Quality Analysis of Physics Questions in Class 10 High School with Iteman 4.0: Case Study

Rosidah[1], Risky Setiawan[2], Zafrullah[3], Annisa Fitriani[4] and Ersa Mayola[5]

[1234]Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[5]Madrasah Aliyah Pivate Madani, Riau Islands, Indonesia
rosipradanarh2124@gmail.com

Abstract. One way to measure learning success is through assessment activities, which require analysis as an important step to ensure the quality of the evaluation tools used. This analysis is needed to ensure that the evaluation tool is effective in measuring learning achievement and assists in assessing the success of the educational process. This research aims to examine the quality of the Final Semester Assessment questions in Physics at one of the Madrasah Aliyah schools in Tanjung Pinang, Riau Islands Province, Indonesia in the form of item analysis. The data used in this research are end-of-semester assessment questions, answer keys, and student responses obtained through document analysis techniques. Data were analyzed using Iteman 4.0 to determine the level of difficulty of the questions, the discriminating power of the questions, and the effectiveness of distractors. Based on the research findings, several important findings were identified regarding the level of difficulty of the questions, discrimination of the questions, and the effectiveness of distractors. It was found that 50% of the questions had a very easy level of difficulty, while the other 50% had a medium level of difficulty. However, it is worth noting that the majority of questions had poor item discrimination (55%), with only 10% of questions having moderate item discrimination and 35% having good item discrimination. Meanwhile, the effectiveness of distractors is also a concern, where 55% of distractors are considered ineffective, 43% are considered quite effective, and only 2% are considered effective. Based on these findings, it is recommended to make improvements to questions with a very easy level of difficulty and consider improving item discrimination and distractors to improve the quality of the evaluation tool.

Keywords: Analysis, Question Quality, Classical Test Theory, Physics.

1 Introduction

Education is one of the most important components in life. Education has an important role in forming individuals who have potential, make positive contributions to society, and are able to face future challenges with confidence and courage [1]. One of the supporting components of education is evaluation, in fact evaluation is one of the important things in it. Evaluation needs to be carried out thoroughly at every level of education [2]. Education requires evaluation as a means or activity to control, guarantee and determine the quality of education

P. C. Kuswandi et al. (eds.), *Proceedings of the 6th International Conference on Current Issues in Education (ICCIE) 2023*, Advances in Social Science, Education and Humanities Research 847,
https://doi.org/10.2991/978-2-38476-245-3_22

Evaluation is carried out with the aim of finding out the extent to which educational goals have been achieved [3]. Learning activities cannot be separated from evaluation [4]. In the world of education, assessment is often considered the same as evaluation. Evaluation activities are regulated in Law of the Republic of Indonesia Number 20 of 2003 concerning the National Education System Chapter. Thus, learning outcomes are used to evaluate the results of increasing student achievement of understanding, skills and attitudes which are then used to improve learning plans, learning processes and ways of evaluating learning outcomes. Evaluation can be used by teachers to determine learning outcomes.

The forms of evaluation that are often used can be tests and non-tests [5]. Tests are one of the measuring tools that teachers often use to measure student learning outcomes. By using tests teachers can find out to what extent the goals that have been set can be achieved from the results of tests that have been carried out by students. A test can be said to be good as a measuring instrument and must meet several requirements, namely validity, reliability, objectivity, practicality and economics. The validity of a test means that a test can be said to be valid if the test can measure something that will be measured accurately. The reliability of a test means that a test is said to be reliable if the test results show consistency. The objectivity of a test means that in carrying out the test there are no subjective factors that influence it. The practicality of a test means that the test is easy to carry out, easy to check and is equipped with clear instructions. The economy of a test means that carrying out the test does not require expensive costs, a lot of energy and a long time [6]. Therefore, it is necessary to have a test that will be used to find out the extent to which a test can measure what it will measure.

