminating power, namely 9 that have poor discriminating power and 5 that have sufficient discriminating power, while the good discriminating power is 4 questions. This shows that these questions can distinguish the ability levels of high-ability and low-ability students. In line with Mania's statement, that good questions are questions that can distinguish between groups of high-ability and low-ability students. According to Sudijono, questions with good and very good discriminating power can be used and entered into the question bank, questions with sufficient discriminating power can be reused if the questions have been revised or improved and stored in the question bank, while questions with poor and very poor discriminating power should not be used again and replaced with new questions.

Questions with a very good category, namely questions number 16 and 18, can be used and entered into the question bank. Questions with a sufficient category, namely questions 1, 2, 12, 15, and 17, can still distinguish between students who understand the material and students who do not understand the material. These questions can cause students who do not understand the material to accidentally guess the correct answer. These questions can be reused if the questions have been revised and stored in the question bank. Meanwhile, questions in the bad category, namely questions number 5, 7, 8, 9, 10, 11, 13, 14, and 20 are questions that cannot distinguish between students who have mastered and those who have not mastered the material. These questions should not be used again and replaced with new questions.

Bad category questions can be caused by many students who do not master the material, but can answer the questions given. This is because the distractors do not function, allowing students who do not master the material to find the answer key. For example, multiple-choice question number 5 has poor discriminatory power because all the distractors are bad. So it is better to pay attention to the answer choices made when making questions, especially multiple-choice questions. The answer choices must come from the same material as that contained in the main question so that the answer key is not easily found.

5. Effectiveness of the Deception

Each multiple choice question or Multiple Choice alternative answers or options are called distractors. To conclude the effectiveness of the functioning of distractors on each question item, we can use the criteria adapted from the Likert scale as follows:

e. If three distractor answers work, then the question is said to have very good distractor effectiveness.
f. If two distractor answers work, then the question is said to have good distractor effectiveness.
g. If there is only one distractor answer that works, then the question is said to have poor distractor effectiveness.
h. If all the distractor answers do not work, then the question is said to have bad distractors.

From the results of the analysis of the quality of distractors in the final exam of the odd semester of Islamic Religious Education and Character Education for grade VI at SD Surya Kids Bukittinggi in the 2023/2024 Academic Year using the Anates application program version 4.09 with the results of 8 out of 20 multiple-choice questions that were compiled contained effective distractors, while 12 questions contained ineffective distractors. The following is a summary of the results of the analysis of the quality of the questions as follows:

Table 5. Results of the Analysis of Question Item Distractors

No	Effectiveness of the Deception	Question Number	Amount	Percentage
1.	Very good	3 and 6	2	10%
2.	Good	1, 2, 4, 12, 15, 16, 17, 18, 19 and 20	10	50%
3.	Not good	5, 8, 9, 10, 11, 13 and 14	7	35%
4.	Bad	7	1	5%
	Amount		**20**	**100%**

Based on the table, questions number 3 and 6 are questions with very good distractor effectiveness, because all the distractors function well and are chosen by at least 5% of all students, these questions can be stored in the question bank. Questions number 1, 2, 4, 12, 15, 16, 17, 18, 19 and 20 have distractor effectiveness in the good category, because in the question there is only one distractor that does not function well because it is chosen by less than 5% of all students who take the exam, these questions can be stored in the question bank on the condition that the distractor that does not function needs to be revised. Questions number 5, 8, 9, 10, 11, 13, and 14 have distractor effectiveness in the less good category, because two distractors do not function well because they are chosen by less than 5% of all students, these questions must be revised until they meet the criteria for good questions. Meanwhile, question number 7 is categorized as not good because three distractors do not function properly, because they were chosen by less than 5% of all students who took the exam, the question does not have all its distractors functioning properly, so the question must be discarded and replaced with a new question.

Based on the research results, it can be seen that the question items with poor distractor effectiveness are fewer than the question items with good distractor effectiveness. Seeing the existence of poor distractors is by Widiyanto's statement, that if the distractor contained in the question item is not selling, meaning that no one out of the many testees is interested in choosing the distractor, then this implies that the distractor is not functioning properly (Widiyanto, 2018). Seeing this, teachers should provide homogeneous and logical answer choices by the material contained in the main question when making questions, especially multiple choice questions. Then place the answer key randomly. And try to ensure that each question item is not interdependent or connected to other questions, so that students are not fooled into choosing the answer choice.

6. Follow-up of Question Item Analysis

After previously conducting item analysis based on validity, reliability, level of difficulty, discriminatory power and distractor effectiveness, it is necessary to follow up whether the items that have been created need to be revised or discarded by first knowing the overall quality of the items. The purpose of item analysis is to help improve the test through revision or discarding ineffective questions. With the follow-up, the quality of the items can be corrected as the next step in achieving improvement after this item analysis is complete. It can be analyzed using the Ikert scale which is grouped into 5 categories as follows (Oktanin & Sukirno, 2015):

No	Number of criteria met	Quality of Question Items	Revision	Enter Question Bank
1.	11	Very good	No need	Yes
2.	1	Good	Revision	Not yet
3.	7	Currently	Revision	Not yet
4.	1	Not good	Thrown Away	No
5.	-	Very Bad	Thrown Away	No

The results obtained show that there are 11 questions with very good quality, 1 question with good quality, 7 questions with moderate quality and 1 question with poor quality. Based on the description above, it can be concluded that the multiple-choice questions for the final semester exam for Islamic Religious Education for grade VI at SD Surya Kids Bukittinggi in the 2023/2024 Academic Year have good question quality because the questions with very good quality and good quality are more dominant or more than the questions with moderate, poor and very poor quality.