The form of the test can also vary, for example in the form of multiple choices and essays or descriptions. In practice, teachers often use essay questions during daily tests and multiple choice questions during mid-semester or final semester assessments. This is done to make the exam more efficient. The questions used in school assessments also require validation or analysis to state whether the questions are good or bad. One way is to analyze the questions that have been applied. The quality of the questions that have been applied can be determined using item analysis. [7] as well as improving the quality of the results of Haladyna and Rodriguez (2021) [8]. Apart from that, it can also be used as a comparison to determine the selected question items [9].

Things analyzed include the level of difficulty, differentiating power, and effectiveness of choices or distractors. The form of questions analyzed is multiple choice questions. The results of the analysis will determine whether the question is included in the good category or not based on the fulfillment of the criteria for level of difficulty, distinguishing power, and effectiveness of choices using quantitative analysis. This is confirmed by [10] who stated that the use of multiple choice questions also requires an understanding of their characteristics.

Item analysis is important as a means of evaluating the use of multiple choice questions. The results of the analysis can be used as an evaluation tool to improve the questions to make them better. Improvement questions will be used more effectively in determining the quality of learning that has been implemented, such as research conducted [11]. On the other hand, the analysis carried out will provide specific information about the item. The findings of this special information and analysis are

one of the efforts to improve learning and increase the quality of education so that educators are not careless in carrying out evaluations because they try to measure it using several categories. In analyzing it, we still prioritize the specified standards [12].

The research conducted was also based on relevant previous research findings, such as research [13]–[18] who conducted item analysis research with various types of questions consisting of several different subjects. There is also research that focuses on developing questions or developing supporting components for evaluation, such as research conducted by [9], [19]–[23]. Research is the basis for improving the quality of the questions and testing them well. Several studies carried out reconstruction and validation of questions such as research [12], [24], [25]. Apart from that, there is item analysis research that uses applications or Excel in the analysis process, such as research [26], [27]. From several of these research topics, real differences can be seen from the research we conducted even though we both carried out evaluation research.

Item analysis includes quantitative item analysis and qualitative item analysis. In this article we will review qualitative and quantitative item analysis. The results of the analysis include: question difficulty, question distinguishing power, answer distribution statistics, test reliability, measurement error (standard error), and score distribution and score acquisition for each test taker. Based on the description above, the problem formulation in this research is what the characteristics of the item are. The aim of this research is to determine the characteristics of the items. The category and type of research is quantitative descriptive research. Quantitative descriptive research is intended only to describe, explain or summarize various conditions, situations, phenomena, or various research variables according to actual events that can be photographed, interviewed, observed, and can be revealed through documentary materials. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance with electronic requirements facilitating simultaneous or later production of electronic products, and (3) stylistic conformity across a conference proceeding.

2 Method

The type of research carried out is quantitative descriptive research which focuses on data analysis using numerical or statistical calculation methods [28]. This research was conducted on a limited basis at one junior high school in the Riau Islands in 2023. The subjects in this research were 33 grade 10 students at one of the Madrasah Aliyah in Tanjung Pinang, Riau Province, Indonesia. The questions used are Final Semester Assessment (PAS) questions in the Physics subject. Data collection was carried out by conducting interviews with teachers at the elementary school. To determine the level of difficulty, distinguishing power, effectiveness of distractors and reliability, researchers used the Iteman 4.0 application.

The criteria for the level of difficulty of the questions can be seen in Table 1 below:

Table 1. Classification of question difficulty levels.

Difficulty Level	Information

0.00-0.30	Difficult
0.31-0.70	Currently
0.71-1.00	Easy

Source: Istiyono (2018) [29]

The differentiation or discrimination index for question items can be seen in Table 2 below.

Table 2. Discrimination power criteria.

Criteria	Information
$D \leq 0.199$	Bad (rejected)
$0.200 - 0.299$	Good enough (need revision)
$0.300 - 0.399$	Medium (Not necessary revise)
$D \geq 0.400$	Very good

Source: Istiyono (2018) [29]

The criteria for distractor effectiveness can be seen in the proportional endorser in Iteman 4.0. Meanwhile, to see the reliability, you can use Table 3 below.

Table 3. Reliability criteria.

NO.	Criteria	Information
1	0 - 0.20	Very Low
2	0.21 - 0.40	Low
3	0.41 - 0.60	Simply
4	0.61 - 0.80	Tall
5	0.81 – 1.00	Very high

Source: (Arikunto, 2021) [6]

3 Results and Discussion

3.1 Results

The aim of this research is to look at the quality of multiple choice questions in the Final Semester Assessment (PAS) in Physics subjects in high school. By analyzing using Iteman 4.0, the analysis results can be seen in Figure 1 below.

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
Scored Items	40	28.364	4.457	15	37	0.70	0.210

Fig. 1. Overall “summary statistics” results scalar

From Figure 1 it can be interpreted that the average Final School Assessment shows a number of 28.364, the Standard Norm is 4.457, the smallest value obtained is that students get the correct answer of 15, and the largest value obtained is that students get the correct answer of 37. From the Mean value P shows the number 0.70 which indicates that the level of difficulty of the Final Semester Assessment questions is in the "Medium" category. Mean Rpbis shows the discriminating power of all questions, with a Mean Rpbis value of 0.210, this shows that the Discriminating Power value has a "Pretty Good" picture.

The results of the analysis of the level of difficulty per question item can be seen in Table 4 below.

Table 4. Results of difficulty level analysis per question item.

Category	Grain	Total
Easy	1, 2, 3, 4, 6, 7, 8, 9, 15, 17, 23, 26, 29, 30, 31, 33, 34, 35, 36, 40	20
Currently	5, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 24, 25, 27, 28, 32, 37, 38, 39	20
Hard	-	-

Source: Modification Researcher

Questions that fall into the "Easy" and "Medium" categories are considered balanced because an even distribution of difficulty levels will help ensure the test or exam is fair and able to measure different levels of understanding or ability. With 20 questions covering both categories, teachers or test administrators can test basic understanding as well as more challenging problem-solving skills. This can provide a more comprehensive picture of participants' knowledge and skills, as well as minimizing the risk of bias that may arise from focusing only on one particular level of difficulty.

It is important for teachers to add some questions in the "Difficult" category because this will give participants the opportunity to test their abilities in facing higher challenges. With varying levels of exam difficulty, teachers can better identify participants who have deeper understanding and higher problem-solving abilities. Apart from that, difficult questions can also motivate participants to continue learning and improving their skills because they will feel like they are being tested thoroughly. This will also help prepare them for more complex challenges in the future, such as final exams or real-world situations where they will need to apply their knowledge and skills at a higher level of difficulty. Thus, adding questions in the "Difficult" category will provide a more complete and accurate picture of the participant's ability to master the material being taught.

The results of the differentiation analysis per item can be seen in Table 5 below.

Table 5. Results of differentiating power analysis per item.

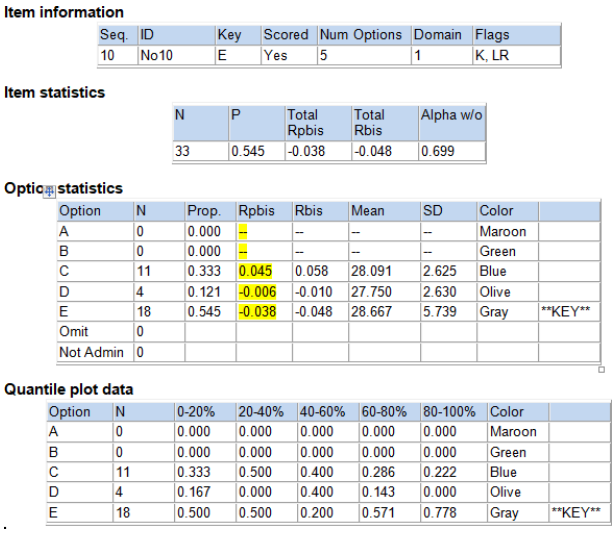
Category	Grain	Total
No Good	1, 4, 10, 11, 12, 13, 16, 18, 20, 21, 22, 25, 26, 28,31, 32, 34, 35, 36, 37, 39, 40	22
Good Enough	19, 24, 27, 38	4
Good	2, 3, 5, 6, 7, 8, 9, 14, 15, 17, 23, 29, 30, 33	14