CONCLUSION

The results of the analysis of multiple-choice questions for the final semester exam for Islamic Religious Education and Character Education for grade VI are as follows:

1. Validity

   Based on the results of the analysis of multiple choice questions using the Anates version 4 application program, the overall validity of the questions or XY Correlation is 0.82 when compared to rtable, then the multiple choice questions as a whole can be said to be valid. Meanwhile, if the validity of the questions is analyzed, there are 12 questions or 60% declared valid on numbers 3, 4, 5, 6, 13, 14, 15, 16, 17, 18, 19 and 20 with the criteria rcount> 0.455, while 8 questions with a percentage of 40% are declared invalid on numbers 1, 2, 7, 8, 9, 10, 11 and 12 with the criteria rcount <0.455.
2. Reliability

   Based on the results of the analysis of multiple-choice questions, a reliability of 0.90 was obtained, which means that the questions have a high level of reliability.
3. Difficulty Index

   Based on the results of the analysis, it was found that 19 questions or 95% were classified as easy questions, 1 question item or 5% were classified as medium questions, but there were no or no questions classified as difficult.
4. Distinguishing Power

   Based on the calculation results, 9 questions or 45% have poor discriminatory power, 5 questions or 25% have sufficient discriminatory power (statistical), 4 questions or 20% have good discriminatory power, 2 questions or

10% have excellent discriminatory power, and no questions have negative discriminatory power.

5. Effectiveness of the Deception

   Based on the results of the analysis of questions 3 and 6, they are questions with very good distractor effectiveness, because all the distractors function well and are chosen by at least 5% of all students, these questions can be stored in the question bank. Questions 1, 2, 4, 12, 15, 16, 17, 18, 19 and 20 have distractor effectiveness in the good category, because in the question there is only one distractor that does not function well because it is chosen by less than 5% of all students who take the exam, these questions can be stored in the question bank on the condition that the distractor that does not function needs to be revised. Questions 5, 8, 9, 10, 11, 13, and 14 have distractor effectiveness in the less good category, because two distractors do not function well because they are chosen by less than 5% of all students, these questions must be revised until they meet the criteria for good questions. Meanwhile, question number 7 is categorized as not good because three distractors do not function properly, because they were chosen by less than 5% of all students who took the exam, the question does not have all its distractors functioning properly, so the question must be discarded and replaced with a new question.

6. Follow-up of Question Item Analysis

   The results obtained show that there are 11 questions with very good quality, 1 question with good quality, 7 questions with moderate quality and 1 question with poor quality. Based on the description above, it can be concluded that the multiple-choice questions for the final semester exam for Islamic Religious Education for grade VI at SD Surya Kids Bukittinggi in the 2023/2024 Academic Year have good question quality because the questions with very good quality and good quality are more dominant or more than the questions with moderate, poor and very poor quality.

REFERENCES

Azwar, S. (2012). Reliability and validity. Student Library.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. ERIC.

Fitrianawati, M. (2017). The Role of Question Item Analysis to Improve Question Item Quality, Teacher Competence and Student Learning Outcomes. Proceedings of the National Seminar and Call for Papers on Education 2017 (PGSD UMS & HDPGSDI Java Region).http://publikasiilmiah.ums.ac.id/handle/11617/9117

Hambleton, R. K., Shavelson, R. J., Webb, N. M., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Vol. 2). Sage.

Hikamudin, E., & Hairun, Y. (2021). Analysis of Apparent Score Disparity and Pure Score Estimation with Normative Reference Categorization on Student Learning Outcome Tests. DeltaPi: Journal of Mathematics and Mathematics Education, 10(1).

Mardapi, D. (2008). Techniques for compiling test and non-test instruments. Mitra Cendikia Press.

Mardapi, D. (2012). Measurement of educational assessment and evaluation. Nuha Medika.

Prijowuntato, SW, Mardapi, D., & Budiyono, B. (2015). Comparison of Standard Setting Measurement Error Estimates in SMK Accounting Competency Assessment. Journal of Educational Research and Evaluation, 19(2), 176–188.https://doi.org/10.21831/pep.v19i2.5578

Purnanto, AW, & Mahardika, A. (2017). Interactive question creation training with Wondershare quiz creator program for elementary school teachers in Magelang City. Warta LPM, 19(2), 141–148. Retnawati, H. (2014). Item response theory and its application: For researchers, measurement and testing practitioners, postgraduate students. Nuha Medika.

Rofiah, E., Aminah, NS, & Ekawati, EY (2013). Compilation of high-level thinking ability test instruments for physics in junior high school students. Journal of Physics Education, 1(2).

Ruslan, R. (2017). Estimation of Standard Error of Measurement of USBN Chemistry Try Out Questions for Senior High Schools in Makassar City. Makassar State University.

Suseno, I. (2017). Comparison of characteristics of multiple-choice test items reviewed from classical test theory. Factor: Scientific Journal of Education, 4(1), 1–8.

Van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of item response theory. Taylor & Francis Group.

Widayati, CSW (2009). Comparison of several methods of measuring error estimation. Journal of Educational Research and Evaluation, 13(2).

Zellatifanny, CM, & Mudjiyanto, B. (2018). Descriptive research types in communication science. Diakom: Journal of Media and Communication, 1(2), 83–90.