Source: Modification Researcher

The power differential results shown in the table show that there are problems with 22 questions in the "Not Good" category. This may indicate that the questions were not effective in gauging participants' understanding or ability, or that the questions may have been too easy to differentiate between participants with different levels of understanding. "Not Good" questions need to be analyzed to find out the specific problems that exist to be corrected, this needs to involve a joint review of the questions that are not good to get different points of view and the solutions needed. On the other hand, there are 4 questions that fall into the "Medium" category, which shows that although there is a slight separation between participants, there is room for improvement in the preparation of the questions. Meanwhile, there are 14 questions in the "Good" category, this shows that the questions are effective in measuring participants' understanding and abilities well. Therefore, it is necessary to revise or improve the questions in the "Not Good" category so that the exam is more accurate and can provide a more precise picture of the participant's knowledge and skills.

Discriminating power refers to the ability of a question to distinguish between students who have different levels of understanding. Questions with low discriminating power tend to give similar results between students, so they do not provide significant information about individual abilities. On the other hand, questions with high discriminating power can provide different results between students who have different levels of understanding. Therefore, attention needs to be paid to improving or replacing questions with low discriminating power so that evaluations can provide more accurate information about students' abilities.

For the effectiveness of distractors, researchers only took three examples, namely distractors with poor differentiation power, and quite good ones.



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
10	No10	E	Yes	5	1	K, LR

Item statistics

N	P	Total Rpbis	Total Rbis	Alpha w/o
33	0.545	-0.038	-0.048	0.699

Option statistics

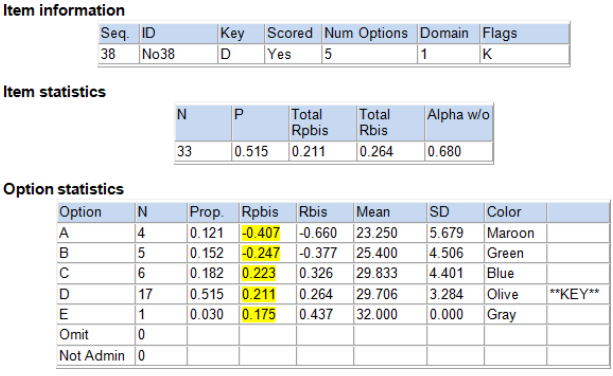
Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	0	0.000	--	--	--	--	Maroon	
B	0	0.000	--	--	--	--	Green	
C	11	0.333	0.045	0.058	28.091	2.625	Blue	
D	4	0.121	-0.006	-0.010	27.750	2.630	Olive	
E	18	0.545	-0.038	-0.048	28.667	5.739	Gray	**KEY**
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
A	0	0.000	0.000	0.000	0.000	0.000	Maroon	
B	0	0.000	0.000	0.000	0.000	0.000	Green	
C	11	0.333	0.500	0.400	0.286	0.222	Blue	
D	4	0.167	0.000	0.400	0.143	0.000	Olive	
E	18	0.500	0.500	0.200	0.571	0.778	Gray	**KEY**

Fig. 2. Results of distractor effectiveness in the "not good" category.

In this case, if item number 10 has a point biserial correlation value ($r\ pbis$) of -0.038 for the answer key, while the $r\ pbis$ value for alternative answers is positive, this indicates that there is a weak and negative relationship between this item and the participant's correct performance. - really understand the material being tested. In this context, a negative $r\ pbis$ (-0.038) indicates that participants who answered this question correctly tended to have a lower overall score on the exam, while those who answered incorrectly had a higher overall score. This could indicate that item number 10 may be measuring a different concept or skill than expected or that there is a problem in the construction of the item. However, keep in mind that a weak relationship (-0.038) can be considered a very low relationship, so its effect on the participant's overall test results may not be significant. Further evaluation needs to be analyzed to understand the reasons behind this negative relationship and whether any improvements need to be made to the questions to more accurately measure participants' understanding of the material being tested.



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
38	No38	D	Yes	5	1	K

Item statistics

N	P	Total Rpbis	Total Rbis	Alpha w/o
33	0.515	0.211	0.264	0.680

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	4	0.121	-0.407	-0.660	23.250	5.679	Maroon	
B	5	0.152	-0.247	-0.377	25.400	4.506	Green	
C	6	0.182	0.223	0.326	29.833	4.401	Blue	
D	17	0.515	0.211	0.264	29.706	3.284	Olive	**KEY**
E	1	0.030	0.175	0.437	32.000	0.000	Gray	
Omit	0							
Not Admin	0							

Fig. 3. Results of distractor effectiveness on one of the questions in the "pretty good" category.

The point biserial correlation value ($r\ pbis$) of 0.211 for question 38 against the answer key indicates that there is a positive relationship between the performance of examinees and the way they answer the question. In other words, participants who answered question 38 correctly tended to have a higher overall test score, while participants who answered incorrectly tended to get a lower score. This shows that question 38 is an effective question in measuring the knowledge or skills tested in that question, and the results can be relied upon as an indicator of participant performance in the exam context. In addition, the positive $r\ pbis$ value for the answer alternative indicates that participants who choose the correct answer alternative tend to have better performance compared to participants who choose the wrong answer alternative. This confirms that item number 38 is able to distinguish well between participants who have a correct understanding of the material being tested and those who do not. Overall, question number 38 can be said to be a fairly good question in the context of the exam, because it makes a significant contribution to measuring participant performance and is able to differentiate between high and low achievers in relation to the material being tested. goods.

To see overall reliability, see Figure 4 below.

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First- Last)	Split-Half (Odd- Even)	S-B Random	S-B First- Last	S-B Odd-Even
Scored items	0.687	2.494	0.532	0.523	0.478	0.695	0.687	0.647
Score	Alpha	SEM	Split-Half (Random)	Split-Half (First- Last)	Split-Half (Odd- Even)	S-B Random	S-B First- Last	S-B Odd-Even
Scored items	0.687	2.494	0.532	0.523	0.478	0.695	0.687	0.647

Fig. 4. Overall reliability results.

The Alpha reliability value of 0.687 which is in the "High" category for the PAS Physics questions shows that the questions have a good level of consistency in measuring the knowledge or skills being tested. This is because Alpha is a statistical measure that shows the extent to which the questions in this test or exam are consistent in measuring the same concept or ability. In this case, the high Alpha value (0.687) indicates that the questions on this exam tend to provide consistent results for participants with comparable levels of knowledge or skills. With high reliability, this exam is more reliable in providing an accurate picture of participants' abilities in the field of Physics. The results obtained from this exam have a high level of consistency, so they can be used with more confidence to make decisions regarding the assessment or evaluation of participants. It also shows that the questions on this exam have been well designed and statistically tested to ensure that they consistently measure the desired abilities. As a result, this exam can provide more accurate information about participants' understanding of the Physics subject.

3.2 **Discussion**

High quality questions are questions that are tested by applying the principles of Classical Test Theory. This shows that they have undergone a rigorous evaluation process and follow recognized standards to accurately measure a person's abilities or knowledge. Thus, questions that adhere to the principles of Classical Test Theory can be relied upon as an effective tool in assessing and evaluating various cognitive and competency aspects.

Based on the initial analysis of the Final School Assessment, an average score of 28.364 was obtained with a Standard Norm of 4.457. The lowest score achieved by a student is 15, while the highest score is 37. With a Mean P of 0.70, the difficulty level of the questions in this Final Semester Assessment can be classified as "Medium". Furthermore, the Mean Rpbis value of 0.210 indicates that the overall level of differentiation is classified as "Medium".

Questions that are divided into the “Easy” and “Medium” categories are considered balanced in their distribution of difficulty, this is important to keep the test or exam fair and able to measure different levels of understanding and ability. With 20 questions covering both categories, teachers or test administrators can test basic understanding as well as more challenging problem-solving skills. This approach provides a more comprehensive picture of participants' knowledge and skills, while reducing the risk of bias that may arise from focusing only on one particular level of difficulty.

The results of the differential power evaluation in the table show that 22 questions are in the "Not Good" category. This may indicate a lack of effectiveness in measuring participants' understanding or a level of difficulty that is too low to differentiate participants with different understandings. Meanwhile, in the "Medium" category there are 4 questions which indicate the need for improvement in the preparation of these questions. However, with 14 questions in the "Good" category, this shows the effectiveness of these questions in measuring participants' understanding and abilities. Therefore, it is recommended to revise or correct

questions in the "Not Good" category to ensure the exam provides a more accurate picture of participants' knowledge and skills.

The Alpha reliability value of 0.687 on the PAS Physics questions shows that the questions have a good level of consistency in measuring the knowledge or skills being tested. Alpha reliability is a statistical measure that shows consistency in measuring the same concept or ability in a test or exam. In this case, a high score (0.687) indicates that the questions on this exam provide consistent results for participants with comparable levels of knowledge or skills. With a high level of reliability, this exam becomes more reliable in providing an accurate picture of participants' abilities in the field of Physics.

From the paragraph above, the questions that have been analyzed still require significant improvement. The average score obtained is not too high, the difficulty level of the questions is at the "Medium" level, and some questions are classified as "Not Good", which shows that most of the questions are less effective in measuring participants' understanding or abilities. Although there are

If there are several questions that are classified as "Good", then corrections and revisions need to be made, especially for questions in the "Not Good" category, to ensure that the exam provides a more accurate picture of the participant's knowledge and skills. With fairly high Alpha reliability, item revision can increase consistency and reliability in measurement.

4 Conclusion

From the results of the research and discussion that have been described, it can be concluded that the initial evaluation of the Final School Assessment shows significant data. The average student score is 28.364 with a Standard Norm of 4.457 showing the diversity of test results. Although there are questions that are classified as "Not Good", the majority of questions fall into the "Good" category, which shows effectiveness in measuring participants' understanding and abilities. The difficulty level of the questions is considered "Medium", which allows for a more challenging test of basic understanding and problem-solving skills. The high Alpha reliability results on Physics questions indicate good measurement consistency, thereby increasing the reliability of the test. Therefore, it is necessary to improve the questions in the "Not Good" category to ensure a more accurate picture of participants' knowledge and skills. Such as unclear instructions, ambiguous wording, or inadequate alignment with learning objectives. Addressing the root of this problem can prevent similar problems in future assessments. Based on the results of interviews with teachers, the reason for the poor quality of the questions is that teachers have very limited time to develop questions and lack of knowledge to check the quality of the questions before they are used by students. Then, based on interviews with students, it was found that the number of questions given was too large and did not correspond to the time given to do it, the length of the questions made it difficult for students to understand the questions and the command sentences used were a little ambiguous. Overall, these data provide important insights for the development and improvement of administering exams that are more effective and accurate in measuring student proficiency in Physics. This research is limited to one high school in the Riau Islands,

for further research it is hoped that it can research the entire Riau Islands province or beyond.

References

1. R. Rosidah, S. Suyanto, and Z. Zafrullah, "Analysis of students' learning interest using E-LKPD based on liveworksheet class VIII junior high school," J. Res. Educ. Res. Eval., vol. 12, no. 1, 2023, doi: https://doi.org/10.15294/jere.v12i1.68227.
2. V. N. I. Sari, A. P. Y. Utomo, and S. Sumarwati, "Kualitas soal bahasa Indonesia di SMP muhammadiyah 1 Pontianak: Analisis butir soal," J. Pendidik. Bhs. dan Sastra Indones., vol. 11, no. 2, pp. 112–119, 2022, doi: https://doi.org/10.15294/jpbsi.v11i2.58091.
3. A. Hamzah, Evaluasi pembelajaran matematika. Jakarta: Rajawali Pers, 2014.
4. T. Kurniawan, "Analisis butir soal ulangan akhir semester gasal mata pelajaran IPS sekolah dasar," J. Elem. Educ., vol. 4, no. 1, pp. 1–6, 2015, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/jee/article/view/7488
5. M. Purwanti, "Analisis butir soal ujian akhir semester gasal mata pelajaran akuntansi keuangan menggunakan microsoft office excel 2010," J. Pendidik. Akunt. Indones., vol. 12, no. 2, Dec. 2014, doi: 10.21831/jpai.v12i2.2710.
6. S. Arikunto, Dasar-dasar evaluasi pendidikan, 3rd ed. Jakarta: Bumi Aksara, 2021.
7. Y. Verawati, F. S. Siskawati, and T. Susilaningtyas, "Analisis butir soal ujian akhir semester (UAS) mata pelajaran matematika pada tahun ajaran 2020/2021 kelas VII SMP islam at tanwir Kecamatan Ledokombo Kabupaten Jember," J. Jendela Pendidik., vol. 3, no. 01, pp. 114–121, Feb. 2023, doi: 10.57008/jjp.v3i01.422.
8. T. M. Haladyna and M. C. Rodriguez, "Using full-information item analysis to improve item quality," Educ. Assess., vol. 26, no. 3, pp. 198–211, Jul. 2021, doi: 10.1080/10627197.2021.1946390.
9. Goleman. et Al., "Analisis butir soal," J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2019.
10. M. Fitrianawati, "Peran analisis butir soal guna meningkatkan kualitas butir soal, kompetensi guru dan hasil belajar peserta didik," in Prosiding Seminar Nasional dan Call for Papers Pendidikan 2017 (PGSD UMS & HDPGSDI Wilayah Jawa), 2017, pp. 282–295. [Online]. Available: https://publikasiilmiah.ums.ac.id/xmlui/bitstream/handle/11617/9117/25.pdf?sequence=1&isAllowed=y
11. N. Ramadhani, "Analisis butir soal ulangan harian pada mata pelajaran bahasa indonesia semester genap tahun ajaran 2018/2019 di SD negeri 7 Jagong Jeget," Universitas Bina Bangsa Getsempena, 2020. [Online]. Available: https://repository.bbg.ac.id/handle/835
12. L. Hamimi, R. Zamharirah, and R. Rusydy, "Analisis butir soal ujian matematika kelas VII semester ganjil tahun pelajaran 2017/2018," MATHEMA J. Pendidik. Mat., vol. 2, no. 1, p. 57, Jan. 2020, doi: 10.33365/jm.v2i1.459.
13. D. J. Ratnaningsih, Isfarudi, and N. Soleiman, "Analisis butir soal pilihan ganda ujian akhir semester mahasiswa universitas terbuka menggunakan pendekatan teori tes klasik," J. Pendidik. Terbuka dan Jarak Jauh, vol. 12, no. 2, pp. 92–99, 2011, [Online]. Available: https://jurnal.ut.ac.id/index.php/jptjj/article/view/431
14. A. S. Halik, S. Mania, and F. Nur, "Analisis butir soal ujian akhir sekolah (UAS) mata pelajaran matematika pada tahun ajaran 2015/2016 SMP negeri 36 Makassar," Al Asma J. Islam. Educ., vol. 1, no. 1, p. 11, May 2019, doi: 10.24252/asma.v1i1.11249.

15. M. Nurjanah, I. Istiningsih, and H. Mangkuwibawa, “Analisis kualitas butir soal pilihan ganda tema 7 indahnya keragaman di negeriku kelas IV madrasah ibtidaiyyah,” J. Paedagogy, vol. 9, no. 4, p. 817, Oct. 2022, doi: 10.33394/jp.v9i4.5299.
16. S. S. N. Mukrimaa et al., “Dasar metodologi penelitian,” J. Penelit. Pendidik. Guru Sekol. Dasar, vol. 6, no. 128, 2016.
17. F. F. Ida and A. Musyarofah, “Validitas dan reliabilitas dalam analisis butir soal,” AL-MU’ARRIB J. Arab. Educ., vol. 1, no. 1, pp. 34–44, Dec. 2021, doi: 10.32923/al-muarrib.v1i1.2100.
18. A. Sukmafitriani, F. S. Munjariyati, W. Wagiran, and D. L. Naryatmojo, “Analisis butir soal penilaian keterampilan apresiasi sastra pada soal UAS materi puisi kelas VII tahun pelajaran 2019/2020 di SMPN 1 Kandanghaur,” Asas J. Sastra, vol. 10, no. 2, pp. 111–122, Jul. 2021, doi: 10.24114/ajs.v10i2.26277.
19. H. Retnawati, “Analisis butir soal dengan pendekatan teori tes klasik dan teori respon butir. pelatihan analisis butir soal dan pemanfaatan hasil ujian bagi guru,” in Pelatihan Analisis Butir dan Pemanfaatan Hasil uJian Bagi Guru di SMK 2 Tarakan Kalimantan Timur 30-31 Maret 2012, 2012, pp. 1–19.
20. A. Muhson, B. Lestari, Supriyanto, and K. Baroroh, “Pengembangan software analisis butir soal yang praktis dan aplikatif,” J. Ilmu Pendidika, vol. 20, no. 2, pp. 207–216, 2014, [Online]. Available: https://media.neliti.com/media/publications/110163-ID-pengembangan-software-analisis-butir-soa.pdf
21. Mahapsari, “Landasan teori bab 2,” Prediksi, vol. 66, no. 1997, pp. 37–39, 2013.
22. E. Prasetyo and Wahyudi, “Pengembangan instrumen penilaian kognitif pada pembelajaran tematik terpadu kelas 4 SD,” Pionir, vol. 9, no. 2, pp. 165–182, 2020, doi: http://dx.doi.org/10.22373/pjp.v9i2.9280.
23. B. I. Sappaile and T. Pristiwaluyo, “Analisis butir soal ujian sekolah berstandar nasional dengan pendekatan klasik dan teori respon butir mata pelajaran matematika,” 2019. [Online]. Available: https://ojs.unm.ac.id/semnaslemlit/article/view/11384
24. W. Widayanti, Bistari, and Suparjan, “Analisis butir soal pilihan ganda penilaian tengah semester pada pembelajaran tematik kelas V sekolah dasar negeri 39 Pontianak Kota,” J. Didika Wahana Ilm. Pendidik. Dasar, vol. 9, no. 2, pp. 279–296, 2023, doi: https://doi.org/10.29408/didika.v7i2.4370.
25. D. Saepuzaman, E. Istiyono, Haryanto, H. Retnawati, and Yustiandi, “Analisis karakteristik butir soal fisika dengan pendekatan IRT penskoran dikotomus dan politomus,” Radiasi J. Berk. Pendidik. Fis., vol. 14, no. 2, pp. 62–75, 2021, [Online]. Available: https://garuda.kemdikbud.go.id/documents/detail/2315487
26. S. Nurinda, E. Rudyatmi, and S. Ridlo, “Analisis butir soal olimpiade biologi SMA tingkat kabupaten/kota tahun 2013,” J. Biol. Educ., vol. 3, no. 1, pp. 77–84, 2014, doi: https://doi.org/10.15294/jbe.v3i1.4161.
27. S. Azzahroh, F. L. Iman, B. Anwar, and R. Aziz, “Analisis butir soal ujian akhir semester mata kuliah psikologi belajar menggunakan software anates,” J. Indones. Psychol. Sci., vol. 2, no. 2, pp. 226–252, Dec. 2022, doi: 10.18860/jips.v3i2.17228.
28. Sugiyono, Metode penelitian pendidikan pendekatan kuantitatif, kualitatif, dan R&D. Bandung: Alfabeta, 2017.
29. E. Istiyono, Pengembangan instrumen evaluasi dan analisis hasil belajar fisika dengan teori tes klasik dan modern. Yogyakarta: UNY Press Yogyakarta, 2018.